1-1-2015

What You Know Counts: Why We Should Elicit Prior Probabilities from Experts to Improve Quantitative Analysis with Qualitative Knowledge in Special Education Science

Tyler Aaron Hicks
University of South Florida, jrhicks7@gmail.com

What You Know Counts:

Why We Should Elicit Prior Probabilities from Experts to Improve Quantitative Analysis with Qualitative Knowledge in Special Education Science

by

Tyler A. Hicks

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Curriculum and Instruction with emphases in
Measurement/Evaluation and Special Education
Department of Educational and Psychological Studies
Department of Teaching and Learning
College of Education
University of South Florida

Co-Major Professor: Phyllis Jones, Ph.D.
Co-Major Professor: Jeffery Kromrey, Ph.D.
Douglas Jesseph, Ph.D.
Lyman Dukes III, Ph.D.
Jennifer Wolgemuth, Ph.D.

Date of Approval:
March 3, 2015

Keywords: Induction, Bayesian statistics, Bayesian epistemology

DEDICATION

Soli Deo Gloria.

ACKNOWLEDGMENTS

My dissertation would be incomplete without an acknowledgement of those folks who united behind me in its production. First and foremost, I wish to acknowledge my beloved wife, Jacky. She was a refuge of sanity, and a beautiful reminder that life is bigger than dissertations. I also thank my daughter, Kara, who, though still in the womb, gave me extra motivation to finish. Moreover, I owe a debt to my parents, Janet and James. They encouraged me to be curious about reality, and my childhood inquisitiveness was nurtured in conversations I had (and still have) with my smarter siblings, Kyle and Jenna, about God, immortality, and freedom.

I also wish to acknowledge my esteemed professor, James Paul. He greatly influenced the trajectory of my studies that eventuated in my dissertation. He also enlarged my thinking of the linkage among philosophy, methodology, and special education, and masterfully articulated the prevailing need to prioritize the examination of the foundations of education science.

I thank my faithful colleague, Mike Riley. His invaluable edits clarified and improved my dissertation. I acknowledge Kyle Mills and Stephen Wright too for their constant intellectual stimulation. They are *amateur* scholars in the best sense of the word ("they do it for love rather than the money"). I also acknowledge the generous help of my fellow participants in a student-directed writing group, Joshua Barton, Wendy Bradshaw, Noel Cherasaro, Pakethia Harris, and, Mehmet Ozturk. I am also grateful for insights elicited from faculty into my dissertation topic. In particular, I must acknowledge David Allsopp, Leonard Burrello, Robert Dedrick, Jeannie Kleinhammer-Tramill, Alex Levine, and Liliana Rodriguez-Campos.

I especially acknowledge my peers, Elly Beak and Seang-Hwane Joo. Their observations improved the quality of my philosophical argument in Chapter 2. I also thank my former statistics professor, John Ferron, for catching an embarrassing error I made in an early draft of Chapter 2. I alone, of course, am responsible for in any remaining errors in Chapter 2.

Next, I would like to acknowledge the use of services provided by Research Computing at the University of South Florida. Without their services, my Chapter 3 would not have been finished in timely fashion. I am indebted then to George MacDonald who brought their services to my attention, and then persuaded me with stories drawn from his own personal experience that this really was the most efficient means at my disposal for simulating data.

I would also like to acknowledge David Hoppey, Bill Black, and Ann Mickelson. They supplied me with an unpublished manuscript about a qualitative case study they had conducted comparing two schools. Their exemplary analysis of the cases prompted me to undertake the study in Chapter 4.

I close out this acknowledgement with inadequate words of thanks to all my committee members. To my Co-Major, Phyllis Jones, thank you for improving my dissertation's relevance. To my Co-Major, Jeffrey Kromrey, I am so thankful that a scholar with your formidable analytic skills took an interest in my dissertation. To Douglas Jesseph, thank you for helping me add a layer of philosophical depth to my analysis of statistical inference. To Lyman Dukes, thank you for inviting me to pursue doctoral studies in special education, and for getting me ready and equipped to complete them. To Jenni Wolgemuth, thank you for an ongoing dialogue throughout the dissertation process about the possibility of bridging some gaps between quantitative and qualitative analysis with Bayesian inferences. I have yet to figure out if such bridges can hold the weight or not, but I look forward to more conversations about it with you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Qualitative knowledge is about types of things, and their excellences. There are many ways we humans produce qualitative knowledge about the world, and much of it is derived from non-quantitative sources (e.g., narratives, clinical experiences, intuitions). The purpose of my dissertation was to investigate the possibility of using Bayesian inferences to improve quantitative analysis in special education research with qualitative knowledge.

It is impossible, however, to fully disentangle philosophy of inquiry, methodology, and methods. My evaluation of Bayesian estimators, thus, addresses each of these areas. Chapter Two offers a philosophical argument to substantiate the thesis that Bayesian inference is usually more applicable in education science than classical inference. I then moved on, in Chapter Three, to consider methodology. I used simulation procedures to show that even a minimum amount of qualitative information can suffice to improve Bayesian *t*-tests' frequency properties. Finally, in Chapter Four, I offered a practical demonstration of how Bayesian methods could be utilized in special education research to solve technical problems.

In Chapter Five, I show how these three chapters, taken together, evidence that Bayesian analysis can promote a romantic science of special education – i.e., a non-positivistic science that invites teleological explanation. These explanations are often produced by researchers in the qualitative tradition, and Bayesian priors are formal mechanism for strengthening quantitative analysis with such qualitative bits of information. Researchers are also free to use their favorite qualitative methods to elicit such priors from experts.

CHAPTER ONE

INTRODUCTION

Is it ever advisable to let an expert with qualitative knowledge influence our statistical analysis? I think so. My guess is their potential contribution to the work of quantitative analysts is far reaching. Their judgments may even improve how we make sense of experimental findings in special education research. The purpose of this dissertation is to test my hunch.

Imagine the principal investigator of a research project hands over the results of an experiment to her statistical analyst, Jen. She asks Jen to see if the intervention of interest raises state test scores. Jen knows nothing about the context of the experiment, besides what is in the data set. She never visited the school site where the study took place, saw the new intervention implemented, or watched how participants reacted.

In contrast, Lisa is a renowned qualitative researcher with a wealth of knowledge. She makes ongoing visits to the school site where the experiment was done. Wholly incidental to her personal research interests, she happened to observe educators using the selected intervention in classroom settings. She never saw test scores, but she understands how the intervention changed the classroom's learning culture better than anyone else.

In this scenario, I propose that Jen should incorporate Lisa's expertise into her statistical analysis. Ignoring such expertise needlessly impoverishes inferences. The experimental data are one piece of evidence; Lisa's expert judgment another. Estimates of intervention effects should be based on all the evidence. The purpose of this dissertation is to shed some light on the possibility of drawing on qualitative knowledge to improve quantitative analysis.

Some Preliminaries

I hate it when an author begins making distinctions before the discussion gets underway. Why not explain everything along the way? With one exception, that is what I plan to do. This section, however, is the exception. It discusses an issue central in the philosophy of statistics which if explained first helps everything else go much more smoothly.

On my preferred account, statistical inferences are formal inductions. A young child encounters several white swans swimming in a lake. She induces all swans in the world are white. Thereafter she sees more white swans, so she now holds her induction more firmly. But inductions are risky. In this case, her induction is dead wrong – black swans exist in New Zealand. Cogent inductions, despite our every effort, sometimes go astray.

Statisticians invoke probability to make the risk of error mathematically precise (Hand, 2008). Probability theory is now a respectable branch of modern mathematics; its theorems derivable from axioms (Kolmogorov, 1933/1956). The math itself, however, underdetermines probability's nature: chance or credence (Hacking, 2001). Chance is an objective force of nature. The outcome of any *chance setup*, such as a coin toss, is random. But, in the limits, repeated flips of the same coin produce a certain proportion of heads and tails (Von Mises, 1936). This empirical convergence happens regardless of who is observing.

Alternatively, probability is construable as a subjective entity. A group of young people bet on the World Series. One bets the Yankees will win; others think her confidence misplaced. We represent (not measure) degree of opinion using probabilities. We assign probabilities to each belief an agent holds. The young polymath, F.P. Ramsey (1926) showed adherence to the axioms of the formal probability calculus yields consistency among beliefs. Probability, then, is equally interpretable as objective chance or subjective credence.

Statement of the Problem

Education research, especially in the area of special education, underwent an astonishing qualitative revolution in the last couple of decades (Denzin & Lincoln, 2008; Pugach, 2001). Qualitative researchers borrow and adapt methods from the social sciences, fine arts, and humanities to study education (Hatch, 2002). This revolution has gone a long way towards building a desirable romantic science of special education – a science inclusive of qualitative knowledge (Smith, 2003). In the last couple of decades there has also been an appreciable rise in computational power (Lynch, 2007). New algorithms, like the Markov Chain Monte Carlo method, make it possible to derive otherwise intractable integrals (Muthen, 2013).

Building the special education knowledge base entails good stewardship of all our new assets (Odom, et al., 2005). The problem is figuring out how to wisely apply our growing qualitative knowledge and computational power to improve statistical inferences in education research. I wish, therefore, to investigate whether special education researchers should now boldly embrace tools established in mainstream statistics which are capable of harnessing both qualitative knowledge and computational power to make the most of small data sets.

The Significance of the Problem

The above problem matters to the field of special education. To see this, only consider the tug of war between *clinical pedagogy* and *evidence-based pedagogy* – a name obviously chosen for rhetorical effect (Hacking, 2006). Evidence-based pedagogy means randomized experiments, meta-analyses, and statistical regularities (Glass & Hopkins, 1996). Clinical pedagogy is about teachers basing their decisions on personal interactions with their students. Such decisions hold a distinct qualitative flavor defying entrapment in numbers.

Some in the United States champion evidence-based pedagogy in special education (Kauffman, 2011; Odom et al., 2005). This is the culmination of a long historical process (Lagemann, 2000). The evidence-based pedagogy platform is admittedly enticing for politicians hoping to garner public support. Evidence-based pedagogy is not a panacea, but somehow it seems consistent with our longings for neutral public education. We are alarmed at the thought of teachers imposing their own private values on students. Labaree (2011) explains our romance with statistical regularities in education research is a consequence of our cherished democratic aspirations rather than as a result of technical progress.

I think the either-or dilemma of clinical pedagogy or evidence-based pedagogy is ill-posed. Why not both? The mathematical apparatus and techniques needed to integrate them exist. All we need to do is apply them. Such integration can increase our power to detect treatment effects, estimate complex models, and capitalize on small data sets (Gill, 2002; Iversen, 1984; Kruschke, 2011). The incorporation of qualitative knowledge in our statistical analyses will allow us to ask new questions and answer old questions in new ways.

The enlistment of informants with qualitative knowledge is most useful when researchers wish to (1) see how credible their hypothesis is after observing the data (which is not what the *p*-value does), (2) calculate an interval containing a range of probable values for the invisible parameters (which is not what the confidence interval does), or (3) evaluate the evidence for and against statistical models in their model competition (Muthen, 2013). These three claims may seem extravagant to those only familiar with traditional *p*-values and confidence intervals, but there is nothing controversial about these promises *per se* - at least from a purely mathematical point of view, granting of course, that our informants are accurate (Duke Evidence Based Practice Center, 2009).

Theoretical Perspective

Philosophy of inquiry, methodology, and methods are an interlinked chain in the research process (Guba & Lincoln, 1994; Morgan, 2007; Crotty, 1998; Paul, 2005), yet a chain is only as strong as its weakest link (Krathwohl, 1998). Method only describes what scientists do while doing research. Methodology prescribes what scientist should do while researching. Philosophy of inquiry encompasses a constellation of problems, such as: (a) What is a science of education? (b) How closely, if at all, should it resemble the natural sciences? and (c) Are teleological explanations applicable in education research? These are questions requiring much philosophical investigation to answer. At the intersection of philosophy of inquiry, methodology, and methods then stands the problem of inputting qualitative knowledge into statistical analysis in education research.

The relationship between philosophy of inquiry, methodology, and methods is depicted in Figure 1. A philosophy of educational inquiry is needed to intelligibly make the seemingly reckless move from qualitative expertise to statistical inference. We also need a methodology showing how and when to make such bold moves. Practical demonstrations of such methods are also needed to bring such a method into the reach of special education researchers.

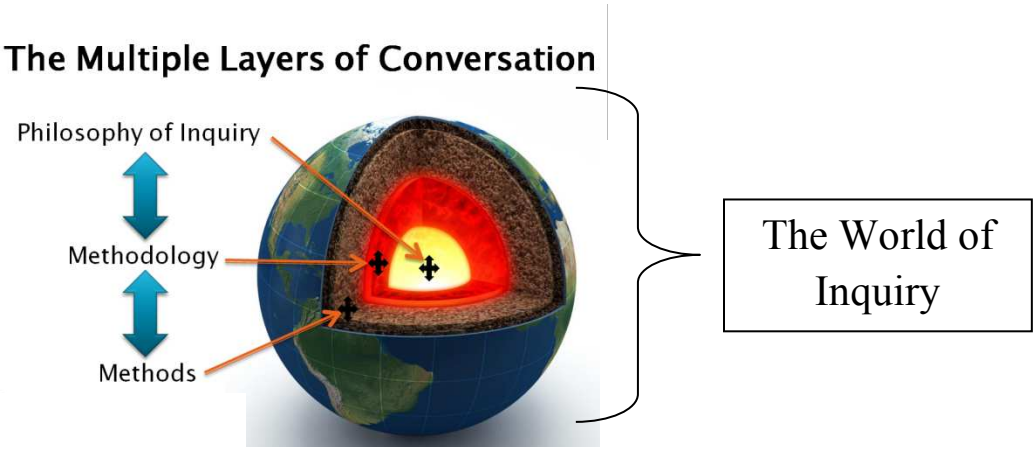

Figure 1. *Relationship between philosophy of inquiry, methodology, and methods*

As a historical observation, the interdependence between philosophy of inquiry and methods is appreciated among qualitative researchers more so than quantitative researchers (Yu, 2006). But when it comes to the issue of inputting qualitative information into statistical analysis the intersection of philosophy and statistics becomes visible. I am not claiming that the methods evaluated in my dissertation are only for philosophers. Knowledge of the philosophy of statistics and educational inquiry is helpful, but not essential. To put the same point another way, one does not need to know that one knows in order to know something (Plantinga, 1993).

To see why the above statement is true, pretend a naive researcher conducts an experiment. Now imagine she learns her intervention works. She does not understand the methodology or philosophy behind experimental methods. She used randomized assignment, but does not know why. She drew causal inferences, but has no theory of causation. In this scenario, our researcher, despite unfortunate gaps in her philosophy and methodology, knows her intervention works. She knows this without knowing why she knows it.

Some philosophical and methodological obligations, of course, must be met before researchers can input qualitative knowledge into their statistical analyses. To appraise the value of incorporating qualitative expertise into statistical analysis, researchers need a sensible philosophy of educational inquiry (Phillips, 1987). But researchers can outsource here. They can let philosophers who investigate such problems articulate passable theories on their behalf. Researchers can also outsource to methodologists. Methodologists are better positioned to address technical questions than researchers who may lack specialist training. So long as philosophical and methodological obligations are competently fulfilled on behalf of researchers, they can trust the research process. This division of labor is essential to building up the education knowledge base, if not all of science (Williams, 2001).

Overview of the Literature

Scientific Realism

Educational processes resemble natural lotteries, their outcomes probability distributions. Only think about how IQ scores follow a normal distribution for an example. Equations – i.e., statistical models – capture the laws of probability controlling such distributions. Researchers try to induce such models from samples. Statistical models are useful for prediction. But scientific realists insist models can give us clues about the nature of reality too.

Realism is not a precise thesis, but a vague picture. In this depiction, the world is *outside* us, and the purpose of theorizing – or a significant segment of it - is to describe its ontology (theory of things) in its own terms (Sider, 2011). To borrow a metaphor from Plato's writings, we want theories to carve up the world at the joints. We realists think successful theories are likely true, but we also think their ontological-ideology likely fits reality's structure. The epistemology (theory of knowledge) of such a privileged ontology is murky, yet Quine's (1969) advice seems sensible: Believe in the ontology of your best scientific theories.

Many realists make space for both credence-type and chance-type probability in their metaphysics. David Lewis connected theses probabilities together in his formulation of the principal principle. Roughly speaking, suppose we know the chance that event A will occur at some future time *t* is x, then – regardless of other admissible evidence we examine before time *t* – our credence in an hypothesis, H, stating A will occur at time *t*, should be x too, briefly:

$$C(H|X, E) = x$$

Where C is any reasonable credence function, x is any real number in the unit interval, E is any other admissible evidence, and H is the hypothesis A occurs at time *t*. Note, credence functions assign probabilities to judgments about the credibility of propositions.

To illustrate, Lewis asks us to consider the proposition a fair coin tossed at noon falls heads. Lewis' rule clearly states we should be 50% confident it will fall heads at the appointed time. Suppose, however, we become privy to additional information, such as we observed this fair coin land tails rather than heads approximately 90% of the time in past experiments. No matter. So long as we know this coin is really fair (and its fairness is key), Lewis counsels us, to be 50% confident it will fall heads. Our uncertainty about whether the coin lands heads at noon tomorrow, thus, partly depends on our certainty the coin is really fair. So formulated, this principle provides ample justification for statistical modeling in education.

In real-life, however, we never know with certainty whether a coin is fair or not. Likewise, in education research, we never know if our favorite statistical models are true or not. We are wise then to lavishly buttress models with as much evidence as possible before accepting them or even acting on them (Phillips & Burbules, 2000). But how to do so is a matter of no small debate. Some influential methodologists, however, resort to Bayesian epistemology to select and assess models (Gelman & Shalizi, 2013).

Bayesian Epistemology

Let S denote a sample and M_i a statistical model of interest. In the chance scheme, we can imagine S is a sample drawn randomly from some population. This population is depicted by M_i. Consequently, $P(S|M_i)$ becomes a palatable notion, but $P(M_i|S)$ none-sense on stilts. There is no chance about M_i being true or false given the sample; it is or it is not. Its truth value is not a matter of chance. But Bayesian statisticians can make sense of $P(M_i|S)$. To them, it is a matter of credence-type probability – not chance-type probability. It represents their subjective credence towards M_i given S.

Some Bayesian statisticians are dogmatic. They reject all talk of chance-type probability as myth, but Bayesian statisticians need not be so exclusivist. They can co-opt chance-type probabilities. They can, for example, think chance-type probability is a physical reality, and $P(S|M_i)$ can be understood within a chance-type probability scheme when S is a random sample. The world is a magical place, and quantum mechanics seems to warrant the reality of chance-type probability.

All Bayesians, regardless of their metaphysical outlook, warn us it is an instance of the gambler's fallacy to equate inverse probabilities, such as $P(S|M_i)$ and $P(M_i|S)$. Generally speaking, $P(S|M_i) \neq P(M_i|S)$. For example, the probability that someone playing a game of cards is dealt a queen given she is dealt a heart is not the same as the probability she gets a heart given she gets dealt a queen. But Bayesian statisticians know a fancy trick for transforming $P(S|M_i)$ into $P(M_i|S)$. They, thus, can admire the ways statisticians reach a consensus about $P(S|M_i)$ using chance. But, unlike them, they will go on to transform it into $P(M_i|S)$.

A one-line theorem from the probability calculus about the relationship between inverse probabilities stands at the core of the Bayesian paradigm; briefly:

$$P(M_i|S) \propto \lambda(S|M_i)P(M_i)$$

where this equation is read as the "posterior" probability of the i^{th} statistical model given sample S is proportional to the product of the likelihood of S given the i^{th} model and "prior" probability of the i^{th} model $P(M_i)$. The likelihood function $\lambda(M_i|S)$ represents how probable S is given M_i. The prior $P(M_i)$ represents our initial credence towards some M_i before examining the sample data. This theorem was first discovered by the eponymous English Puritan, Thomas Bayes (1701-1761), to solve some esoteric problem in probability theory. But it has since become the foundational formula of Bayesian epistemology.

The engine of Bayesian inference is the prior probability, $P(M_i)$. But formal mechanisms underdetermine our choice of prior density in specific applications. One Bayesian might select a prior representing a skeptical stance. A more trusting Bayesian, however, might select a prior representing a credulous stance. This is not a mathematical mistake, but a mere difference of opinion. Outside of the extreme cases, what constitutes the best possible prior in any particular circumstance is always a matter of qualitative judgment (i.e. judgment informed by good sense rather than a mechanical formula). This subjectivity makes many squeamish, but selecting an α-level for the familiar *p*-value (ex., .1, .05, .01) is likewise subjective.

To critics, however, prior probabilities introduce an unacceptable degree of subjectivity into the selection and assessment of models. But the sensational track record of Bayesian statistics reversed this verdict (Silver, 2012). In her book, "The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy," McGrayne (2012) retells how one Bayesian triumph after another, finally forced the majority of people to acknowledge the positive role priors can play in even rigorous scientific investigation.

Applications of Bayesian statistics, albeit established in mainstream statistics, continue to lag behind in special education research (Hicks & Knollman, 2014). Statistical analysis without priors dominates quantitative research in special education. It is impossible to be sure, but this state of affairs may change in the future as interest in Bayesian applications seems to be on the ascendency among researchers. But there are far too many contingencies to make any bold predictions about the future with much certainty. It is also difficult to envision how a popular version of Bayesian statistics will soon materialize in education research, even if we guess it eventually will gain a foothold.

Methodologists who subscribe to some version of Bayesian epistemology are very far from being a unified lot of folks. They have their fair share of disputes, like any other family. Some argue statistical inferences are fancy decisions. They argue we will never know if our best models are true, but we can know if they are useful. They prefer models that balance the dangers of Type I and Type II errors. But, of course, as Bayesians they do this within a comprehensive Bayesian decision framework. They, consequently, try determine how probable each model is given the sample evidence, and then evaluate the risks and rewards of acting on any one of the models should they really be true or false (the choice not to act is also a decision).

Other Bayesians, in contrast, claim statistical inferences are inductions - not decisions. They then assess whether a model viewed in relative isolation is plausible or not given the data. This model is usually considered to be the null model. It represents a model, which if true, is so dull it entails nothing of substantive interest to researchers. Researchers, thus, hope to show that this model is probably false. To accomplish this masterstroke, these Bayesians can compute the posterior probability of the null model (Kruschke, 2011). If this posterior probability is too low, then they deem the data significant (otherwise, insignificant). Significance here only entails this area of research is probably capable of repaying future efforts (Mohr, 1990).

Subjective Priors in Bayesian Inference

There are two types of priors, objective and subjective. Objective priors are set using formulas. They are objective because everyone using the formula reaches the same prior. But there is no consensus about which objective prior among all the possibilities, if any, is obligatory in any situation. Subjective priors, in contrast, encapsulate agents' background knowledge. They are agent-relative: Different agents, different priors. Again, it takes qualitative judgment rather than a fancy formula to set up a cogent subjective prior.

A teacher claims she can tell whether a full moon occurred last night by tallying up student misbehaviors. A drunken man boasts he can predict the outcome of a fair coin toss. Highly esteemed testing experts predict any 1st grader who passes their scholastic aptitude test above a certain threshold will go on to pass the 1st grade state reading test too. Skeptical, we put each claim to the test. In ten out of ten trials, each one accomplished their boasted feat.

A classical (non-Bayesian) significance test would assess the null hypothesis (H_0) that $\theta = .5$ (i.e., the claimant is only guessing). In all three situations this hypothesis is rejected at the .05 significance level as classical tests omit priors altogether, but reaching the same conclusion in each case seems counter-intuitive. Bayesian epistemology shows why thing went wrong: We forgot about the priors.

The prior probability of each claim differs. We are open-minded to the claims of a trusted expert. The evidence then only strengthens our initial conviction. But, regardless of the evidence, most of us will remain skeptical towards the drunken man's claim. It goes against everything we know about the effect of high concentrations of alcohol on the brain. In regards to the teachers claim about full moons, it is not quite clear what to think. Some teachers swear full moons influence student behavior. Others surmise it is superstitious nonsense. In this case reasonable people will find the evidence inconclusive because of their personal priors.

In special education research settings, we can inform subjective prior distributions with clinical knowledge. This knowledge is distributed among practitioners, experts, and other key stakeholders. The qualitative revolution in special education research has shown that there is an abundance of such knowledge at our disposal. Using qualitative methodology, it may be possible to elicit such knowledge from key informants. This knowledge, albeit perhaps gathered through qualitative methodology, can itself later be quantified by Bayesians.

The statistical literature on eliciting priors from experts is scanty, but as a general rule researchers can use their favorite methods for doing qualitative research to understand their informants' background knowledge. Quantifying this knowledge afterwards, however, is challenging, albeit, not impossible (Christensen, Johnson, Branscum, & Hanson, 2011). Several methods can assist Bayesian analysts in the process of constructing representative priors of prevailing expert opinion (Buckley, 2004). Again, the goal is only to imperfectly represent – not precisely measure – their judgment about the probability of some proposition.

There are three steps to quantifying knowledge elicited from informants. Bayesian analysts should: (1) specify a distributions family for the prior density (ex. normal-shape, *t*-shape, uniform-shape, etc), (2) identify parameter values (ex. mean, standard deviations, etc) for the prior, and then (3) cross-validate their final choice of prior. The last step involves double-checking with the informant to make sure the final prior density fairly and defensibly represents her judgment on the matter (Gill & Walker, 2005).

To illustrate, imagine classroom teachers possess valuable background knowledge about the youth they teach on a regular basis. They anticipate how their students will respond under intervention and control conditions. But they, of course, are not infallible informants. Their best clinical intuitions sometimes go astray. So, we need to counter-balance their judgments with empirical data. To do this, we setup a little experiment in the classroom.

In the experiment, we randomly assign youth to intervention and control conditions and observe what happens. To analyze the data, we use a Bayesian independent means *t*-test to assess whether these youth differ on average on some dependent variable of interest. To set appropriate subjective priors for this analysis, we will just use our favorite qualitative methods to understand the clinical judgments of these students' teacher.

Nothing hinges on what qualitative method researchers use to understand their informants perspective so long as gain enough understanding of their partial beliefs about all the hypotheses of interest to fairly represent them. We can then quantify beliefs with a probability distribution using steps 1-3: We may for instance specify a normal-shape for the prior density, identify a certain mean and variance for it, and then cross-validate it with the teacher. Using this subjective prior, Bayesians can then "conditionalize" on the objective sample evidence to compute the requisite posterior probability distribution. Conditionalize is Bayesian-talk for transforming prior probabilities into posterior probabilities using Bayes' formula.

All of the above steps for quantifying an expert's qualitative knowledge can be done in brief, one-on-one interviews between Bayesian analysts and informants. In situations where there are multiple experts at hand, Bayesian analysts may need to get more sophisticated in how they elicit priors. A skeptical audience may be more impressed with a prior set by a panel of experts than a prior set by a *rouge* expert. So, despite the added costs in complexity of constructing a single prior representative of the opinion of multiple experts, these costs are compensated by the added credibility with readers.

There are different ways one might approach the problem of inputting several expert opinions into one representative prior. One option is to apply steps (1)-(3) with the panel of experts in a focus group setting. This procedure is risky as it depends on expert collaboration and consensus. But the Bayesian analysts can facilitate the discussion to keep conversation on topic. When consensus cannot be reached then this uncertainty can be modeled with second-order priors over first-order priors. The other option is to apply steps (1)-(3) individually. Then one can take expert responses and apply regression analysis to aggregate responses to get a representative prior of the collective opinion (Gill & Walker, 2005).

Assessing Subjective Prior Densities

The subjective prior is the engine of Bayesian inference (Gill & Walker, 2005). Fisher's famous effort to formally constrain the *subjectivity* in the prior proved intractable. Given that statistics tends to draw the formally inclined this has been one reason why the use of priors has often *felt* to classically-minded statisticians like subjectivity run amok (Sider, 2011). But Bayesian statisticians figured out some ways of ruling out *garbage* priors. These methods are informal and not foolproof, but they do help. These methods include transparency, triangulation of multiple priors, and posterior predictive checks.

Full disclosure of how priors were set offers an important constraint on prior choice. Bayesians seeking to publish in peer-reviewed journals must persuade critical reviewers who are often skeptical of their findings that their priors are indeed sensible. This means that analysts must formulate some explicit, written rationale for their choice of priors. These justifications empower readers to reach their own conclusions about the reasonableness of the choice of prior. Kruschke (2011) argues that this transparency in setting priors actually makes applications of priors in Bayesian statistics less prone to abuse than the alternative "canned" α-levels researchers routinely invoke using non-Bayesian methods.

Triangulation of multiple priors for comparative purposes offers yet another way of constraining priors (Pollard, 1986). Convergences in the posteriors of Bayesian analyses done with *skeptical* priors, *neutral* priors, and *advocacy* priors toward some hypothesis shows that choice of prior had only minimal influence on the posterior. This often happens in situations where large samples are at our disposal (Kruschke, 2011). If convergence fails to obtain, however, and choice of prior matters then readers can pick the prior that best represents their own stance on the issue.

The posterior predictive check is actually a method for testing the overall goodness-of-fit of the posterior distribution rather than for constraining prior choice (Lynch, 2004). But we can use this method as a backwards check on our choice of prior after the analysis is over. Imagine we used Bayesian statistical analysis to get a posterior distribution of a population mean. We can then simulate drawing random samples from our posterior distribution to determine the chances (yes, chances) of getting our obtained sample mean. This is a posterior predictive check. If this chance is high we have corroborated our statistical model, else, we produced evidence against it. Again a *bad fit* only entails that the choice of prior *might* possibly have been the culprit. Other culprits such as a fluke random sample may instead be to blame.

But there is one more informal way of constraining the choice of prior yet to be discussed: Ask experts for help in setting up priors (Buckley, 2004). Expert opinion represents prevailing opinion on hypotheses, before taking into account the sample information. Besides experts, we could use previous quantitative studies, common sense, or "theoretical" knowledge to inform our choice of priors. Incidentally, basing priors on known statistical regularities is fully compatible with the Bayesian spirit (Lewis, 1987). But, in their absence, special education researchers might try the route of content experts. It seems only natural to make priors represent expert opinion (Christensen, et al., 2011).

Buckley (2004) proposes Bayesian analysts exploit the knowledge of qualitative researchers to set defensible priors (see also Gill & Walker, 2006). Some qualitative researchers are choice informants in this regard because they are aware how human agency and social institutions may moderate the chance that an event occurs. In instances where past quantitative studies have been done, it may make sense to even blend this past quantitative information with qualitative expertise when setting priors.

Dissertation Objectives and Organization

My priority in this dissertation is to test my hunch that qualitative knowledge can improve how we analyze quantitative data sets in special education. Based on the linkage of philosophy of inquiry, methodology, and practice in research, three objectives guide this dissertation research: (1) philosophically speaking, I will build an argument that, despite its risks, Bayesian inference can cope better than classical inference with the hazards of doing education research (ex. frequent inability to use random assignment). (2) Methodologically speaking, I will test my hunch qualitative knowledge can improve statistical analysis. Using the performance of classical *t*-tests as a baseline, I will evaluate the Type I and Type II error rates of Bayesian *t*-tests under different conditions, such as tiny sample sizes. (3) Practically speaking, I will demonstrate how I envision using Bayesian methods in special education science. Specifically, showing how Bayesian analysis can sometimes produce results more informative and helpful than traditional classical analysis.

The three chapters comprising the body of my dissertation each addresses one objective. Table 1 offers a summary of fit between article and objective. These articles form a cumulative case for fostering a romantic science of special education using Bayesian inference. Moreover, I think Bayesian inference will still satisfy those researchers among us who demand technical rigor in statistical analysis at all times. This is because it provides a kind of logic of quantitative analysis that is formal yet comfortable with qualitative knowledge.

Table 1. *Objectives and Methods*

Objectives	Description of Objectives	Chapters
#1	Build an argument that, despite risks, special education should befriend Bayesian inference	Philosophy Piece
#2	Determine whether even limited qualitative knowledge can improve statistical analysis	Methodology Piece
#3	Show how to apply Bayesian statistics in special education to solve otherwise intractable problems	Pedagogical Piece

In chapter 2, I will cautiously venture into the depths of statistical philosophy. Using John Rawl's method of reflective equilibrium, I marshal an argument for the thesis that Bayesian inference is generally more applicable than classical inference in education research. This is because classical inference needs probability samples or random assignment to keep its promise of delivering objective inferences grounded in chance-type probability. I argue that this is true whether we are inducing population characteristics or causation. Bayesian inference, in contrast, delivers a different version of objectivity to its user, and it does not need probability samples or random assignment to keep its promises. The upshot, Bayesian and classical inference are different versions of statistical inference. One is not better than the other, but Bayesian inference is more generally applicable in education research.

In chapter 3, I report the results of a Monte Carlo experiment comparing several methods for setting convenience priors when doing Bayesian *t*-test analysis. Convenience priors are respectable default settings for priors in Bayesian analysis. The purpose of this simulation study was to help guide researchers in selecting reasonable convenience priors in research contexts involving tiny samples. I begin this chapter surveying the problem of setting up priors in pioneering research. Setting up priors is an art rather than an exact science. One idea is to setup priors that remain invariant under re-parameterization. This is Jeffrey's prior, and it is objective because it is based on a math formula. But I wish to propose another method of convenience prior setting: select a mixture of subjective priors representing the different possible options (ex. no practical effect, small practical effect, medium practical effect, large practical effect). Then let the sample data "select" which prior in the mix is most appropriate. Using classical analysis as a baseline, I used simulation methods to assess the performance of Bayesian *t*-test with different priors. The results show mixed priors worked well in significance testing.

In chapter 4, I provide a practical demonstration of the usefulness of Bayesian analysis in special education. I sought to quantify the effect of different grouping methods on annual reading outcomes. Specifically, I tested whether enacting inclusion in 5th grade elementary classrooms using a natural proportioning method rather than a clustering one yielded a difference in annual reading gain scores. But, I only had a tiny judgment sample at my disposal and my groups were balanced with propensity score matching. To preserve the logic then of my statistical inferences, I resorted to Bayesian methods.

But, even if the requisite sampling conditions had been met for classical analysis, I still might have preferred Bayesian. Given tiny sample sizes it is extraordinarily difficult to definitely eliminate the null hypotheses using classical *p*-values. Refuting the null hypothesis is just too lofty a goal in such cases. Instead, a more realistic goal is assessing how probable it is given the sample. We may discover it is less probable than other hypotheses, even if we cannot eliminate it altogether from our hypothesis competition. This later feat, however, is naturally accomplished using Bayesian analysis.

In chapter 5, I summarize the findings of all three chapters. I return to my opening question: Is it ever advisable to let experts influence our statistical analysis? The purpose of this dissertation was to investigate this possibility in more depth. I argue in this chapter that my findings substantiate my original hunch, but more work still needs to be done in this area. As enticing as undergoing a Bayesian revolution in special education might be to me, I acknowledge it is also a risky proposition. One troublesome worry is that it makes the knowledge base reliant on the "good sense" of researchers who setup priors. To address this concern, I argue that all researchers, including classically-minded ones, must rely on their "good sense" too in statistical analysis – it is an inescapable problem.

Definition of Key Terms

Bayesian Statistical paradigm: In essence, statistical inferences modify prevailing credence-type probabilities given objective sample information as Bayes' rule prescribes. Methodologists who subscribe to this system usually reference the pioneering statistical work of Frank Ramsey, Harold Jeffreys, and Bruno de Finetti when developing their own statistical methods (Hacking, 2001; Howson & Urbach, 2006).

Classical Statistical paradigm: In essence, statistical inference should preserve the objectivity of chance-type probability. Methodologists who subscribe to this system usually reference the pioneering statistical work of figures such as Karl Pearson, R.A. Fisher, and Jerzy Neyman when developing their own statistical methods (Hacking, 2001; Lehmann, 2011).

Methodology: Normative account of the protocols researchers should use when carrying out research in theory (Paul, 2005).

Methods: Descriptive account of protocols researchers do use when carrying out research in actual practice (Paul, 2005).

Monte Carlo Simulation Techniques: Finding the chance-type probability of some event by observing the outcome of many probability experiments (Ware, Ferron, & Miller, 2013).

Qualitative Knowledge: In the usual statistical parlance, it is knowledge about qualities rather than quantities. Qualities are all about essences (types, categories) of things and their degrees of excellences (virtues, traits) (Bhaskar, 1979).

Philosophy of Inquiry: The philosophical investigation into a constellation of problems surrounding research methodology (Yu, 2006).

Pragmatism in Educational Research: The claim that "truth" is not a viable criterion for theory selection in educational research. Consequently, we use methods or procedures of research to "settle," "fix," or "stabilize" our beliefs about education to improve human happiness, democracy, or some other useful goal (Danforth, 2006).

Realism in Educational Research: The claim that a legitimate aim of educational research is to *understand* causal mechanisms operating in the world that influence education. Moreover, educational research is (or can be) successful at identifying true causal mechanisms, even if the "perfect" theory is beyond grasp (House, 1991).

Romantic Science: The bold thesis teleological explanations belong in science, and scientists cannot reduce key qualities to quantities without impoverishing their own disciplines (Plantinga, 1993; Nagel, 2012; Sider, 2011).

CHAPTER TWO

TWO VERSIONS OF STATISTICAL INFERENCE

There is no universal theory of statistical inference (Hacking, 2001), diverse paradigms underlay our statistical imaginations. The term paradigm is much overused today. But, stripped of the excesses of Kuhn's (1962) grandiose theory of science, paradigm becomes an apt choice of word. Before Kuhn, paradigms were heuristic exemplars. When learning Latin by rote, a person first learns to conjugate amare (to love) as amo, amas, amat… and then follows this familiar pattern, called *the paradigm*, to conjugate unfamiliar verbs of the same class. Likewise, renowned applications of statistics are held up as paradigms by subsequent analysts.

There are two key paradigms for statistical inference in education research, classical and Bayesian. Each offers its own version of inference. Those all too familiar *p*-values belong to the classical paradigm, but not Bayesian. Classical inferences require that studies include a chance setup somewhere in their design – ex., a subset of the population selected using some physical randomization device. A chance setup is a physically occurring probability experiment, such as the roll of a die. The Bayesian paradigm, in contrast, does not mandate chance setups.

Probability samples are often infeasible in education. Youth are vulnerable and school districts protective. Youth are also often scattered across countless locations. To compensate for this extra complexity, probability sampling has evolved into different species, including "simple random sampling," "stratified random sampling," and "cluster random sampling." This plurality helps researchers a little bit but bureaucratic headaches, prohibitive costs, and logistic nightmares still prevent probability sampling from becoming routine in education science.

To illustrate, consider the case of a survey researcher who wants to investigate whether caretakers sending their kids to a certain school perceive themselves as "insiders" or "outsiders" to its school community. She develops a fancy questionnaire to accomplish this feat. But she has no power to force caretakers, randomly selected off some master list, to complete it. Instead, she recruits volunteers. Suppose her advertisement campaign pays off, and she recruits a sizable sample. Even so, she does not have a true probability sample.

In light of the rarity of true chance setups in study designs in education research, I find it a bit odd classical inference is so prevalent in education research and Bayesian inference so rare. The purpose of this paper is to argue for reversing this. When chance setups are available, there is warrant for using classical inference. Outside these special cases, however, its logic is mostly inadmissible. Bayesian inference, in contrast, proceeds without hiccup.

A secondary purpose of this paper is to familiarize education researchers with some of the important philosophical contrasts between classical and Bayesian statistical inference. I suspect unfamiliarity among education researchers with this difference partly explains why Bayesian inference is underutilized in education research (Muthen, 2013). The debate between Bayesian and classical statisticians, after all, is philosophical in its nature (They actually agree on all the mathematics).

The structure of this paper is divided into three parts. In Part I, I provide a synopsis of classical and Bayesian versions of inference. In Part II, I critically compare their approaches to extrapolating from samples to populations. In Part III, I go on to consider inferences to causation. In all these cases, I find the logic of Bayesian induction is more generally applicable in education research than its classical counterpart because of the absence of chance setups in study designs. In my closing remarks, I discuss the implications of this for education science.

Part 1

Two Flavors of Objectivity

Humans have the extraordinary capacity to think objectively about reality. Objectivity is unbiased subjectivity, and there are multiple ways of establishing it. Methodologists argue about how best to accomplish this feat in science. Roughly speaking, methodologists are Bayesian or classical in their statistical philosophy. I elaborate on these two philosophies below.

The Classical Philosophy of Statistical Inference

Objectivity is prized among scientists, and naturally they want their statistical inferences to exhibit this virtue too. Politicians also value objectivity. They have a much easier time, after all, selling their campaign platform when objective research supports it. Politicians need this perception of objectivity to sound democratic, and non-elitist. Classical statisticians use chance-type probability to make their statistical inferences objective.

Recall, chance-type probability is about relative frequencies in the long run; credence-type probability, quantified uncertainty. Chance is highly objective. Two people flip a fair coin countless times. In the limits, they each observe a balanced proportion of heads and tails. To inject this type of objectivity into inductions, however, classical statisticians need probability samples (i.e. every member in a population must have some non-zero chance of selection) or a sample closely approximating a probability sample.

With a little metaphysical imagination, envision using a physical randomization devise, such as a spinner, to repeatedly select samples from the same population. Then record all results in a data set. In the limits, this data set will become the sampling density. Sampling densities, like random samples, are then all about objective chance – the chances scientists will observe this or that random sample in a given study.

Statistics computed from probability samples have sampling distributions too. Using the central limit theorem, for example, statisticians can easily prove means ($\bar{X}s$) computed from random samples of a certain size (*N*) have normal sampling distributions with a mean equal to the population mean and a variance proportional to the population variance. Armed with this knowledge, classical statisticians can then proceed to make all kinds of substantive inferences about unknown population means from known sample means. This is an impressive triumph, but it is not the only version of statistical inference around.

The Bayesian Philosophy of Statistical Inference

Bayesians inspect posterior probabilities – not sampling distributions – to draw their inferences about statistical hypotheses of interest. They assign posterior probabilities in fact to all hypotheses in the competition. Researchers, for example, may conjecture all sorts of possibilities for a population mean μ_X. Some guesses will be more probable than others given sample data. Bayesians can represent (not measure) such judgments with a normal posterior density $\mu_X \sim N(\mu_0, \sigma_0^2)$. If so, the hypothesis with most density (i.e., μ_0) represents the most credible value to scientists given the sample. The variance σ_0^2 of the posterior represents their uncertainty: larger variances, greater uncertainties.

Credence-type probabilities are subjective, and posteriors consist of nothing but credence. To compute a posterior probability for a hypothesis, Bayesians must specify a prior probability for it. Prior probabilities represent scientists' (informed) judgments about hypotheses probability before they even have an opportunity to examine the sample evidence. But scientists may differ in their judgments about which hypotheses are probable, and by how much. Reasonable people, after all, sometimes disagree. Bayesians, thus, must think carefully about what prior to use in their analysis: Garbage in, garbage out.

Classical statisticians worry priors might smuggle bias into inferences. Bayesians might settle on priors slanted towards their favorite hypotheses, and thus contaminate their posteriors. Without the constraints of samples, Bayesians must setup priors that seem sensible to them. But instead of sensible priors Bayesian might be selecting priors that reinforce their own prejudices about reality, and no one would ever know it.

The possible effect of an ill-chosen prior is concerning, but Bayesians are risk-takers. They risk some bias to gain the rewards of using priors. Bayesians find the risk acceptable since they can eventually filter out such bias from their posterior probabilities. Bayes rule specifies how to modify prior probabilities in light of objective sample evidence. Using it, Bayesians will, under normal conditions at any rate, eliminate bias in posteriors.

To illustrate this process, imagine two open-minded researchers hold contrasting opinions. One thinks it highly probable phonics instruction is better than whole language. The other thinks this claim is improbable. Notwithstanding the sizable gap in their priors, if each researcher uses Bayes Theorem to update their priors with sample evidence, they will eventually settle on the same posterior given a growing body of evidence. Mathematical proofs guarantee convergence in the limits, and mathematical proofs are pretty objective.

There are now two density-types to keep track of when considering inferences. All Bayesian inference depends on posterior densities; classical inferences on sampling densities. Sampling densities consist of constant parameters and varying data possibilities. They are all about objective chance. Posterior densities, by way of contrast, consist of constant data and varying parameter possibilities. They are all about subjective credence. Classical statisticians eschew credence-type probability. Bayesian thinkers, however, freely invoke both probability types as they see fit. This is a philosophical rather than mathematical argument.

Part II

Scientific Inferences to Populations

In part II, I move on to consider applications of classical and Bayesian inference in education research. Researchers often must extrapolate from samples to populations. Both paradigms promise objectivity to their users, but in this section we see classical inference only delivers on its promise in cases involving true probability samples. Bayesian inference, in contrast, can keep its promise with or without probability samples.

The Power of Probability Sampling

Imagine education researchers, for the sake of illustration, are interested in inducing the population mean μ_X from a sample. Classical statisticians usually want their inferences about μ_X to carry the objectivity of chance-type probability – i.e., researchers in the same circumstances always draw the same inferences. This is a high standard for objectivity, but classical statisticians found a way to make this bold dream obtainable.

To facilitate the whole induction process, classical statisticians use point estimates, interval estimates, and hypothesis tests. When studying a population with an unknown mean μ_X, they may produce a point estimate of μ_X. A point estimate ($\hat{\mu}_X$) is a single guess at the unknown parameter. An interval estimate, in contrast, gives an upper and lower-bound for μ_X,. Hypothesis tests, finally, help determine whether a value, such as $\mu_X = 50$, is tenable or not.

In the 1930s, the classical statistician, R.A. Fisher, neatly cast point estimates, interval estimates, and hypothesis tests alike within the chance-type probability mold. To accomplish this feat, he first used a physical randomization devise, such as a table of random numbers, to select samples from populations. He then capitalized on this chance setup in the design to compute all three types of inferences within the classical system.

Classical statisticians, like Fisher, evaluate possible point estimation methods, in the final analysis, using chance-type probabilities. There are multiple classical estimators (maximum likelihood estimation, method of moments, ordinary least squares), and they do not always produce the same estimate. Consistent with their philosophy, classical statisticians thus choose estimators based on their frequency properties, such as being "unbiased" in the limits. But chance-type probability presumes random samples and sampling distributions.

Two classical statisticians using the same methods in parallel research projects will reach the same point estimates – granting, of course, they observed identical samples. But probability samples gathered from the same population using the same protocols are expected to vary from one another. Differences accrue because of sampling error. Fisher figured out how to turn even this oddity into an asset in the classical system.

Researchers often carry out investigations hoping to find something worthwhile to report. "Worthwhile" may be, and in education research often is, just a new line of inquiry to explore. Fisher suggested using *p*-values to determine if samples were indeed significant or not (Significant data only entails this area of research is probably capable of repaying our efforts). To compute a *p*-value, one must first setup a null hypothesis (H_0). The null hypothesis expresses a situation where there is nothing of substantive value in this area of inquiry.

We might know, for example, youth on average score a 50 on some test of interest. There is nothing for researchers to explain then if a subgroup of this population also averages 50. But researchers would have something to query if a subgroup of interest did significantly better or worse than the overall population. They would need to explain, for example, why this subgroup differs from everyone else. Given this setup, our null hypothesis should be the subgroup average matches the overall population average, $H_0: \mu_X = 50$.

After successfully setting up the null, Fisher next computes the sampling density for the test statistic given the null. He carves up the sampling density into significant and insignificant regions. Conventionally, he demarcated these regions so that the selected 5% region of sampling density is furthest away from the most likely data possibility under the null. In the case of the study involving the population mean, if we grant a population variance $\sigma_X^2 = 10$ and a sample size *n*=10, we designate any $\bar{X}$ between 51.96 and 48.04 insignificant (otherwise, significant). This is because only $\bar{X}$s falling outside of this region produce *p*-values small enough to make them significant. If we find $\bar{X} = 60$ in our study, then we simply compute the probability of getting $\bar{X} = 60$ (or a weirder sample mean) given H_0: $\mu_X = 50$ to find the *p*-value.

Interestingly, Fisher never gave a rationale for picking the 5% region furthest away from the highest density of the sampling density. Technically, using 5% of any part of this density should keep the significance level at 5%. It seemed only too obvious to Fisher, however, that it made much more sense to settle on the region furthest away. Fortunately, Neyman and Pearson later supplied Fisher's intuitions with a justification using the likelihood ratio principle - a principle fully compatible with classical philosophy.

Fisher considered the customary 5% significance level to be convention; a matter of clinical judgment not an objective formula. This surprising intellectual move seems more appropriate for Bayesians. Again, Neyman and Pearson came to the rescue. They placed statistical inference into a decision framework, and sought to pick significance levels in such a way so as to balance Type I (α) and Type II (β) error rates. For his part, Fisher never accepted their "solution" as he claimed inferences from samples to population are about inductions rather than decisions – but this family feud among classical statisticians need not detain us.

Neyman also invented confidence intervals. Using knowledge of the sampling density, Neyman figured out how to produce a 100%(1-α) confidence interval. Given researchers computed $\bar{X} = 60$ from a random sample (n=10) and $\sigma_{\bar{X}}^2 = 10$, researchers can say using Neyman's method they are 95% confident the true parameter is anywhere between 58.4 and 61.96. Their confidence here rests on the grounds that the procedure captures the true population mean 95% of the time in repeated applications.

The (mistaken) claim that there is a .95 chance-type probability the true population parameter is between 58.4 and 61.96 is a common misinterpretation of the confidence interval. It betrays the classical insistence on frequency in the long run. The true parameter either is or is not in the interval, this is not a matter of chance. Classical statisticians make no claims then about the chances the parameter is in this or that interval. Their .95 confidence rests on the procedure's .95 coverage rate in repeated applications in the limits.

The masterstroke behind all classical inference from samples to populations is random selection. When they use a physical randomization devise to select samples from populations they can invoke chance-type probability. Everything hinges on this. Classical statisticians often compute sampling densities with fancy analytics. Using the central limit theorem, for example, they proved the sampling distribution of $\bar{X}$s is normal.

Sometimes, however, it is not possible to analytically compute sampling distributions for statistics of interests, such as medians (m). Simulation methods, such as bootstrapping, provide classical statisticians with an empirical way of finding the right sampling distributions in these special cases. Simulation methods are computer-intensive, but they are lifesavers. Statisticians can use them to construct sampling distributions for any statistic of interest provided they know what probability sampling technique to simulate when 'drawing' samples.

The Scope and Limits of Classical Inference

Within its natural boundaries, the logic of classical inference survives all critical scrutiny. Its version of objectivity surpasses the Bayesian version in protecting inferences from bias. But outside its proper boundaries, it loses its entire luster. Without probability samples, its inferences fail to deliver on their promises of objectivity. They become wholly arbitrary, and so unfit for a science of education.

Someone might object, however, that I am being too rash. They might counter me by saying something like classical statisticians can treat non-probability samples as if they were random. They could then compute *p*-values, and conclude, 'Hypothetically speaking, these data are significant at the 5% level.' This information might function as a measure of practical effect, and so inform the knowledge base. And all classical statisticians would need to do is alert the reader to their "pretense" in the limitation sections.

I am not satisfied, however, with the above reasoning. Without probability sampling, the choice of sampling distribution is too underdetermined to be an adequate measure of practical effect. One classical statistician might come along and impute one sampling distribution for the sample and then another classical statistician might come along and impute a different one. They will then reach different effect sizes, but there would be no matter of fact about which imputation choice was correct. The math works equally well in both cases.

To illustrate, consider the case of a drunken man who claims to his friend he can usually predict the outcome of the toss of a fair coin. The friend sets up an experiment to test his claim. In 9 out of 12 trials, the drunk succeeded in predicting whether it would land heads or tails. The friend then asks two statisticians to independently determine if the data are significant or not, but she forgot to specify why she halted the experiment after 12 trials.

The stopping rule matters. Without this information, statisticians wishing to conduct classical analysis must guess it to select an appropriate sampling distribution. The first statistician, therefore, guessed the experimenter planned from the outset to stop after 12 trials. She, therefore, selects the binomial sampling distribution. The second statistician instead guesses the experimenter planned from the outset to stop after observing 9 correct predictions. It just happened to take 12 trials to reach 9 correct predictions. She, thus, selects the negative binomial sampling distribution rather than the binomial.

Each classical statisticians setup the same null hypothesis, $H_0: \theta = .5$. This represents the case where the drunk is randomly guessing. Using the binomial sampling distribution, the first statistician computes a one-tailed *p*-value of .075. She declares the data to be insignificant at the 5% level, and recommends no further investigation. Using the negative binomial sampling distribution, the second statistician computes a one-tailed *p*-value of .035. She declares the data to be significant at the 5% level, and recommends future investigation.

The friend is confused. But, unless she supplies the true reason she stopped at 12 trials, both analyses stand on equal mathematical footing. But suppose she clarifies that she stopped at 12 trials because the drunk passed out on the 13th trial, and she was forced to stop. Technically, both imputations were wrong. If the experiment was repeated again the drunk might pass out earlier or later, and so this warrants a different sampling distribution altogether.

If we assume time to passing out randomly varies from experiment to experiment, then the classical statisticians might rethink their analysis plans. They might use bootstrapping to simulate the requisite sampling distribution. They will then have a rationale for why they used this sampling distribution in the analysis. They know it models what actually happened in the experiment. But this all entails a known element of chance in the sampling process.

The objectivity of classical inference, thus, hinges on knowledge of the stopping rule. Without it, classical inference is vulnerable to bias. The classical statistician who hates the idea of a drunk predicting the tosses of fair coin might select the binomial sampling distribution to reinforce their bias. The classical statistician who loves the idea might instead pick the negative binomial sampling distribution. Classical inference no longer safeguards against bias.

One might try to salvage the situation by making up a new convention. Always treat a non-probability sample as if it were a certain type of probability sample. Perhaps, treat them as ones drawn with a fixed number in mind from the outset. This is in fact what is happening in a lot of education research. But this does not really solve the deeper problem: The whole point of doing classical inference was to prevent such subjectivity (Kruschke, 2011).

Bayesian Inferences without Probability Samples

All classical inferences are drawn from sampling distributions. Sampling distributions grant classical inference the objectivity of chance-type probability, but make them vulnerable to the so-called stopping rule problem. The stopping rule problem is overcome when we know what physical random mechanism produced the data. Otherwise, it is intractable. The stopping rule problem, thus, bounds classical inference to the realm of true probability samples.

Bayesian inferences are drawn from posterior distributions – not sampling. The formula for sampling distributions requires stopping rules; the formula for posterior distributions does not. Bayesians only care about the visible data, and Bayesians will reach the same conclusion regardless of what stopping rule may or may not have been used to gather the sample. Thus, Bayesian inference is immune to the stopping rule problem. They will reach the same inference given the visible data, regardless of whether it was really a binomial or negative binomial sampling distribution.

Bayesians, however, can disagree with one another. They need to specify priors, and if they specify different priors they may compute posteriors entailing opposite conclusions. This is troublesome to classical statisticians, because bad priors might bias inferences. This, however, is not the Bayesian version of the stopping rule problem. The stopping rule problem without true random samples is intractable within the classical system. The problem of priors biasing inferences is solvable within the Bayesian system.

Imagine the friend had instead asked two Bayesians to analyze her experimental data: nine successes, three failures. It is likely both Bayesians would be skeptical of the claim, and their priors would equally represent skeptical stances. Such a skeptical stance is consistent with background knowledge, and it seems unwise to exclude such background knowledge. The prior specifications need not be exactly identical to each other to produce a comparable posterior so long as both priors reasonably represent a skeptical opinion. In this case, both Bayesians should produce slightly different, but still very, skeptical posteriors.

No formal mechanism in Bayesian analysis, of course, guarantees two Bayesians will in fact set priors encapsulating skeptical stances. One Bayesian might let her wishful thinking cloud her judgment, and so specify a credulous prior – i.e. a prior biased towards the drunken man's claim. The other Bayesian may show more caution, and specify a more sensible (skeptical) prior. Each will then proceed with the analysis, but arrive at opposite conclusions.

In this case, the friend might decide to collect more evidence to resolve the difference. She might redo the experiment, and share her findings again with both Bayesians. They will then update their opinions. They will use their old posteriors as their new priors, and compute a new posterior with the new evidence. Overtime, their two posteriors will converge as the influence of their priors is washed away by a growing body of evidence.

Judgment Samples

In the case of samples, Bayesians have multiple options on the table. They can use probability or judgment samples. Classical statisticians are limited to probability samples. Probability samples are about chance-type probabilities. Judgment samples, in contrast, are about credence-type probability. We select samples that we (or experts) judge to represent the population of interest (Howson & Urbach, 2006).

The stopping rule problem prevents classical statisticians from capitalizing on judgment samples. Bayesians, however, have no such restrictions. They can embrace any sample they deem to be representative of the population. All that matters are the visible data, and they do not care about how they were selected so long as they are representative. Judgment samples, within the Bayesian system, then are viable alternatives to probability samples.

The quota sample is a well-known instance of judgment sampling. It is the credence-type analog of stratified random sampling. This sampling technique involves giving interviewers quotas of people to interview. Quotas are deliberately picked to ensure the sample represents the population of interest in key ways. The interviewers are then instructed to fill up their quotas using their best judgment. It is already being utilized in marketing research to inform high-stakes decisions.

Imagine researchers have access to a visible sample. It could be either a quota sample or a stratified random sample. Bayesians do not care. The relevant information to them is in the visible data themselves Classical statisticians, in contrast, need to know. They can proceed with objective inferences only if it was a stratified random sample. If it was a quota sample then they must either pretend like it was really a stratified random sample and engage in a bit of fiction or – given the stopping rule problem – forgo analysis altogether.

Summary

The argument of this section is that the logic of classical induction is inadmissible in the realm of non-probability samples. The warrant for classical inferences comes from their high objectivity. To achieve this objectivity, however, classical statisticians are constrained to probability samples. Bayesian inference gets its objectivity from an epistemic rule, Bayes' Theorem. The Bayesian version of objectivity lacks some of the luster of classical inference, but it imposes no constraints on sample type.

Part III

Scientific Inference to Causation

Classical and Bayesian extrapolations from samples to populations are both valid. But Bayesian methods are more generally applicable in education than classical ones. I argue something similar happens in the case of inferences to causation. Such inference occurs in intervention studies, and for ill or good they hold a high status in education science.

Intervention Studies

All intervention studies involve a manipulable variable, but only those with adequate controls count as experimental. Otherwise, they are *quasi*-experimental. The prefix "quasi" here intimates an epistemic status of second best. Researchers settle for them, wishing for more. But what constitutes "adequate" controls?

Sometimes experimental units function as their own controls. In single case design, for example, researchers observe the same individual under both control and intervention conditions, and draw comparisons. They monitor rates of misbehavior among students to establish baselines. They then introduce behavior modification technology into the environment and observe changes in behaviors. This design establishes adequate control.

It is not always possible, of course, to use experimental units as their own controls. To compare the effect of two math curriculums on children, for example, researchers cannot simply make children take both curriculums consecutively. They want to know what happens if children are given either this or that curriculum, not both. They have no choice then but to use two different groups of students.

But groups had better be comparable on extraneous variables to warrant inferences to causation. Researchers, however, are never certain they controlled for all extraneous variables. Perhaps, they omitted an unknown one. But certainty is too high a standard, they can settle for probable (i.e. the groups are probably matched on all extraneous variables). This is as good as it gets in education science. We are satisfied then with a research design making the comparability of groups more probable than not.

Selection and Assignment

To probe the topic further, methodologists routinely distinguish between two types of procedures, selection and assignment. The selection-type entails sampling from a population. Take a list of the whole population, select a subset of it. Its function is to produce samples representative of the population of interest. The assignment-type, in contrast, divides existing sample members into subgroups. Its function is to produce balanced groups – or more precisely, groups probably matched on extraneous variables in the requisite way.

Classical statisticians need an assignment procedure compatible with their system. To accomplish this, they invoke random assignment – i.e., using a physical random devise, or computer simulation of one, to divide sample members into subgroups. But it is a misconception to think random selection is now moot. Cogent inferences to causation in classical analysis, I argue, usually need both random assignment and selection working together in concert.

The Logic of Classical Inference to Causation

Random selection seems like it should be moot. Assignment, after all, is what makes groups balanced – not selection. But, as we shall see, this is not exactly so. Statisticians using parametric statistics, such as *t*-test, need both random assignment and selection in their designs to establish causation (Recall, parametric entails interest in invisible population parameters). This proviso for random selection may be unfamiliar, and so I explain why.

To illustrate, suppose a novel reading intervention raises students' test scores by a lot. To discover this effect, classical statisticians must consider two populations of test scores. One consists of scores of all students *with* the intervention (intervention population); the other scores of students *without* it (control population). If classical statisticians had physical access to these populations, they could draw a random sample from each, and test for mean difference. Perhaps, using the independent means *t*-test, for example, to establish causation.

Classical statisticians, however, never have physical access to both populations. Instead, they use random assignment to compensate. If they can physically draw a random sample from the control population – i.e. students without exposure to intervention - then they can randomly assign all members of this big sample into two smaller subgroups. These subgroups simulate two random samples drawn from the control population on distinct occasions. Each subgroup, therefore, probably resembles the control population.

Classical statisticians can then physically transform one of these two groups into a random sample from the intervention population, by subjecting it to intervention conditions. Together random assignment and selection, thus, give researchers something akin to random samples drawn from both control and intervention populations. This circumstance is all statisticians need to classically induce causation.

This procedure illustrates two admirable virtues of random assignment: it (a) controls known and unknown extraneous variables and (b) preserves the high objectivity of classical inference. But why does the classical parametric *t*-test need both random assignment and random selection in the design to deliver on its promise of objectivity? To answer this question, it is worthwhile to investigate the above two claims in a little bit more depth.

Controlling All Extraneous Variables at Once

Let us consider, therefore, the claim random assignment handles both types of extraneous variables, known and unknown. To understand why one must recall that randomly dividing one random sample into subgroups produces two smaller random samples. But there is a chance both samples will resemble their parent population, and this chance increases with sample size. There is, thus, a chance groups are comparable on all extraneous variables. We never know the exact chance, but we know this chance will increase as sample size rises.

Without the special combination of random selection and assignment, however, chance calculations are inadmissible. Random assignment only takes advantage of operating chance. It cannot produce it. If one randomly divides a non-random sample into subgroups, there is no chance groups will resemble their parent population. They either do or do not. Chance does not enter into this picture in the requisite way.

In real applications, classical statisticians can check group balance on visible extraneous variables to see if random assignment did its job. Perhaps, gender is an extraneous variable. Girls may have more natural aptitude than boys for example. Statisticians can then use a logistic regression to determine if groups are significantly imbalanced. If they are indeed so imbalanced, they can (a) redo randomization till groups become balanced, (b) physically control for it, or (c) statistically control for it.

Preserving the Logic of Classical Inference

The need to do significance testing to check group balance then brings us to the second claim made in support of random assignment – i.e. it underpins the logic of significance testing. It is by far, the strongest reason for classical statisticians to use it. But, obviously, it holds no weight among Bayesians who have a different inference scheme. Let us move on, therefore, to consider the claim that random assignment keeps classical systems intact.

Without random selection in experimental designs the physical operation of chance-type probability in parametric versions of statistical inference becomes a fiction. Classical statisticians can only conduct a parametric independent means group *t*-test to evaluate null hypothesis by treating each group like a random sample drawn from the same population. Without random selection, however, this premise is not defensible – i.e., the two groups do not really behave like random samples from populations.

Random assignment alone – at least, in parametric *t*-test contexts – undermines the choice of sampling density. This is because of the stopping rule problem mentioned in the previous section. Recall, the formula for the *p*-value depends on the sampling density, and different specifications of it can shift the final *p*-value in incompatible directions. Classical statisticians, thus, need guidance from a physical random selection mechanism to construct the requisite sampling densities for control and intervention groups.

The above criticism, of course, only applies to "parametric" versions of classical statistics (ex. *t*-tests, ANOVAs, etc). One can instead use "non-parametric" statistics to infer causation with just random assignment (Edgington & Onghena, 2007). But there is a catch: one must then delimit conversation about causation to samples – and only samplea. But random assignment, even divorced from probability sampling, is still frequently infeasible in education.

Bayesian Analysis and Propensity Score Matching

Random assignment is difficult to implement in education research. We cannot usually, for example, randomly assign one group of youth to college and forbid another group from going to college only to measure the effect of college on life satisfaction. This is unethical, and it would never pass muster with an IRB board. But the classical proviso for random assignment and random selection to induce causation in populations makes the task epic.

Education researchers, therefore, often resort to propensity score matching (hereafter, PSM) to make observed groups comparable. PSM involves credence-type probability – not chance-type. This is because researchers must exercise their own judgment about what covariates are probably (credence-type) important, and so ought to be included in the match up process. But PSM nicely safeguards against researcher introducing selection-bias into studies.

Consider this skeptical hypothesis: imbalance between groups on an extraneous variable really explains experimental findings – not intervention effects. Bayesians can assure us that as the prior for this skeptical hypothesis shrinks the posterior for the causation hypothesis rises. The Bayesian, consequently, use PMS to shrink the prior for the skeptical hypothesis and raise the posterior for the causation hypothesis, and this is their rationale for PSM. Methodologists often refer to this as using historical controls rather than random controls.

Classical statisticians deploying PSM to balance groups, like Bayesians, inadvertently rely on credence-type probability. They find this arrangement suboptimal, however, demoting it to a non-experimental status. On the other hand, Bayesians are not embarrassed by credence-type probability at all, and can adopt a more welcoming posture. They may award an experimental status to observational studies using adequate historical controls (Howson & Urbach, 2006). Bayesians, of course, decide what counts as ‘adequate’ on a case-by-case basis.

Conclusion

I do not fault classical inference because it is not Bayesian any more than I fault Bayesian inference because it is not classical. They are apples and oranges. But they have different logics, and we are wise to respect their natural differences. I claim we can use whatever version of statistical inference we prefer, so long as we remain faithful to its inner logic.

There is elegance in classical inferences. Their objectivity prevents (or else minimizes) a certain type of bias. But the price it pays for this objectivity is some inflexibility. It is only applicable in study designs involving chance setups. Bayesian inference does not have the objectivity of classical inference. But it does have another version of objectivity, and we need to settle for it when chance setups in designs are not available.

The conclusion then is not that Bayesian inference is superior to classical inference. The conclusion here is much more modest and defensible than that. My conclusion is that Bayesian inference is generally more applicable in education research than classical inference. But we can still do classical inference in education whenever the requisite sampling conditions prevail.

Discussion

But what should we do with all the education research out there that used classical analysis on judgment samples to inform their conclusions? One possibility is to re-evaluate the *p*-values as measures of practical effect. And, as a measure of practical effect, it would be given a pragmatic justification rather than underwritten by a logical principle. But reporting *p*-values as effect sizes is bound to create confusion. Notwithstanding, it hardly seems like a viable solution to just throw away education research with non-probability samples merely because it reports *p*-values. That is too drastic a step. I, consequently, argue that a much more sensible solution to this problem is to charitably impute Bayesian logic onto these studies.

To make this translation, all we need to do is note that the difference between classical confidence intervals and Bayesian credible intervals is probability type. But one of the biggest problems in applying Bayesian inference in education research is setting up appropriate priors. It is easy to setup priors when there is an abundance of background knowledge. But in contexts of pioneering research there is only minimal background knowledge.

One recommendation for setting up priors with minimal background knowledge is to setup objective priors. Certain kinds of objective priors produce posteriors with inferences for parameters of interest with classical frequency properties. We can then compute Bayesian intervals with both a chance-type and credence-type probability interpretation. So, we can interpret a classical 95% confidence interval the way we would interpret a 95% Bayesian credible interval – i.e., we are 95% sure the true parameter is in this interval given the sample and our prior probabilities. We can, thus, interpret education research using classical parametric analysis on non-probability samples as if they had used really used Bayesian analysis with objective priors.

This is a sensible policy. It matches up existing practice and sound philosophy. As things stand now, much statistical practice in education research is out of accordance with its exposed classical philosophy. But classical philosophy is not the only game in town. The Bayesian system provides a worthy alternative, and I think that as more education researchers become familiar and comfortable with it, they will start explicitly espousing it too. But I certainly do not claim to know the future, and I can only guess.

CHAPTER THREE

CONVENIENCE PRIORS FOR BAYESIAN *T*-TESTS

Scientists love experiments. In education science, it is their wont to compare the average outcome of control and experimental groups on some outcome of interest. If it is a continuous variable, they invoke the celebrated independent means *t*-test to substantiate causation. This test formally decides whether group means significantly differ or not. Significance entails causation is a plausible conjecture (i.e., the null hypothesis is untenable).

There are two versions of the independent means *t*-test, classical and Bayesian. The inner logic of the classical version of this parametric test requires both random assignment and random selection (Chapter Two). Randomization for Bayesians, however, is not even held up as the *Golden Standard* (Howson & Urbach, 2006). This flexibility in study design makes Bayesian *t*-tests more generally applicable in social science. But Bayesian *t*-tests carry their own fair share of risks, and care needs to be taken when using them.

The centerpiece of all Bayesian statistics is a one-line theorem of the probability calculus, Bayes rule. To Bayesian epistemologists, it provides an elegant framework for induction; briefly:

$$P(H|D) \propto \mathrm{P(H)P(D|H)}$$

The posterior density of hypothesis H given data D is proportional to the product of the prior density of H and the likelihood of the data given the hypothesis. This can also be expressed:

$$(posterior\ density) \propto\ (prior\ density) \times (liklihood)$$

The posterior density represents how credible the hypothesis is to us after adjusting our initial beliefs towards the hypothesis in light of sample information.

The success of the whole Bayesian enterprise depends on posteriors. Some posteriors are sensible, others absurd. Likelihood functions dominate posteriors as sample size rises, and this is reassuring to researchers. It means sample data influences posteriors more than their personal priors. But priors take on a prominent role in posterior construction as sample size shrinks. This makes priors critical in cases involving small samples.

Subjective priors represent judgments about credibility (Gill & Walker, 2005), and they are usually informed with background knowledge. We do not need perfect priors, but we do need sensible ones. And what counts as sensible is a matter of qualitative judgment. The rewards of subjective priors can be great, but so too their dangers (Baldwin & Fellingham, 2013). When consensus prevails, the prior's sensibility is obvious. But it is not always so.

Pioneering scientists cannot appeal to consensus to defend their prior. No such consensus exists. This raises the frightening specter of rogue scientists using their select "sensible" priors as they see fit. Priors, thus, become vehicles for injecting unwarranted bias into inference. This fear of abusing priors is perhaps the oldest criticism of Bayesian statistics (Fisher, 1930).

To prevent such abuses, Bayesians resort to convenience priors – i.e. priors customarily chosen as defaults in initial analyses (Christensen et al., 2011). Bayesians invented objective priors to function as defaults. Objective priors are determinable by formal rules and logical principals, such as *the principal of entropy* (Jaynes, 1968). But, even after much technical work, we have yet to arrive at a single route for picking objective priors (Kass & Wasserman, 1996). I, therefore, want to buck the general pull towards objective priors. Instead, I propose we investigate the possibility of using subjective priors as convenience priors. To this end, I conducted a Monte Carlo study to evaluate the performance of a Bayesian *t*-test with a subjective prior under different conditions with small sample sizes.

Theoretical Background

To facilitate the rest of our discussion it is helpful to differentiate among models, estimation theories, and computational algorithms (Raudenbush & Bryk, 2002). Models are equations specifying parameters – or constants of substantive interest – needing to be estimated from sample information. Estimation theories define the estimation process. Computational algorithms implement our favorite estimation theories. They are mechanical rules for getting output given input.

The Model

Social processes resemble natural lotteries, their outcomes probability densities. Normal densities are often useful for modeling continuous outcomes in education science, briefly:

$$Y_i \sim N(\mu, \sigma^2)$$

Where the outcome of the variable Y on the i^{th} participant is stochastically determined according to a normal density with mean and variance –i.e., μ and σ^2, respectively. Such variables often take on the normal-shape because the central limit theorem states any random variable that arises as the sum of a large number of independent, small effects will be normal.

Suppose we conduct a clinical trial to infer causation. We compare control group means, and discover the intervention group's mean is superior to the control's, $\overline{X}_T > \overline{X}_C$. This finding is interesting, but it only pertains to the sample. What we usually care about is whether there is a similiar difference in the population, $\mu_T > \mu_C$. Using statistical modeling techniques, such as regression, we draw probabilistic conclusions about true population mean difference given the sample information. Ideally, we want valid point estimate, probability intervals, and hypothesis tests. Classical and Bayesian inference offer alternative pathways to get all these types of inference procedures (Hacking, 2001).

Classical Estimation Theory

All classical inferences are based on sampling densities (Mohr, 1990). Sampling densities contain the chances of any sample possibility given a parent population (or lottery). To build a sampling density, we randomly sample from the same population again and again using the same protocol. The frequency each data possibility is observed and recorded in a data set. This data set becomes the sampling density in the limits.

In the classical scheme, the computed statistic, $\overline{X}_T - \overline{X}_C$, is the best point estimate for the true parameter quantity, $\mu_T - \mu_C$. It has the best chances of matching this parameter value in the sampling density. To assess whether this point estimate is statistically significant or not, it is conventional practice among scientists, when population variances are unknown quantities, to compute the *t*-statistic – i.e., the ratio of the observed difference to the standard error (i.e. the standard deviation of a sampling density):

$$t(\boldsymbol{y}_{\mathrm{T}}, \boldsymbol{y}_{\mathrm{C}}) = \left[(\overline{X}_{\mathrm{T}} - \overline{X}_{\mathrm{C}}) - (\mu_{\mathrm{T}} - \mu_{\mathrm{C}})\right] \Big/ S_{(\overline{X}_{\mathrm{T}} - \overline{X}_{\mathrm{C}})}$$

Where $t(\boldsymbol{y}_T, \boldsymbol{y}_C)$ is the t-statistic and $S_{(\overline{X}_T - \overline{X}_C)}$ is the standard error estimate. In practice, we use this ratio to rule-out sampling error as the cause of sample mean difference.

The requisite sampling density for the *t*-statistic is a *t*-density with certain degrees of freedom. If the null hypothesis is true we expect $t(\boldsymbol{y}_T, \boldsymbol{y}_C) = 0$. So, we check to see if the computed *t*-statistic value is close to this expectation. For mathematical precision, we use *p*-values to accomplish this feat – i.e. the chances of obtaining our actual *t*-statistic value or one weirder given the truth of the null hypothesis. If this *p*-value is too improbable given a stipulated cut-off criterion, such as $\alpha = .05$, we decide the null hypothesis is false (5% chance our decision errs given all our model assumptions).

Classical statisticians build chance-type probability intervals for the true population mean-difference with any α-level desired:

$$(\overline{X}_{\mathrm{T}} - \overline{X}_{\mathrm{C}}) \pm t_{2/\alpha}(S_{(\overline{X}_{\mathrm{T}} - \overline{X}_{\mathrm{C}})})$$

here *t* is a critical value with certain degrees of freedom. Any possible interval of values either contains the true parameter or it does not. There is no chance about it. So, we interpret a 95% chance-type probability interval as an interval generated by a mechanism known to produce intervals covering the parameter 95% of the time in the limits.

Classical Computational Algorithms. There are two principal computational algorithms for the *t*-statistic ratio, the pooled standard deviation and Satterthwaite method. Each has its unique advantages and disadvantages making them both indispensable tools in the classical statisticians' full armament.

Pooled Variance Method. When $\sigma_T = \sigma_C$, we pool information across sample groups to get an overall estimate of their singular quantity. The algorithm for implementing the pooled standard deviation method is:

$$S_p = ({}^{n_1}/_{(n_{\mathrm{T}} + n_{\mathrm{C}})})S_{\mathrm{T}} + ({}^{n_{\mathrm{C}}}/_{(n_{\mathrm{T}} + n_{\mathrm{C}})})S_{\mathrm{T}}$$

Where S_p is the pooled standard deviation and n_T and n_C are respective group sample sizes. Then we compute the requisite standard error for the *t*-test by using an inverse relationship between it and the ratio of sample group sizes and their parent populations' standard deviations:

$$S_{(\overline{X}_{\mathrm{T}} - \overline{X}_{\mathrm{C}})} = S_p\sqrt{{}^{1}/_{n_{\mathrm{T}}} + {}^{1}/_{n_{\mathrm{C}}}}$$

Setting $n_T + n_C - 2$ as the degrees of freedom, we select the needed *t*-shape for the sampling density. The pooled method produces a *t*-test more powerful than its sister Satterthwaite method (discussed next) at any of α-level desired (Ware, Ferron, & Miller, 2013).

Satterthwaite Method. When $\sigma_1 \neq \sigma_2$, we experience a loss of Type I control using the pooled method. To keep control of Type I error rate, therefore, we must use S_T and S_C as distinct estimates of their parent population. The *t*-density shape depends on the ratio of population standard deviations so it is now no longer straightforward what the degrees of freedom should be for *t*-testing purposes. In such cases, Satterthwaite recommends adjusting our degrees of freedom to build a robust *t*-test:

$$\frac{S_{\mathrm{T}}^2/n_{\mathrm{T}} + S_{\mathrm{C}}^2/n_{\mathrm{C}}}{\frac{(S_{\mathrm{T}}^1/n_{\mathrm{T}})}{(n_{\mathrm{T}}+1)} + \frac{(S_{\mathrm{C}}^2/n_{\mathrm{C}})}{n_{\mathrm{C}}+1}}$$

Satterthwaite corrections then account for heterogeneity between population variances. The pooled variances offer a more powerful test of mean difference, the Satterthwaite a more robust one. Both methods, however, equally preserve the inner logic of the classical *t*-test: Interpret all probability-statements in inferential statistics as being about the frequency of outcomes in the limits – not qualitative judgments of the credibility of hypotheses.

Bayesian Estimation Theory

Bayesian *t*-tests, in contrast to classical *t*-tests, are all about posterior densities. Posterior densities show how credible all distinct parameter possibilities are given sample data:

$$Y_{T1}, Y_{T2} \ldots Y_{Tn_1} \sim N(\mu_T, \sigma_T^2) \qquad Y_{C1}, Y_{C2} \ldots Y_{Cn_2} \sim N(\mu_C, \sigma_C^2) \qquad \mu_{\mathrm{T}} - \mu_{\mathrm{C}}$$

Bayesian analysis, unlike classical *t*-tests, focus on the posterior density of $\mu_T - \mu_C$ to evaluate population mean difference. Using this density-type, they draw point estimates, probability intervals, and hypothesis tests. In the Bayesian approach, we need not invoke *t*-test statistics and we rarely even resort to *t*-shaped densities. But since this procedure plays an analogous function in the Bayesian economy, we nonetheless will refer to it as a Bayesian *t*-test.

Classical statisticians choose $\overline{X}_T - \overline{X}_C$ as their point estimate of $\mu_T - \mu_C$. But given the posterior probability for all possible $\mu_T - \mu_C$ values, Bayesians instead pick the single value they find most credible given the objective sample and their personal priors. Probability intervals, drawn from posterior densities, also have nothing to do with coverage rates in the long run. There is a .95 credence-type probability that the true parameter is contained in a 95% credible interval, and they make no commitments whatsoever about its coverage in the limits. Likewise, the user of classical version of the *t*-test has to begin with the supposition the null hypothesis is true, $\mu_T - \mu_C$=0, in order to refute it later on, but the user of Bayesian *t*-tests can simply compute how probable rational agents will judge the null hypothesis to be given the objective sample and their personal priors (Kruschke, 2011).

Bayesian Computational Algorithms. Finding posterior densities posed serious problems for doing Bayesian analysis in the past as posterior integrals were often intractable. This partially explains why many applied researchers never learned about Bayesian analysis in their training. All this changed when statistical software packages, such as SAS, gave us Markov Chain Monte Carlo (MCMC) simulation methods for finding complex integrals. MCMC algorithms sample from posterior distributions, and compute quantities of interest.

In this simulation study, I used PROC MCMC (SAS 9.4) to derive posterior distributions. This procedure uses a Metropolis-Hastings Algorithm, and I used all the standard default settings in SAS to make replication of my study easier. I uniformly used a burn-in of 10,000 to cope with autocorrelation, sampled 100,000 cases, and thinned my posterior samples by a magnitude of 10 to reduce any autocorrelation. This left a total posterior sample size of 10,000. To set the initial values of chains for model parameters, I always started at the mode of its prior distribution. This procedure yielded reasonable results based on fit statistics in a pilot run.

Motivating Example. Imagine a researcher is conducting a Bayesian *t*-test on the outcome variable Y, wanting to detect traces of causation in a tiny data set. The studied control population has been researched in the past on multiple occasions, and it is known that $Y \sim N(50, 10^2)$ under control conditions. But next to nothing is known about what typically will happen to Y under intervention conditions. She is doing ground-breaking research. The researcher, thus, had better setup defensible priors to proceed with Bayesian analysis to estimate the effect. Below I elaborate on a three possible default priors she could use in this case.

Bayesian t-tests with Objective Priors. Jeffreys prior is the most celebrated convenience prior for Bayesian *t*-tests. To see why consider this: A choice on how to parameterize the normal density confronts Bayesian analysts:

$$Y_i \sim N(\mu, \sigma^2) \qquad Y_i \sim N(\mu, \sigma) \qquad Y_i \sim N(\mu, \tau)$$

Where μ is the mean, σ^2 is the variance, σ is the standard deviation, and τ is the precision. These three parameterizations have the same location, but different (albeit, legal) ways to parameterize the scale, σ^2, σ, and τ. We can translate between these schemes using one-to-one functions. The invariance-principle then states parameterization should not change conclusions.

Jeffreys priors produce posterior distributions that satisfy the invariance-principle. Using them, for example, we can draw conclusions about variances from standard deviation posteriors. Jeffreys prior is obtained as the square root of the determinant of the Fisher Information Matrix:

$$P(\theta) \propto \sqrt{detI(\theta)}$$

Specific to *t*-testing, Jeffreys prior densities become:

$$\mu \sim constant \qquad \sigma^2 \sim -log(\sigma^2)$$

These prior densities choices make logical sense because they obey the invariance-principle, but, like classical procedures, they are vulnerable to the stopping rule problem (Chapter 2).

Bayesian t-tests with Both Subjective and Objective Priors. In the above motivating example, our background knowledge only underdetermines subjective priors for the intervention parameters. She actually had a lot of information about the parameters belonging to the control population. It is obvious then what counts as appropriate subjective prior for those parameters. Naturally, when subjective priors encapsulate background knowledge they are called "informed," and even Fisher thought there was nothing controversial about using such subjective priors to 'inform' research so long as they were really obvious (Hacking, 2001).

In this case, she may specify subjective priors for both control population parameters as follows: $\pi(\mu_C) \sim N(50, 1^2)$ and $\pi(\sigma_C^2) = IG(\alpha = 19, \beta = 2000)$, where *N* is the normal density and *IG* is the inverse-gamma. These priors are both conjugate priors, and so can be conceptually interpreted as representing additional sample data about the unknown parameters. She can then use Jeffreys priors for the remainder of the parameters in the model, μ_T and σ_T^2.

Bayesian t-tests with a Mixture of Subjective Priors. Mixed priors represent an alternative to objective priors for μ_T. She cannot setup informed priors for it, of course, for no one knows what will happen under intervention conditions. But she usually she is not clueless about μ_T. Using a measure of effect size, such as Cohen's Δ, she may know how to evaluate it; briefly:

$$\text{Cohen's } \Delta = {}^{(\mu_T - \mu_C)}/_{\sigma_c}$$

There is a harmful or no effect ($-\infty > \Delta > .3$), small positive effect [$.3 > \Delta > .5$), medium effect [$.5 > \Delta > .8$) or large effect if [$.8 > \Delta > \infty$).

An effect size scale may not seem like much to go off on when trying to build default priors for μ_T, but I wager it suffices. Researchers at the pilot stage of inquiry, after all, care little about whether the true Δ is in fact .31 or .3102. What they really want to know is whether a future follow up study into this topic would be worthwhile or not.

Sometimes posteriors resemble neither the prior or the likelihood, and this is troublesome to most researchers. Inferences drawn from such posteriors, after all, are inconsistent with both background knowledge and the sample. To solve this problem, Bayesians can mix subjective and objective priors together to ensure posteriors resemble either the prior or likelihood (Bolstad, 2007). Mixed default in the analysis to whatever prior in the mix best fits the data. Mixed priors, thus, ensure the posterior resembles either the prior or likelihood.

Usually one mixes subjective priors with objective priors, but I propose another kind of mixed prior for μ_T: A mixture of all the priors encapsulating effect size possibilities of interest (i.e. none, small, medium, large). Bayes' postulate bids us to equally weight these priors in the mix. Actually, it demands we assign equal prior probabilities whenever in doubt, but this practice fulfills the spirit of this timeless maximum. In this case, set $\pi(\mu_T) \sim N(\mu' = 50{,}53{,}56{,}59;\ \sigma^{2'} = .333)$. This represents a mix of four normal priors with different locations, but same scale. One can then use Jeffreys prior on the other intervention parameter, σ_T^2, and the same informed priors as above on the control parameters (μ_C, σ_C^2).

Purpose of the Present Study

The purpose of this study is to evaluate Bayesian *t*-tests using convenience priors. They, unlike their classical cousins, are not - as a general rule - justified on the basis of their frequency properties (albeit, it is interesting to know whether they can really beat classical *t*-tests at their own game). But it does make sense to justify ones using convenience priors on such grounds. After all, the usual justification for Bayesian inference – posteriors represent how rational agents will judge a statistical hypothesis' credibility given samples and their personal priors – do not really apply here. Contrary to popular claims, convenience priors do not encapsulate ignorance, and so it is natural to appraise them using classical criterion instead.

Method

I simulated scientists using Bayesian independent means *t*-test to detect intervention effects with only small samples at their disposal. Sections below are organized as follows: simulation design, model specifications, data generation, and analysis plan.

Simulation Design

The design included five factors: (a) group ratio ([1 control: 1 intervention],[1 control: 2 intervention], [2 control: 1 intervention], [1 control: 3 intervention],[3 control: 1 intervention]), (b) total sample size (12,24,36,48,60), (d) population effect size ($\Delta = 0, .2, .5, .8$), (e) normal-like shapes with varying skewness (γ_1) and kurtosis (γ_2) (i.e. [$\gamma_1 = 0, \gamma_2 = 0$], [$\gamma_1 = 1$ and $\gamma_2 = 3$], [$\gamma_1 = 1.5$ and $\gamma_2 = 5$], [$\gamma_1 = 2$ and $\gamma_2 = 6$],[$\gamma_1 = 0$ and $\gamma_2 = 25$]) and finally (e) variance ratio between parent populations ([1:1],[1:1.5],[1:2],[1:3]). This cross factorial design 5X5X4X5X4 provided a total of 2000 conditions for the simulation study.

Model Specifications

Three versions of the Bayesian *t*- test were setup, each had the same five parameters ($\mu_T, \mu_C, \sigma_T^2, \sigma_C^2$, and $\mu_1 - \mu_2$), and the same likelihood functions (L_J) for the two sample groups:

$$L_T(Y_{Ti}|\mu_T, \sigma_T^2) = \phi(Y_{Ti}; \mu_T, \sigma_T^2) \text{ for i=1,...} n_T \quad L_C(Y_{Ci}|\mu_C, \sigma_C^2) = \phi(Y_{Ci}; \mu_C, \sigma_C^2) \text{ for i=1,...} n_C$$

The prior specification schemes differentiated these three versions of the Bayesian *t*-test, and Table 2 summarizes the details.

Table 2. *Specifications for Prior Distributions in Simulation Study*

Convenience Prior Type	Location Parameters (μ)	Scale Parameters (σ^2)
Objective Priors	Jeffreys prior: $\pi(\mu_T) \sim 1$ Jeffreys prior: $\pi(\mu_C) \sim 1$	Jeffreys prior: $\pi(\sigma_T^2) \sim 1/\sigma_T^2$ Jeffreys prior: $\pi(\sigma_C^2) \sim 1/\sigma_C^2$
Objective & Subjective Priors (μ_T has an objective prior)	Jeffreys prior: $\pi(\mu_T) \sim 1$ Informed prior: $\pi(\mu_C) \sim N(50, 1^2)$	Jeffreys prior: $\pi(\sigma_T^2) \sim 1/\sigma_T^2$ Subjective prior $\pi(\sigma_C^2) \sim IG(19, 2000)$
Objective & Subjective Priors (μ_T has a subjective prior)	Mixture prior: $\pi(\mu_T)$ $\sim N^*(\mu = 50,53,56,59, \sigma^2 = .333)$ Informed prior: $\pi(\mu_C) \sim N(50, 1^2)$	Jeffreys prior: $\pi(\sigma_T^2) \sim 1/\sigma_T^2$ Subjective prior $\pi(\sigma_C^2) \sim IG(19, 2000)$

*mixture of four normal distributions (*N*) with equal weights.

Data Generation

All data for this study was generated through Monte Carlo simulation methods. I employed a random number generator using the RAND function in SAS statistical software (SAS institute Inc., 2013). For each condition in the simulation, I generated 5,000 samples. The use of 5,000 replications provides a standard error of .003 when coverage is .95. I used PROC MCMC and PROC TTEST to perform Bayesian and classical *t*-tests, respectively.

Analysis Plans

This study analyzed findings using a Factorial ANOVA. I examined the frequency properties of Bayesian estimators' point estimates, probability intervals, and hypothesis tests. I used this classical metric because Bayesian estimators with convenience priors can be justifiable on the grounds of their frequency properties, even though they retain credence-type probability interpretations. This metric also made it possible to use the frequency properties of the classical *t*-test as an informative baseline.

I examined bias in point estimates, $\hat{\mu} - \mu$. In the context of Bayesian *t*-testing, it makes little difference which of the three measure of central tendency is chosen as the posterior density will be normal. But I used the median value. To test which estimator usually got us closer to the truth, I considered both the mean bias and the root mean squared error (RMSE). I examined the design factors related to their variability by conducting η^2 analysis.

I examined these estimators' probability intervals too, using the same η^2 analysis scheme. I evaluated precision and coverage rate. I did not expect a 95% credence-type probability interval to have a 95% coverage rate. But good (≥ 95) coverage rate can warrant intervals produced from Bayesian estimators with convenience priors. Again, the classical .95 criterion for 95% intervals only provided a baseline for evaluating the performance of the Bayesian intervals.

Finally, I examined Type I and Type II error rates, using the same η^2 analysis procedure. I evaluated the classical estimators Type I error rates on the basis of the liberal criterion for robustness suggested by Bradley (1978). When $\alpha = .05$, a classical estimator is considered robust when the Type I error rate falls between .025 (=.5*.05) and .075 (=1.5*.05). But this criterion was not applied to Bayesian estimators given they deliver inferences with the force of credence-type probability – not chance-type. Notwithstanding, good ($\alpha \geq .05$) Type I error control can warrant Bayesian *t*-tests with convenience priors.

Results

Results are divided into three subsections: (a) point estimates, (b) probability intervals, and (c) hypothesis tests. In favorable experimental conditions (i.e. balanced groups, and normal populations with equal variances), I selected the classical *t*-test with pooled variance as the appropriate baseline. In unfavorable experimental conditions, I selected the classical *t*-test with Satterthwaite corrections as the baseline.

Point Estimates

Favorable Conditions. Figure 2 depicts the frequency properties of estimators' point estimates under favorable conditions. The classical estimator, $\overline{X}_T - \overline{X}_C$, produced unbiased estimates of $\mu_T - \mu_C$. This feat was replicated by Bayesian estimator with Jeffreys priors for all parameters (hereafter, Jeffreys Estimator). The Bayesian estimator that had Jeffreys priors for intervention parameters and informed priors for control parameters also produced unbiased estimates in the limits (hereafter, Partial Estimator). The Bayesian estimator with a mixed prior for the parameter μ_T (hereafter, Mixed Estimator), however, provided biased estimates in every condition. The bias had a positive linear relationship with effect size. Interestingly, when its bias was averaged across all favorable conditions it dropped to zero.

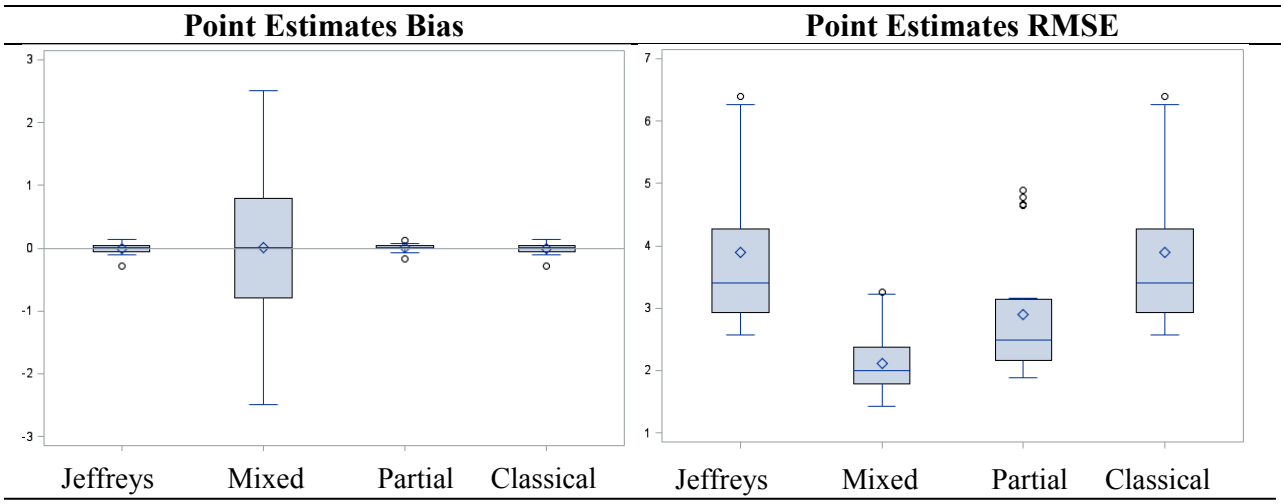

Figure 2. *Frequency Properties of Point Estimates across Favorable Conditions*
Note: The Classical Estimator used the Pooled Variance Method.

The RMSEs of the estimators' point estimates were abysmal, under favorable conditions, even for unbiased ones. It varied with sample size: smaller samples, larger RMSEs. Nonetheless, the Mixed estimator yielded point estimates with the smallest RMSE of the bunch. This pattern is consistent with the known behavior of Bayesian estimators with subjective priors. The Partial estimator also managed to outperform the classical estimator, but not to the same degree as the Mixed. Jeffreys estimator, however, only tied the classical estimator in its RMSE. Such dreadful point estimates can be expected when samples are kept tiny.

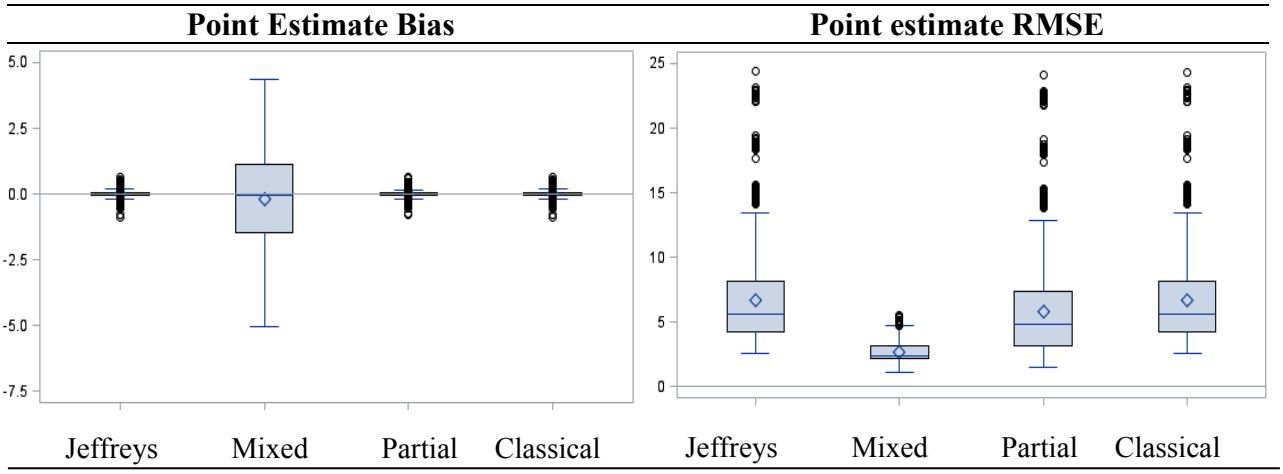

Figure 3. *Frequency Properties of Point Estimates across All Conditions*
Note: The Classical Estimator used Satterthwaite corrections.

All conditions. Moving on to consider unfavorable conditions too, the same pattern as above was again replicated here. Figure 3 summarizes these findings. The classical estimator's point estimates are unbiased, but they have awful RMSEs. This was also the case for the Jeffreys and Partial estimators. Point estimates from the Mixed estimator, again, proved to be even more biased than before, but had the lowest RMSEs. This averaged bias across conditions, however, this time did not drop to zero. Based on η^2 analysis, the pattern in RMSE among point estimates depended on heterogeneity, population effect, and balance (see Figure 4).

In summary, the properties of point estimates from the Jeffreys estimator resembled the point estimates of the classical estimator in both their bias and RMSE. The point estimates from the Partial estimator, however, behaved a little bit differently than the classical estimator. Its point estimates were unbiased, but they had much lower RMSEs. The point estimates of the Mixed estimator were biased, but with significantly lower RMSEs. Bias/RMSE trade-offs often happens when comparing Bayesians and classical procedures.

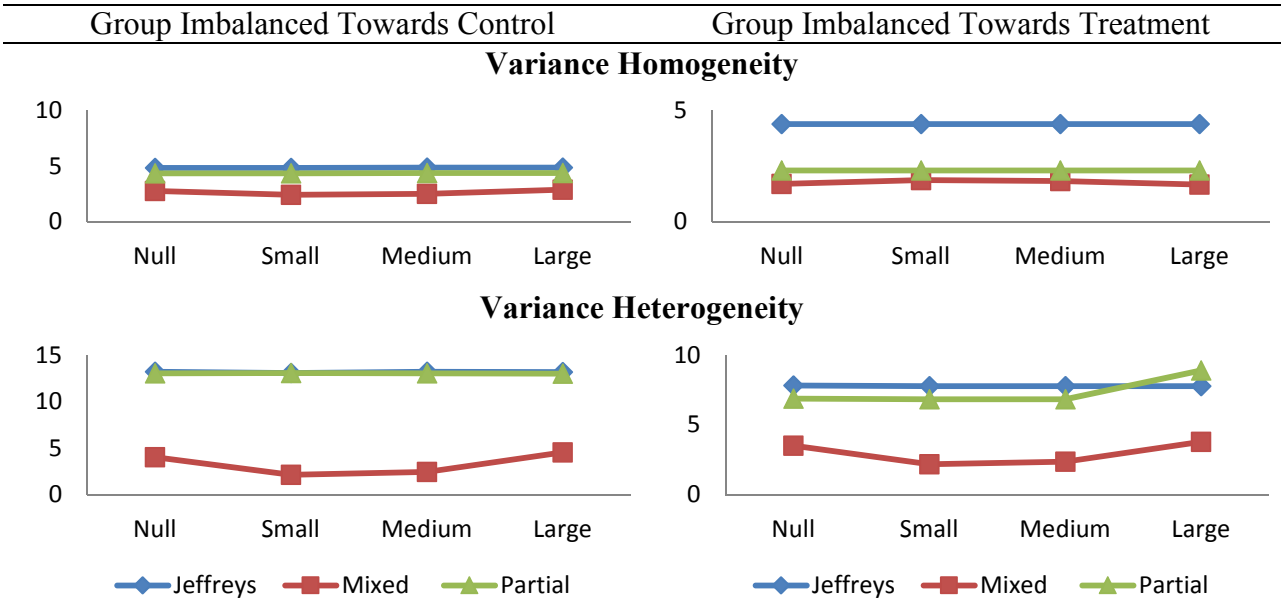

Figure 4: *Interaction between Heterogeneity and Effect Size on RMSE*
Note: Jeffreys and Classical Estimators produced same results under these conditions.

Interval Estimates

Favorable conditions. Figure 5 summarizes the frequency properties of estimators' interval estimates. They all produced imprecise intervals, even under favorable conditions, and based on η^2 analysis, the imprecision was a function of samples: smaller sizes, less precision. The Classical estimator (using pooled method) gave us 95% probability interval for $\mu_T - \mu_C$ with coverage rates just slightly below their nominal-level, but always with abysmal precision. With a sample size of 12 and a medium population effect, for example, it produced on average confidence interval width of 24.6633 and a coverage rate of .9320. Such a width is too imprecise to be useful, but coverage rate adequately approximates 95% (Bradley, 1978).

The Partial estimator, in comparison to its classical cousin, usually improved precision at the expense of coverage. In the same case as above, for example, it produced intervals with an average width of 19.4481 and a coverage rate 0.9164. This makes perfect sense, of course, for a .95 credence-type probability interval to have less than or more than a .95 coverage rate, but nonetheless a 95% coverage rate (or greater) would be a desirable trait for Bayesian estimators using convenience priors.

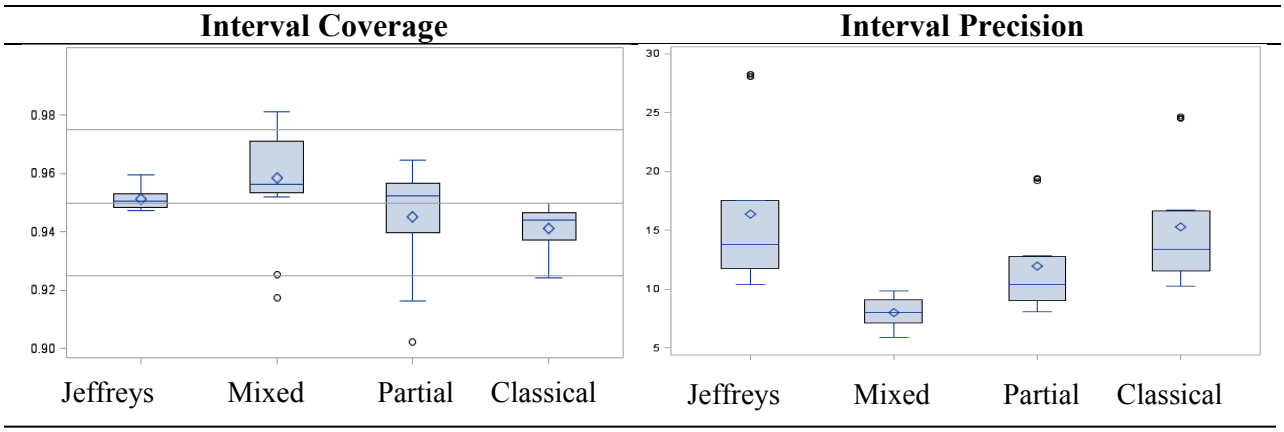

Figure 5. Frequency Properties of Interval Estimates across Favorable Conditions
***Note*:** The Classical Estimator used the Pooled Variance Method.

The Jeffreys estimator, in comparison to the classical estimator, improved coverage but only at the expense of precision. This is the reverse of what happened with the Partial estimator. In the aforementioned example – the one involving a medium effect and sample size of 12 – it produced intervals with an average width of 28.2861 and coverage of 0.9596. Its coverage, thus, matched the classical nominal-level of .95, but its precision was abysmal. This case exemplifies well the general pattern across favorable conditions. Trading off a bit of coverage to gain extra precision sometimes is a necessary evil, but the Mixed estimator did not make such a sacrifice necessary. It usually improved precision and coverage simultaneously.

The Mixed version usually preserved good coverage and precision compared to the classical estimator. But as sample sizes become tiny (ex. 12) the intervals become unstable in their coverage: sometimes worse, sometimes better than classical intervals. In the illustrative case so far considered, it produced interval widths of only 9.8754 and a coverage rate of .9540. There is still a lot of imprecision lurking in such intervals, of course, but this is by far the best precision in the bunch. Gaining this precision also did not require any loss of coverage. This pattern occurred across favorable conditions if sample sizes were kept above 24.

Unfavorable Conditions. The story of interval estimates again unfolds along the familiar lines as before, except in the case of the Partial estimator (see Figure 6 and 7). The Partial estimator's intervals were more precise on average than classical intervals (produced with Satterthwaite corrections), but they had lower coverage rates. The noteworthy exception to this general trend is when sample sizes got bigger than 24 and groups became imbalanced with more people in the intervention conditions. Under these conditions, the Partial estimator usually always managed to outperform the classical estimator in terms of both its precision and coverage rate.

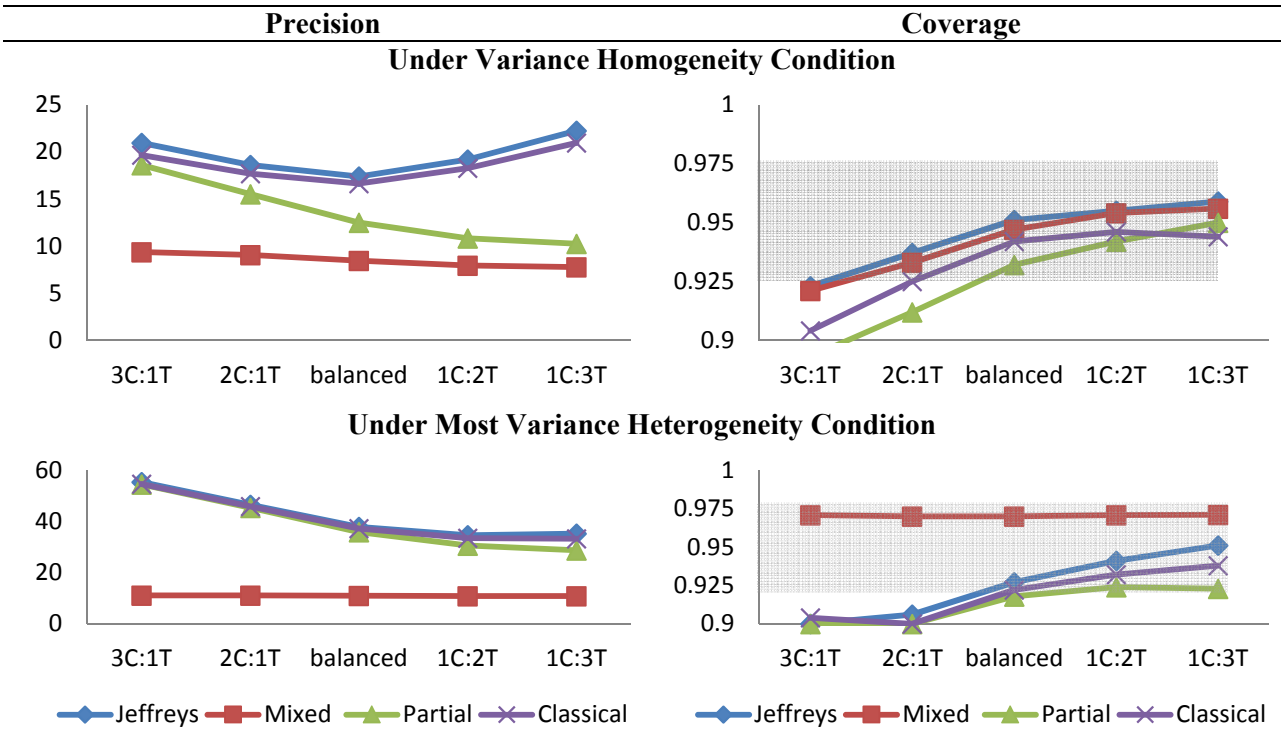

Figure 6: *Analysis of Interactions between Estimator and Balance (N=24)*
Note: Read "2C:1T" represent group ratios, and read as "2 control units to 1 intervention unit."

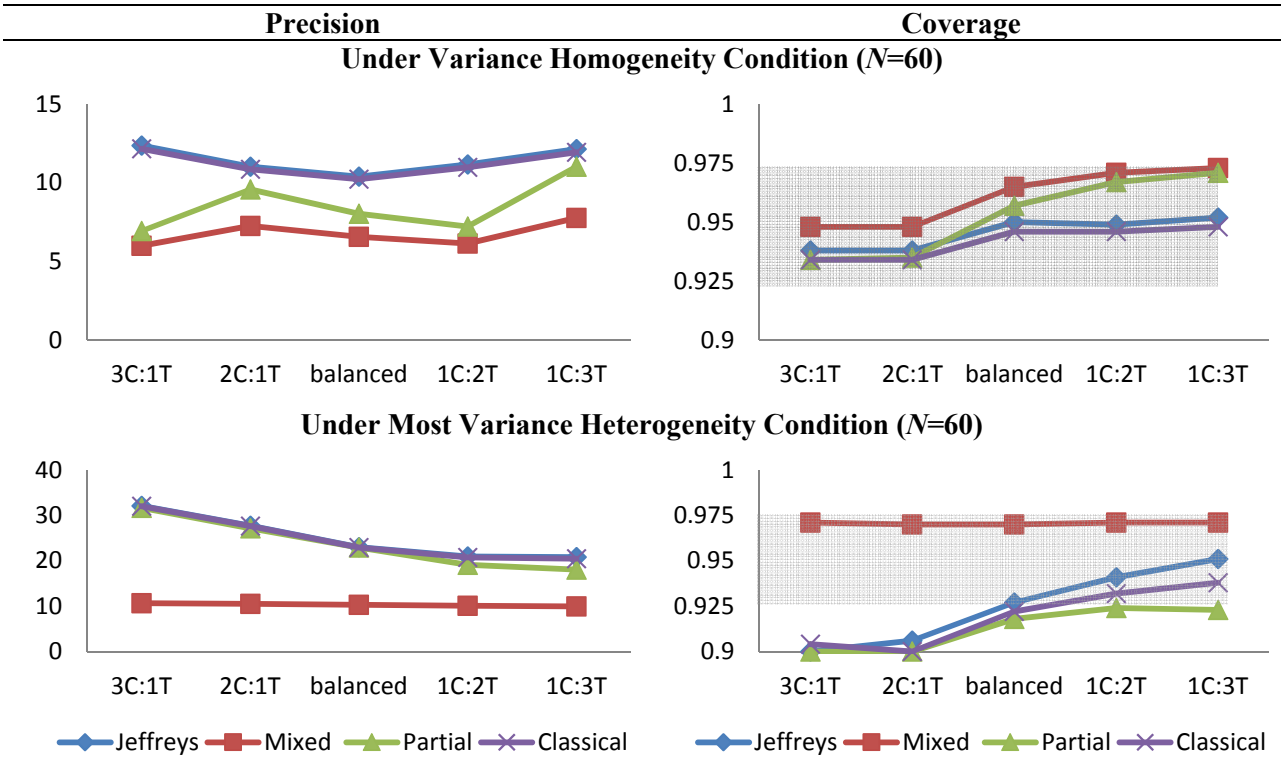

Figure 7: *Analysis of Interactions between Estimator and Balance (N=60)*
Note: Again, "C:T" represents group ratios.

Jeffreys estimator outperformed the classical estimator (using Satterthwaite method) in sticking to a 95% coverage rate (or better), but again only at the expensive of some precision. To preserve a desirable coverage rate, for example, it made intervals more imprecise to account for the extra uncertainty of unfavorable conditions. The Mixed estimator, in contrast, consistently yielded much more precise intervals for $\mu_T - \mu_C$ than all the other estimators while usually sustaining a greater or comparable coverage rate to the classical estimator. Its accuracy rate, however, had more variability, but it stabilized above 95% as sample size rose ($N \geq 36$). Figure 8 provides a visual representation of these findings.

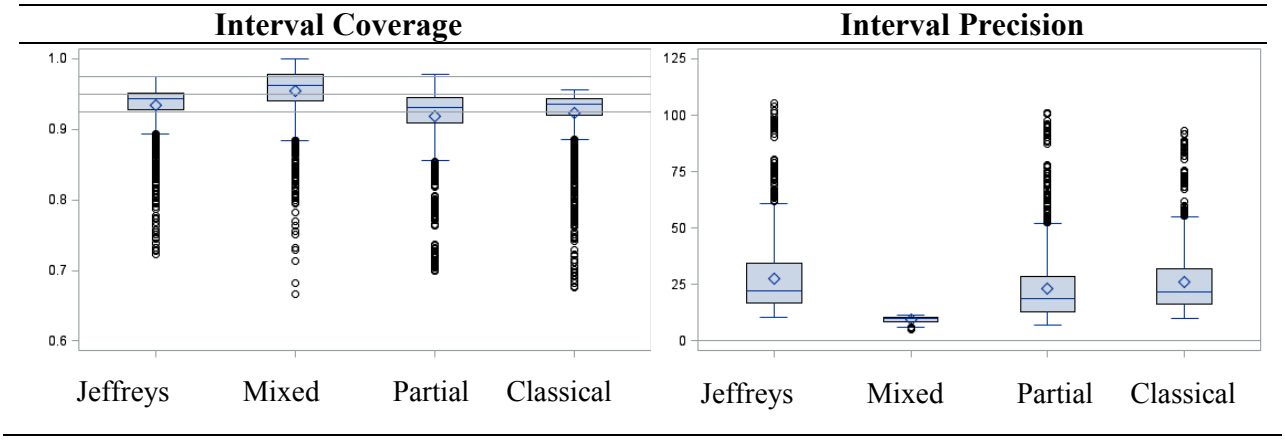

Figure 8. *Frequency properties of Interval Estimates across All Conditions*

Hypothesis Tests

In favorable conditions, classical estimators are known to maintain adequate Type I error control. In the classical scheme, probabilities are only about relative frequencies in the limiting cases. With this strict interpretation, Classical statisticians cannot make sense of a test with a specified nominal 5% Type I error rate ($\alpha = .05$) and an actual $\alpha \neq .05$ on repeated applications of the test in identical situations. Bayesians, in contrast, have no qualms about such mismatches. They cite credence-type probability rather than frequency-type. I, therefore, permitted actual Type I error rates to be lower than the nominal in the case of Bayesian estimators.

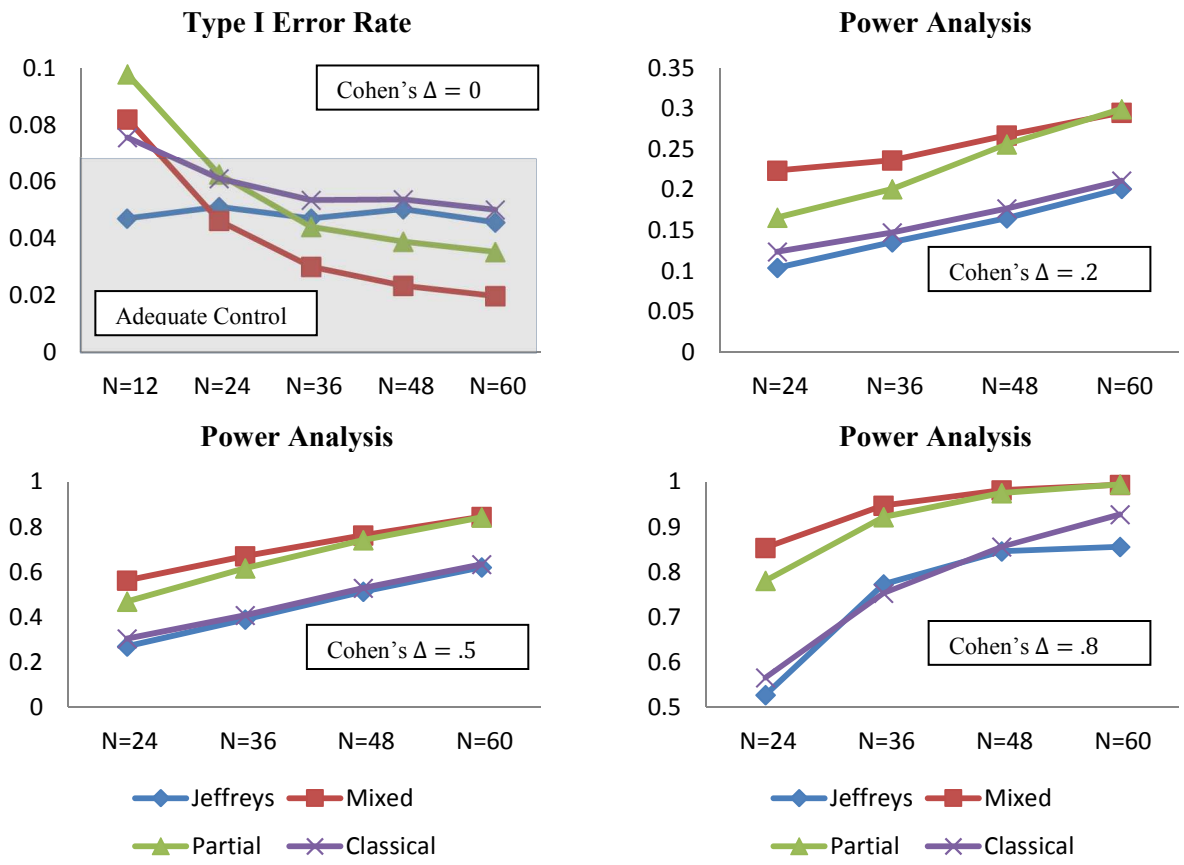

Figure 9. *Analysis of Type I Error Rate and Power across Favorable Conditions*
Note: Sample size (*N*=12) excluded from power analysis, because of inadequate control of Type I error.

Figure 9 reports Type I and Power findings for each estimator in favorable conditions. Again, all the tests had a nominal $\alpha = .05$, and my simulation results for the smallest sample size (*N*=12) condition proved to be problematic as they conflict with the known sampling density of the *t*-test statistic provided by exact mathematical results. This discrepancy is most likely due to simulation error, and I, therefore, discarded this condition from power analysis to play it safe. In terms of my power analyses, the Jeffreys estimator is slightly inferior to its classical cousin. It traded off a bit of power to gain the edge on Type I error control. Mixed and Partial estimators, at least under favorable conditions, consistently outperformed the classical estimator in terms of power, with the Mixed estimator's power prevailing overall.

Unfavorable conditions. When unfavorable conditions were thrown into the mix, I used Bradley's (1978) liberal criterion of robustness to examine the classical estimator's Type I error rates. This criterion is not quite consistent with frequency-type probability, but it is justified on grounds of practical equivalence – i.e. classical statisticians can consider tests with $\alpha = .05$ to be robust to violations of assumptions within tolerance, $025 \leq \alpha \geq .075$. Again, Bayesians do not need to invoke liberal criterions. But to make the comparison fair, I adjusted my criterion for evaluating them to, $\alpha \leq .075$. Figure 10 shows the results of Type I error analysis.

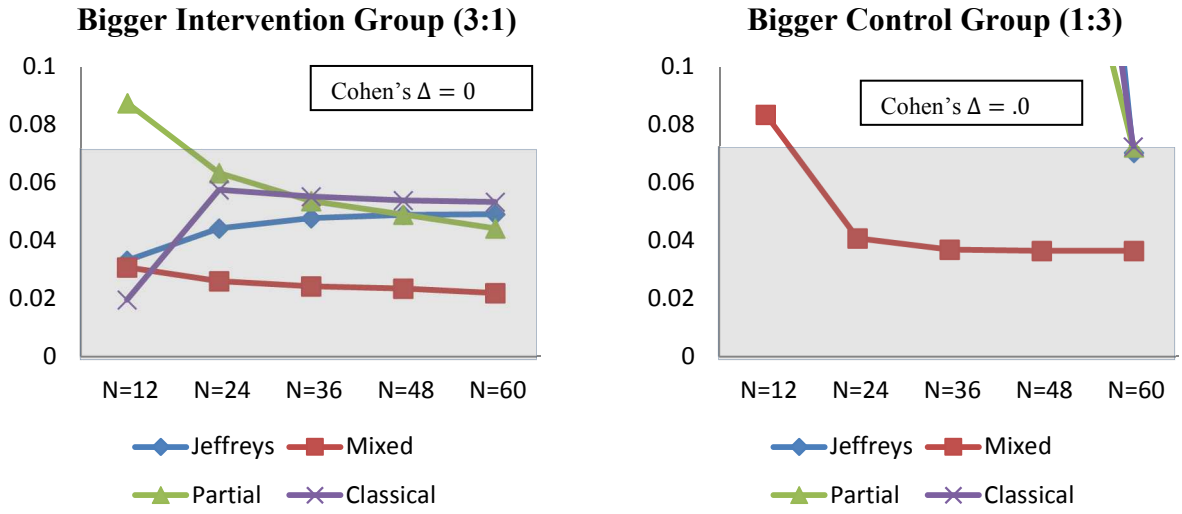

Figure 10. *Analysis of Type I Error Rate given Most Imbalanced Group Condition*

The Mixed prior again stole the show. Figure 10 depicts the interaction between group balance and sample size in the analysis of Type I error control. There were two ways, of course, groups could be imbalanced: bigger (a) control group or (b) intervention group. The Mixed estimator seemed most robust to the problem of imbalanced group. Figure 11 shows power analyses for all conditions wherein Type I error was adequately controlled. Naturally, the power of all estimators dropped to pathetic levels as population variances become more and more unequal and sample sizes shrink. Those depressing results are now shown here.

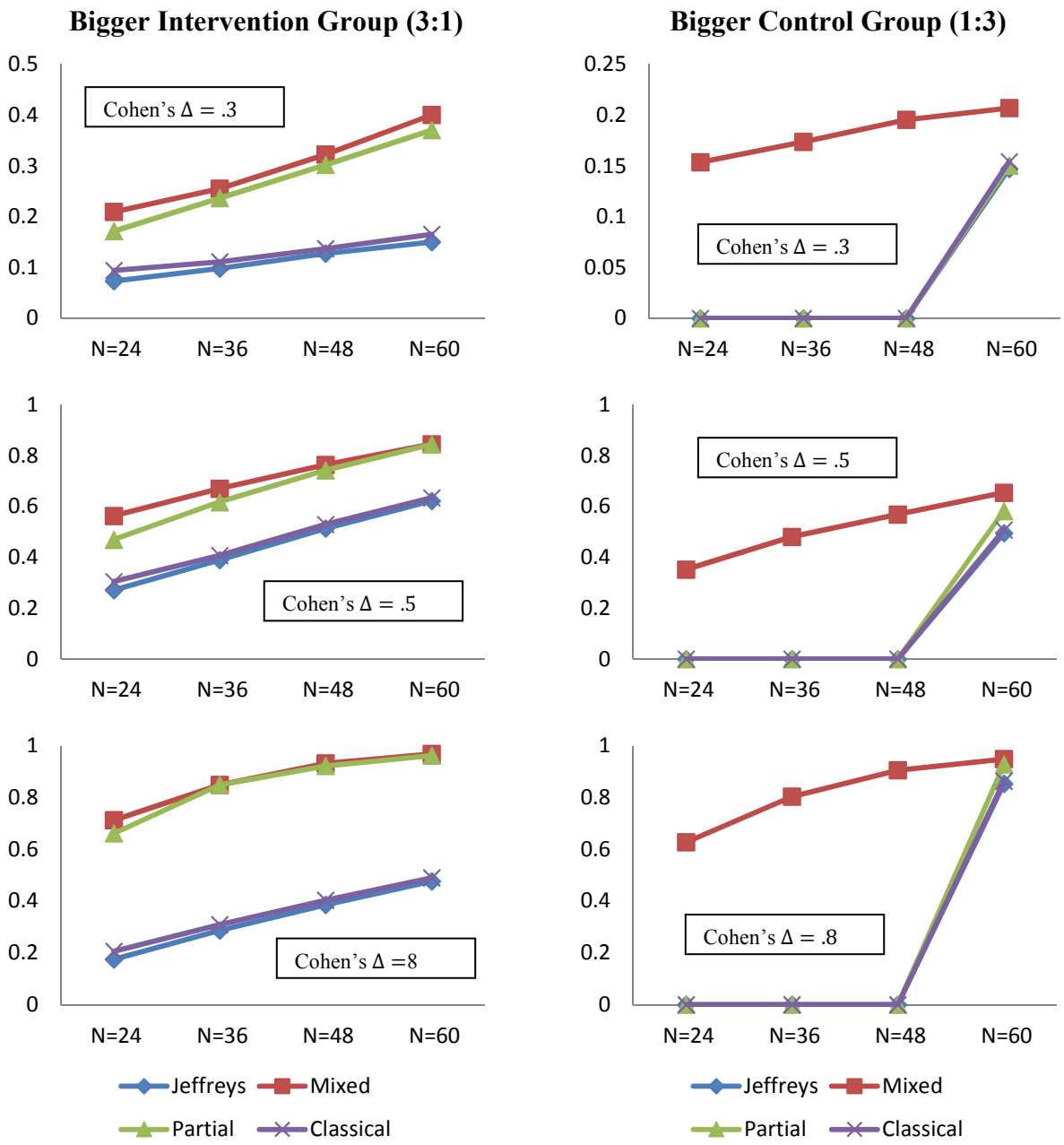

Figure 11. *Analysis of Power Given Most Imbalanced Group Condition*
Note: Only includes conditions where Type I error controlled and variances homogenous.

This power analysis shows Jeffreys estimator underperformed compared to the classical estimator. The Partial estimator's power increased when group size became imbalanced in favor of the intervention group, but had no similar increase in power when control group size was larger. The Mixed estimator, however, had the most power in the group.

Discussion

The purpose of this study was to evaluate whether mixed priors could plausibly function as convenience priors in Bayesian *t*-tests of experiments involving tiny groups using frequency criterion. The risks of using small samples are well known, but scientists often must resort to them – at least, in the beginning of their research projects. Large scale-studies are expensive, and funding agencies usually want solid evidence their substantial fiscal investments into such experiments will likely be worthwhile. To supply the needed evidence, scientists resort to small-scale pilot studies.

Based on small pilot study findings, it is not wise to be dogmatic about point and interval estimates. All the point estimates in this simulation study, for example, had abysmal RMSEs. We can interpret this as entailing that we expect on any given occasion the point estimates to be far away from the truth of the matter. And this expectation holds under favorable and unfavorable conditions alike. In the limits, point estimates of Jeffreys estimator and the Partial estimator, like their classical cousin, were unbiased. "But in the limits," Bayesians like to say, "we will all be dead" so this is little comfort. These findings about point estimates, therefore, support the well known advice to cautiously interpret point estimates when sample sizes are small.

The same caution applies to probability intervals. All the Bayesian estimators had good coverage properties, under the best of conditions, but the imprecision in them made them useless. It is off little help to know the interval accuracy rate is good, for example, if the interval does not eliminate any parameters of interest. This is the peril of working with small samples, and not even the Mixed estimator can compensate for this hazard. It did give the most precise intervals of the bunch, but even so they were still far too imprecise to be of much use. Experimenters using *t*-tests should keep their expectations low for interval estimates.

Fortunately, scientists do not need the information supplied by point estimates and confidence intervals to convince funders their research is worthwhile. Hypothesis tests suffice for this. At the minimum, significant data findings can justify funding for a slightly larger pilot study to get a more exact fix on the unknown effect size. The results of this simulation study suggest the Mixed estimator has certain key advantageous over the others – at least, when it comes to Bayesian *t*-tests. Given favorable conditions, these tests had both adequate control of Type I error and substantial power. It, in fact, they outperformed all the other contestants in the competition.

The only time it stumbled was when total sample sizes shrunk and variance heterogeneity rose in excess. But, in fairness, all the estimators stumbled in such cases. There is just too much uncertainty when this happens to rule out the null hypothesis. Some statisticians, in fact, argue that in these circumstances it no longer makes sense to focus on means (Bolstad, 2007; Hoff, 2009). Instead, the focus should be on population variances. Therefore, the best advice may be to come up with a new statistical test altogether in cases of excessive heterogeneity.

Outside of those special circumstances, however, it makes sense to keep the focus on the population locations. Point estimates and interval estimators are interesting, but the true purpose of the Bayesian *t*-test is hypothesis testing. It is in this area that the Mixed estimator boasts ample power and adequate Type I error control.

Its success here evidences that one does not need extensive background knowledge or even an objective formula to safeguard Bayesian inferences against bias. So long as one does not expect too much, the Mixed version of the Bayesian *t*-test works well. It is simply amazing how far knowledge of what constitutes an effect size of practical significance can carry a Bayesian *t*-test, and this minimal amount of background knowledge is usually obtainable.

The danger of all Bayesian inference, of course, is bias. But Bayesian epistemologists argue that Bayesian statistics has a filter built into it that is capable of eliminating bias. Each time a new sample is drawn, for example, Bayesians can use their previous posterior probabilities as the new prior probabilities. Over the long haul, as the body evidence grows, any bias in the initial priors will be overwhelmed by the objective sample evidence.

Limitations

This study had several limitations worth considering. One limitation was that I had to fix the true population size at certain values, and then setup informed priors for them. This felt like cheating. In real-life, statisticians rarely have the level of certainty I had when setting up priors. Therefore, it is unlikely that researchers using Partial Estimators can produce unbiased estimates. I only pulled off this feat because my informed priors were unbiased. Fortunately, no substantive conclusions in this study hung on whether or not point estimates were unbiased. All the studied estimators produced untrustworthy point estimates in this simulation.

Another limitation of this study is that I did not examine what happens when the practical significance of a true effect size is fuzzy. For example, suppose the true population effect size rests on the border between a small and medium effect sizes. This might then wreck havoc on the success of the Mixed estimator as the samples will sometimes appear small or medium because of sampling error. If so, the mixed estimator may too bounce back and forth between the small and medium effect size hypothesis. This is because it defaults to one of its component priors in the analysis based on what the sample suggests. Additionally, future studies may want to investigate the Mixed estimator's performance in more complex sampling situations. Education researchers, for example, often have nested data, and it would be interesting to see how well the Mixed Estimator fares in a multilevel version of the *t*-test.

Final Remarks

Some argue that the Bayesian *t*-test supersedes the classical *t*-test. Personally, I am eclectic in my statistical philosophy. I do not fault the classical *t*-test because it is not the Bayesian *t*-test (or *vice versa*). There is place for both versions of the test in the researcher's toolkit. But this study is useful because the Bayesian *t*-test is more generally applicable in education research (see Chapter 2).

In education research, for example, it is often unethical or infeasible to gather a probability sample and use random assignment in a small-scale pilot study. But the logic of classical *t*-tests is inadmissible without these design features. The logic of the Bayesian *t*-test, however, is not so constrained. All Bayesians need, to proceed with the *t*-test, is a representative sample of the population, and such samples can be selected using good judgment rather than a randomization devise.

The general applicability of Bayesian *t*-test in education science makes the present investigation into convenience priors necessary. Researchers can use it to make better informed decision about their choice of prior when they proceed with Bayesian *t*-tests in their pilot studies involving small, non-probability samples – an all too common reality in education research. Researchers conducting such small-scale tests, as is well known, should stay away from point and interval estimates. But hypothesis tests were made to function in these environments. If data are significant based on a hypothesis test then there is evidence a follow up study using a bigger sample may find something interesting and get a closer fix on the true effect size, and that is often all we want to know in education research (Mohr, 1990).

CHAPTER FOUR

QUANTIFYING EFFECTS OF GROUPING METHODS ON READING OUTCOMES

In the highly litigated field of special education, placement decisions matter (Yell, 2006). The decisive verdict of the U.S. Supreme Court (Board of Ed. of Hendrick Hudson Central School Dist. v. Rowley, 1982), for example, shows school districts are accountable to the public for such decisions. They may not conveniently place students with disabilities in basements, and hope for the best. Instead, under IDEA (2004), they must place students in special education in their Least Restrictive Environment (LRE), as specified by an IEP team. This clause eventuates in the need for a hierarchical continuum of placement options for students in special education, ranging from the most inclusionary to the most segregated classrooms. Inclusionary settings, unlike their segregated counterparts, are designed to amply educate students with and without disabilities alike in the same physical locale (Gilhool, 1989).

In the last few decades, inclusionary placements eclipsed segregated placements as the norm in the U.S. school system. In 2011, 13% of students in the U.S. school system were eligible for special education services. The overwhelming majority of these students (95%) were placed in regular schools. Likewise, the practice of placing them in general education classrooms for more than 80% of the school day steadily increased from 31.7% in the fall of 1989 to 61.1% in the fall of 2011, an almost 200% increase (U.S. Department of Education, 2013). Trends of this kind have led some scholars to the make the slightly exaggerated claim *full inclusion with co-teaching* is becoming the default model of service delivery (Zigmond, Kloo, & Volonino, 2009). The momentum for inclusion is also not likely to wane anytime soon (Jorgensen, 2005).

Different Grouping Methods for Enacting the Vision of Inclusion

When inclusionary placements become the majority, districts must adapt to fit this new actuality. They should place students in special education at neighborhood elementary schools (Burrello, Sailor, & Kleinhammer-Tramill, 2013; Toson, Burrello, & Knollman, 2012; Marks, 2011). This policy naturally promotes a resemblance between district and school ratios. To illustrate, if 13% of students in special education hold inclusionary placements in a district, then approximately 13% of the students at each schools should also be in special education. Natural proportioning at the district-level, thus, ensures schools have neither too many nor too few students in special education to enact inclusion (McLeskey & Waldron, 2000).

At the school-level, there are two prevailing grouping methods for enacting inclusion, natural proportioning and clustering. Natural proportioning ties classroom ratios to school ratios. For example, if the ratio of incoming 1st graders with disabilities to those without disabilities is 1:9 within a certain school, then administrators should assign 1st graders with disabilities (n=12) to all classrooms (n=6) in ratios replicating the overall grade-level ratio to achieve the natural proportion. Perhaps, given class sizes of 18, they will place 2 students with disabilities in each classroom to achieve the requisite ratio.

McLeskey and Waldron (2000) argue, in contrast, achieving natural proportioning at the school-level is not always best practice. Students and classrooms are not exchangeable goods. And, especially at the early phases of enacting inclusion at a school, some classrooms will have superior service-infrastructures to support students in special education, and clustering such students in these classrooms makes sense. Clustering, thus, maintains instructional quality at the school. Advocates of this method, of course, can still view natural proportioning as a regulative ideal, but they balance its desirability with the need to sustain instructional quality.

The Urgent Need to Improve Special Education Services for Reading

In the area of reading, there is an urgent need to rethink how special education services are being utilized to improve reading in schoolchildren with disabilities in U.S. schools (Morgan, Frisco, Farkas, & Hibel, 2010; Sullivan & Field, 2013). On the basis of a secondary analysis of a large-scale, longitudinal, and nationally representative sample of elementary schoolchildren – namely, ECLS-K – Morgan, et al., (2010) concluded that children in special education in the spring of 2002, aggregated across classroom placements, had lower reading skills in the spring of 2004 than closely matched peers in general education. Further analysis of this data set by Sullivan and Field (2013) found the receipt of services actually had a negative effect on reading skills among matched pairs of kindergarten children, ($d = -.21$).

The above evidence raises the troublesome possibility that - at least, in terms of reading – the U.S. tax payer's investment in extra special education services in the elementary years is not yielding good returns. The above studies suggested, general education is likely outperforming special education in producing higher rates of reading achievement among youth with disabilities (Sullivan & Fields, 2013). Indeed, children placed in special education classrooms sometimes even scored lower on post-test scores (Lane, Wehby, Little, & Cooley, 2005).

Inclusionary classrooms celebrate ability-heterogeneity among students (Jones, Carr, & Fauske, 2011), and, unlike traditional general education classrooms (Anastasiou & Kauffman, 2012), are equipped with a multi-tiered support system to help ensure higher quality instruction for students with disabilities (McLesky & Waldron, 2011). But inclusionary classrooms are not a panacea (Kauffman & Badar, 2014), and it would be worthwhile to investigate whether the different ways of enacting inclusion, such as using clustering rather than natural proportioning, can produce better reading outcomes for elementary schoolchildren.

The Bayesian Historical Control Trials in Special Education

Clinical trials are among the best ways to induce causation in science. To warrant such inductions, researchers usually require studied groups in the clinical trials to be comparable on extraneous variables. In education science, however, researchers cannot be certain their groups are optimally balanced. Instead, they settle for groups more likely balanced than not. But it is possible to increase the probability groups are balanced in clinical trials.

In Randomized Controlled Trials (hereafter, RCT), researchers construct groups using a physical randomization device, such as die, to augment the probability their studied groups are adequately balanced (Fisher, 1926). Such devices invoke a certain type of probability, physical chance. Flip a fair coin countless times, and one observes a balanced proportion of heads and tails in the limits. The logic of significance tests based on classical *p*-values also depends on this type of physical probability (Chapter Two).

Historical Controlled Trials (hereafter, HCTs) provide researchers with an alternative to RCTs. In HCTs, researchers select groups they deem comparable. They appeal then to yet another type of probability, credence. It is probable, given evidence, that Caesar crossed the Rubicon. This is credence-type probability, and it is a matter of judgment. Bayesian inference, thus, can proceed with causal inference without randomization in study designs.

When experimental units belong to a protected population, like schoolchildren, it is often infeasible to use RCTs. Bayesian HCTs, in comparison, are well suited for education science. For example, researchers, using school records, can use HCTs to observe comparable groups of schoolchildren under different interventions (Howson & Urbach, 2006). Such groups can be constructed using propensity score matching techniques from school records (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007).

HCTs are often preferable to RCTs on ethical grounds, and empirical investigation suggests they each can balance groups on extraneous variables (Benson & Hartz, 2000; Worrall, 2007). HCTs are natural experiments. Harrison (2011) defines such experiments as occurring when history supplies researchers groups comparable on their extraneous variables: observed serendipity. The extra difficulty of doing special education research (Odom, et al., 2005), makes researchers in this field well positioned to take the lead on methodological innovation in the education field by implementing and evaluating the usefulness of new methods, like HCTs with Bayesian analysis (Kauffman, 2011).

Study's Purpose

The purpose of my paper is to estimate the differential effects of enacting inclusion using the natural proportions and clustering methods on reading skills acquisition among elementary students, using Bayesian HCTs. Specifically, my objective is to assess the credibility of the null hypothesis stating that the typical effect of using clustering rather than natural proportioning in inclusionary settings is negligible for all practical purposes. To test this null, of course, I will utilize Bayesian analysis rather than classical analysis.

Research Questions

In this study, I address the following questions:

1. Is the effect of the natural proportioning method on annual reading gains for children in *general* education different from the effect of clustering by a significant amount, as measured by official state reading tests?
2. Is the effect of the natural proportioning method on annual reading gains for children in *special* education different from the effect of clustering by a significant amount, as measured by official state reading tests?

Research Design

Study's Context

The studied schools belong to one of the largest metropolitan school districts in the U.S. In 2005, several policy and compliance issues prompted the district to embark on an aggressive campaign to reform its inclusionary policies and practices. A recent state report found the district to be "out-of-compliance" (Barton, Burrello, & Kleinhammer-Tramill, 2012).To remedy the situation, the district decided to increase the percentage of schoolchildren with disabilities in inclusionary settings in regular schools from 57% to 80%. The national average at that time was well above 90% (U.S. Department of Education, 2014).

To achieve this task, they needed to improve the capacity of their schools to fully serve all students in general education classrooms. A reform initiative took root in subsequent years at both district- and school-levels, but the success of inclusion varied across schools and practices looked different - even in successful schools. In 2008, district leadership became particularly interested in how two high performing elementary schools enacted inclusion, Clark Elementary School and Shady Lake Elementary School (pseudonyms).

In 2010, Hoppey, Black, and Mickelson (2015) conducted an external investigation into these schools using qualitative case-study methods. They observed classrooms in the schools and interviewed key stakeholders (ex. administrators, teachers, and parents). They discovered each school was committed to improving their inclusionary practices. Their quality of instruction also seemed comparable, as evidenced by schoolwide grades. But, in the midst of all the observed commonality between the schools, they suddenly realized that there was an important contrast. In essence, Clark had abandoned the clustering method in favor of natural proportioning. Shady Lakes, however, still advocated clustering, and strove to perfect it.

Ethical Considerations

The school district de-identified all data to keep everything confidential, and a university IRB approved all study protocols. My utilization of historical controls in this study also protected the legal and moral rights of youth eligible for special education services as I did not disturb or interfere with the normal educational process to conduct the study.

Study's Data Set

School district personnel provided me a judgment sample for my analysis. Judgment samples resemble populations by design rather than chance (Howson & Urbach, 2006). The data set consisted of two sample groups of 5th grade students (*N*=216) from Clark (*n*=121) and Shady Lake *(n*=95) in the same school year (08-09). The 5th grade cohorts were representative of both general and special education students placed in inclusionary classrooms at the schools. The school district collected and maintained the data set. Table 3 reports descriptive statistics for the data set, and I refer interested readers to Hoppey's, et al. (2015), case study for information about these schools beyond what is supplied by this quantitative data set.

Table 3. *Statistics for Each Sample Group (2008-2009 School Year)*

	Clark	Shady Lakes		Clark	Shady Lakes
Schoolchildren	95 (43.98%)	121 (56.02%)	**Migrant**	0 (0%)	0 (0%)
Service Type			**Female**	38 (40%)	65 (53.72%)
General:	58 (61.05%)	85 (70.25%)	***FRL Eligible**	68 (71.58%)	36 (29.75%)
Special:	37 (38.95%)	36 (29.75%)	**504 Plan**	8 (8.42%)	4 (3.31%)
Exceptionalities			**Race**		
Autism:	0 (0%)	2 (1.65)	Asian:	1 (1.05%)	5 (4.13%)
Behavioral:	0 (0%)	9 (7.44%)	Black:	2 (2.11%)	7 (5.74%)
Gifted:	22 (23.16%)	5 (4.13%)	Hispanic:	10 (10.53%)	12 (9.92%)
Language:	2 (2.11%)	0 (0%)	Islander:	1 (1.05%)	0 (0%)
Learning:	13 (13.68%)	11 (9.09%)	Mixed:	5 (5.25%)	5 (4.13%)
Other:	0 (0%)	1 (.83%)	White:	76 (80%)	92 (76.03%)
Speech:	0 (0%)	8 (6.61%)	**ESOL Eligible**	1 (1.16)	3 (2.97%)
Reading					
*Annual Gains	87 (SD=182)	97 (SD=227)			

***FRL=Free/Reduce Lunch Eligible; I report Mean and Standard Deviation for reading score gains.
.***Note***: Clark implemented Natural proportioning; Shady Lakes clustering.

Analytic Sample

To be included in the analysis, a schoolchild in a clustered classroom needed to be paired up with a schoolchild in a naturally proportioned classroom on selected covariates and needed to have values for the dependent variable. The constrained sample had 60 matched cases in total: 20 cases were in general education; 10 cases in special education.

Study's Measures

I predicted the effects of clustering and natural proportioning methods on state test scores. Specifically, I tested whether I could predict annual gains in reading achievement among 5th grade schoolchildren in the 2008-09 school year on the basis of their placement at either Clark (natural proportioning) or Shady Lakes (clustering).

Reading Achievement. I used gain scores on the State reading tests as proxies for annual reading achievement. Specifically, I used differences in scaled scores from the 07-08 and 08-09 waves of the official state reading test, $Y_{0809} - Y_{0708}$. Gain scores increase the probability that any observed differences between pairs of matched students are not mere statistical artifices of initial learning. Methodologists consider utilizations of gains scores to be appropriate for studies estimating effects (Allison, 2005).

Naturally Proportioning. For my study, I operationalized it as a placement at Clark Elementary School. Naturally proportioning, as a schoolwide policy rather than districtwide policy, enacts inclusion by assigning students to classrooms in ways designed to adequately preserve the grade-level ratio of students with and without disabilities at the school (McLeskey & Waldron, 2000). Within tolerable limits, school administrators trust natural proportioning to produce the conditions in classrooms friendly for highly effective instruction. The premise is that smaller proportions of schoolchildren in special education are usually manageable.

Table 4. *Summary of Groups after Propensity Score Matching*

	Special Education (*N*=20)		General Education (*N=40*)	
	Clark	Shady Lakes	Clark	Shady Lakes
Female	3 (30%)	4 (4%)	8 (40%)	6 (45.45%)
FRL Eligible	6 (60%)	6 (6%)	8 (40%)	7 (36.36%)
504 Plans	0 (0%)	0 (0%)	2 (20%)	1 (4%)
Race				
Asian:	0 (0%)	0 (0%)	1 (5%)	0 (0%)
Black:	0 (0%)	0 (0%)	0 (0%)	1 (5%)
Hispanic:	1 (10%)	1 (10%)	2 (10%)	2 (10%)
White:	9 (90%)	9 (90%)	17 (85%)	17 (85%)
Exceptionalities				
Gifted:	1 (10%)	1 (10%)	0 (0%)	0 (0%)
Language:	2 (20%)	2 (20%)	0 (0%)	0 (0%)
Learning:	7 (70%)	7 (70%)	0 (0%)	0 (0%)
4th Grade Reading State Test				
Failed	4 (40%)	4 (40%)	4 (20%)	2 (10%)
Passed	4 (40%)	3 (30%)	5 (25%)	6 (30%)
Mastered	2 (20%)	3 (30%)	11 (55%)	12 (60%)
4th Grade Math State Test				
Failed	6 (60%)	5 (50%)	4 (20%)	1 (5%)
Passed	1 (10%)	3 (30%)	10 (5%)	11 (55%)
Mastered	3 (30%)	2 (20%)	6 (30%)	8 (40%)

Table 5. *Odds Ratio of Groups Before and After Propensity Score Matching (PMS)*

	Special Education (*N*=20)		General Education (*N=40*)	
	Before PSM	After PSM	Before PSM	After PSM
Female	1.6852	.15556	2.6384	.6429
FRL Eligible	.2398	1	13.0659	.8077
504 Plans	.3122	1	3.3056	.4737
Race				
Asian:	--	--	0	--
Black:	.4667	--	.4198	--
Hispanic:	.3	1	2.4427	1
White:	.3448	1	1.1556	1
Exceptionalities				
Gifted:	9.0909	1	--	
Language:	--		--	
Learning:	1.2158	1	--	
4th Grade Reading State Test				
Failed	1.7858	1	1.2121	.4444
Passed	.56	1.5556	1.7113	1.2857
Mastered	2.375	1.7143	1.8281	.8148
4th Grade Math State Test				
Failed	.535	1.5	1.208.	.2105
Passed	4.6667	.2593	.8374	.8182
Mastered	5.2024	.5833	.6375	1.5556

Note. Cohen's odd ratios effect size: small = 1.50; medium = 2.50; large = 4.30. Read the odds ratio of "1.6852" as the odds of being female in the natural proportioning condition are 1.6852 times the odds of being female in the clustering condition.

Clustering. For my study, I operationalized it as a placement at Shady Lake Elementary School. Clustering, as a schoolwide policy, enacts inclusion without prioritizing classroom ratios representing natural proportions. School administrators, for example, may place students with disabilities into classrooms above or below their natural proportions. They will place reasonable clusters of students with disabilities in classrooms they judge to have superior service-infrastructure in place to support students with disabilities. I considered a reasonable cluster size for an inclusionary classroom to be less than one 1/3 of the total class size.

Data-Analytic Plan

Propensity Score Matching. I used a propensity scores matching technique to improve the internal validity of my study (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). Propensity scores can improve inferences to causation by increasing the visibility of the effect of the types of inclusionary settings of interest in this study. Such effects on reading gain scores among sampled schoolchildren are frequently hidden by extraneous variables. It is usually possible, however, to statistically balance naturally occurring groups on such variables using propensity score matching, if they are known and measured. But researchers must select the set of covariates using their best judgment rather than a formula.

My process for propensity score matching had two stages. Stage 1 involved using logistic regression to calculate propensity scores. The propensity score $p(T)$ is defined as follows:

$$p(T) \equiv \text{P(T=1|S)=E(T|S)}$$

where $p(T)$ is the propensity a child in my sample would be educated in a natural proportions inclusionary setting, *T* indicates a child did or did not receive a natural proportions inclusionary setting, and *S* represents a vector of covariates influencing whether a child did or did not receive a natural proportions inclusionary setting.

In stage 2, I balanced groups with propensity scores. I used a 1:1 caliper strategy without replacement to match them. I specified a caliper of .1 (i.e. a caliper is the maximum allowable propensity score difference of paired up schoolchildren). As a quality control measure, I also randomly ordered schoolchildren before matching them. I used a SAS MACRO to implement this procedure (Lanehart, Rodriquez de Gil, Kim, Bellara, Kromrey, & Lee, 2012).

The school district selected the covariates for analysis. After statistical balancing groups, I had 10 matched pairs in special education and 20 in general education. Tables 4 and 5 provide balance statistics on covariates before and after propensity score matching. The balanced groups were no longer representative of their respective school populations, but they were comparable enough to induce effects. In other words, I weakened the study's external validity to improve its internal validity. I judged this trade off acceptable given the study's purpose.

Bayesian Methods. The Bayesian version of statistical inference offers education researchers an alternative to classical inference. Bayesians dispenses with classical *p*-values (Iversen, 1984), and uses posterior distributions instead. Posterior distributions are statistical distribution representing a scientist's informed judgment about the credibility of hypotheses, under consideration, given sample evidence. All statistical inferences in this study were framed within the Bayesian rather than classical mold.

Classical *p*-values are premised on chance setups in the study design – a proviso too often neglected. The logic of Bayesian posterior distributions, in contrast, encompasses study designs using judgment samples and propensity score matching. Randomization becomes an optional component in study designs for Bayesians (Howson & Urbach, 2006). I, thus, selected Bayesian inference because of its greater flexibility in study designs (see Chapter Two).

I used Markov chain Monte Carlo (MCMC) to sample from posterior distributions, and compute quantities of interest. Specifically, I used PROC MCMC (SAS 9.4). This procedure uses a Metropolis-Hastings Algorithm, and I used all the standard default settings to make replication easier. I inspected available diagnostic information reported in SAS output, and visual analysis of trace plots and statistical tests, such as Geweke Diagnostics, and found no evidence my chain had not properly converged. I used a burn-in of 1,000, sampled 500,000 cases, and then thinned my posterior sample by a magnitude of 10 to reduce any autocorrelation. My posterior samples for effect sizes, under Jeffreys prior, were 47,690.9 and 50,000 for special education and general education, respectively.

Simple Difference of Means. I used a Bayesian version of the dependent means *t*-test to evaluate the credibility of possible effect sizes given the sample. I used the following equation:

$$y_i \sim N(\mu_Y, \sigma_Y^2)$$

where y_i is the difference in reading gains of the i^{th} matched pair of schoolchildren, and values of Y are normally distributed with mean (μ_Y) and variance (σ_Y^2). To estimate the model, I used a normal-shaped likelihood function for my analysis rather than a *t*-shaped one; briefly:

$$f(y_i|\mu_Y, \sigma_Y^2) = \phi(y_i; \mu_Y, \sigma_Y^2) \text{ for } i = 1,2, \dots 35, ,36$$

My choice of normal-shaped likelihood function was supported by descriptive statistics.

In Bayesian analysis one must specify priors for parameters. Jeffreys priors produce posteriors invariant under transformations. I used such priors for my parameters, μ_Y and σ_Y^2:

$$\pi(\mu_Y) \propto 1 \qquad \pi(\sigma_Y^2) \propto {}^{1}/_{\sigma_Y^2}$$

Jeffreys priors are objective. They are based on a math formula, and they represent an idealized scientist who entertains all possible effect size as equally credible before inspecting the sample. I used this impartial stance to draw substantive conclusions in this analysis.

To set up a null hypothesis test, I used Cohen's (1988) measure of practical significance

$$\text{Cohen's } \Delta = {}^{(\mu_Y)}/_{\sigma_Y}$$

I decided to interpret values of Cohen's Δ between -.2 and .2 as being practically akin to the null hypothesis, H_0: $\mu_Y = 0$, and so this became my Region of Practical Equivlence (ROPE). Using a posterior distribution for the paramter Cohen's Δ and my ROPE, I could compute the probability this special null hypothesis is true given my sample (Kruschke, 2011).

Scientists are rarely impartial towards the null hypothesis. But they, of course, differ in their judgements about the crediblity of the null hypothesis, and so I ran the analysis several times using different priors. These priors represented different stances scientists might take towards the null hypothesis.

Cynical Scientists. They posited the null is likely. They conjectured there was likely no efffect or only trival effects ($-.2 < \Delta < .2$). To them, the effects of natural proportioning and clustering on reading scores is neglitiable, and they zealously dismissed conjectures of even a small effect size as proposterous, $\pi(\mu_G) \sim N(0, 40^2)$ and $\pi(\mu_E) \sim N(0, 40^2)$.

Skeptical Scientists. They posited small effects ($.3 \leq \Delta \leq .5$) as likely. But they still fully entertained the null hypothesis too. Anything other than these, however, they regarded as wishful thinking, $\pi(\mu_G) \sim N(120, 40^2)$ and $\pi(\mu_E) \sim N(120, 40^2)$.

Optimistic Scientists. They championed modereate effects ($.5 \leq \Delta \leq .8$). But, of course, they also entertained small or large effect sizes too. They give no credence, whatsover, to the null, $\pi(\mu_Y) \sim N(195, 40^2)$ and $\pi(\mu_Y) \sim N(195, 40^2)$.

Credulous Scientists. They barely tolerated the null hypothesis, and conjectured that a large effect size was the best guess, ($.8 \leq \Delta$). Indeed, they ranked bigger a effect size as always more probable than smaller effects, $\pi(\mu_G) \sim N(310, 40^2)$ and $\pi(\mu_Y) \sim N(390, 40^2)$.

Results

Below, I present results from my main analysis. Difference scores in annual reading gains were computed using the following formula:

$$Y_{Dj} = Y_{Cj} - Y_{Nj}$$

where Y_{Dj} represents a difference score in the j^{th} group, Y_{Cj} the student in a clustered classroom, and Y_{Nj} students in a naturally proportioned classroom. Figure 12 depicts the distribution of gain score differences between matched schoolchildren in general education and special education, and both of these sets of scores approximate normalcy.

The null hypothesis is that the effect of the studied student grouping models – namely, clustering and natural proportioning – on reading gain scores is negligible The difference scores ($\bar{Y}_{DG}$=31.8, $S_{DG} = 309.9$) among schoolchildren in general education ranged from a minimum value of -391 to a maximum of 1019. A visual inspection of Figure 12 seems to favor the null in the case of general education. The difference scores ($\bar{Y}_{DS}$=-156.5, $S_{DS} = 390.6$) among students in special education scores, however, seem less favorable to the null. In the next section, I use a Bayesian *t*-test to formalize these inductions with inferential statistics.

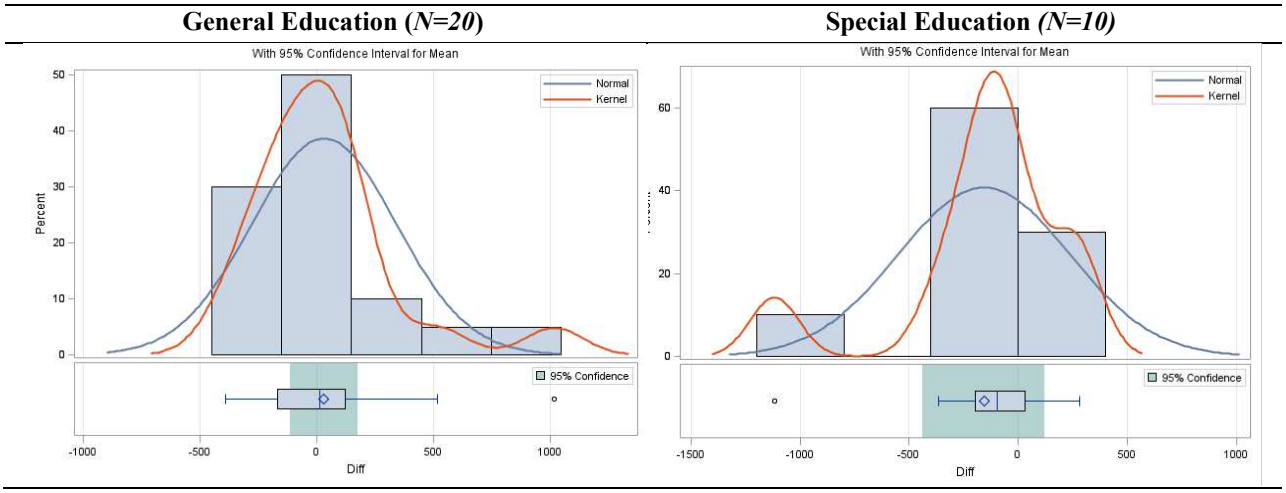

Figure 12: *Description of Differences in Reading Gain Scores*

What happens to Schoolchildren in General Education?

Bayesian scientists, at least ones using Jeffreys priors, induce the null is the most credible given the sample, and they judge posits of moderate or large effect sizes to be too incredible to be live options. They gave the null a posterior probability of .58, but still entertained the option of a small effect (p=.38). Figure 13 shows posterior distributions for μ_{DG} and Cohen's Δ. Table 6 summarizes the posterior probabilities of a representative sample of Bayesian scientists. Despite the tiny sample (*N*=20), a comparison of their posteriors shows a (subtle) convergence towards a cynical stance. Credulous Bayesians, for example, would now posit moderate rather than large effects after studying this sample; optimistic Bayesians, small rather than moderate; and so on. This pattern in posterior distributions supports the null. The DIC statistic (smaller values are better) also suggested the data is most consistent with a cynical stance.

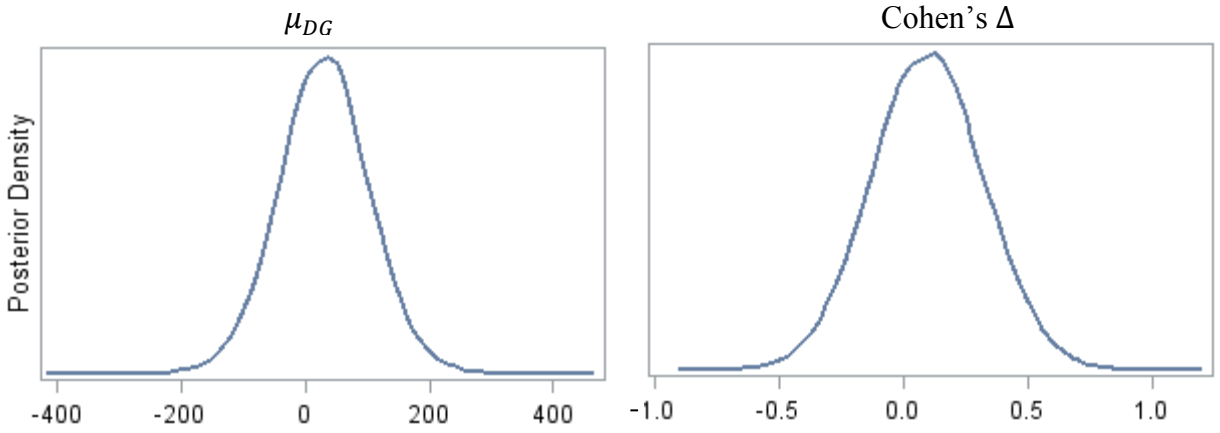

Figure 13: *Jeffreys Posterior Distributions for General Education Parameters*

Table 6: *Different Posterior Distributions for Cohen's Δ in General Education*

				Hypotheses		Probable Effect Size		
Analysis Method	Probability	DIC	Mn (95% HDI)	Null	Alternate	Small	Medium	Large
Jeffreys Priors	Prior	--	--	--	--	--	--	--
	Posterior	289.155	.10 (-.34, .54)	.58	.42	.38	.04	.00
Cynical Stance	Prior	--	--	.88	.12	.12	0	0
	Posterior	287.579	.02 (-.16, .21)	.96	.04	.04	.00	.00
Skeptical Stance	Prior	--	--	.09	.91	.71	.19	.01
	Posterior	288.495	.31 (.09, 54)	.17	.83	.78	.05	.00
Optimistic Stance	Prior	--	--	.00	1	.24	.62	.14
	Posterior	290.666	.47 (.24, .70)	.01	.99	.60	.38	.01
Credulous Stance	Prior	--	--	.00	1	.00	.18	.82
	Posterior	296.229	.67 (.44, .92)	.00	1	.08	.77	.15

What happens to Schoolchildren in Special Education?

In this case of school children in special education, Bayesian scientists, at least those ones using Jeffreys priors, induce the alternative hypothesis to be 4 times more probable than null given the sample, and they judge smaller effect size posits to be more credible than larger. Figure 14 provides a visual depiction of their judgments. But, of course, they still entertain the null as a live option. They need larger samples to fully eliminate it. Table 7 (like Table 6) summarizes the posterior probabilities of the different groups of Bayesian scientists of interest, and, despite much uncertainty and disagreement among groups, a convergence towards the alternative is evident. All scientists, for example, downgraded the credibility of both of the extreme hypotheses, null and large effect size. This provides inductive support for the conclusion natural proportioning had a bigger effect than clustering.

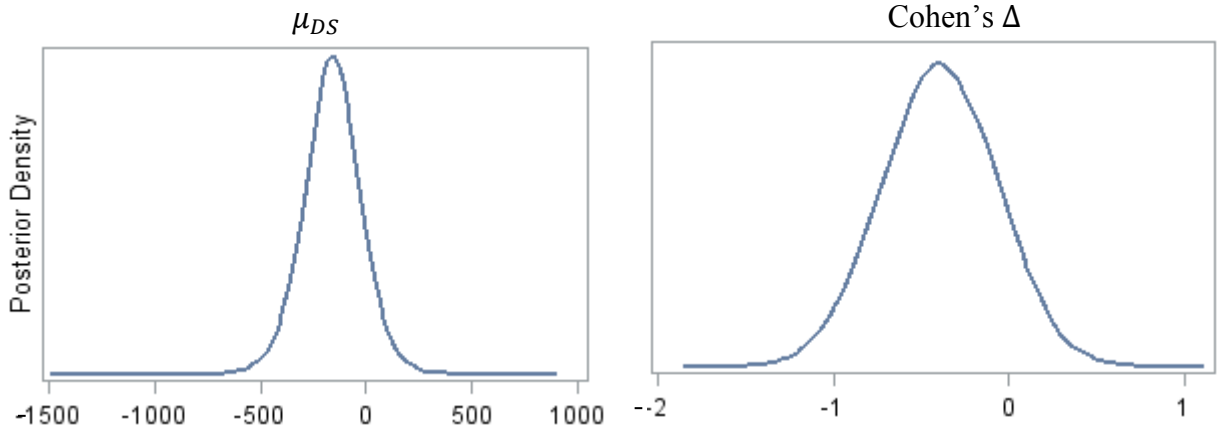

Figure 14: *Jeffreys Posterior Distributions for Special Education Parameters*

Table 7: *Different Posterior Distributions for Cohen's Δ in Special Education*

				Hypotheses		Probable Effect Size		
Analysis Method	Probability	DIC	Mn (95% HDI)	Null	Alternate	Small	Medium	Large
Jeffreys Priors	Prior	--	--	--	--	--	--	--
	Posterior	150.536	-.39 (-1.03, .53)	.25	.75	.39	.26	.10
Cynical Stance	Prior	--	--	.95	.05	.05	0	0
	Posterior	149.984	-.03 .-24, .14)	.94	.06	.06	.00	.00
Skeptical Stance	Prior	--	--	.24	.76	.70	.06	.00
	Posterior	148.783	-.32 (-.57, -09)	.17	.83	.75	.08	.00
Optimistic Stance	Prior	--	--	.00	1	.25	.58	.17
	Posterior	149.214	-.49 (-.78, -.21)	.00	1	.24	.62	.14
Credulous Stance	Prior	--	--	.00	1	.02	..25	.73
	Posterior	149.962	-.72 (-1.07, .-.38)	.00	1	.03	.40	.56

Posterior Predictive Checks

Before moving into the discussion section, I briefly turn to the subject of model fit. I used a posterior predictive checking procedure to assess my models (Lynch, 2004). To check the quality of my joint posteriors, I sampled parameters 500,000 times and then simulated samples from models specified with the sampled parameters. I kept track of the frequency of the occurrence of minimum (min), maximum (max), mean (mean), and standard deviation (sd) in the simulated data sets. Finally, I computed Bayesian *p*-values to formally determine how well my simulated samples matched up with the real one. Bayesian *p*-values, unlike classical ones, should be near .5 and high or low values are troublesome. Figures 15 and 16 depict the results.

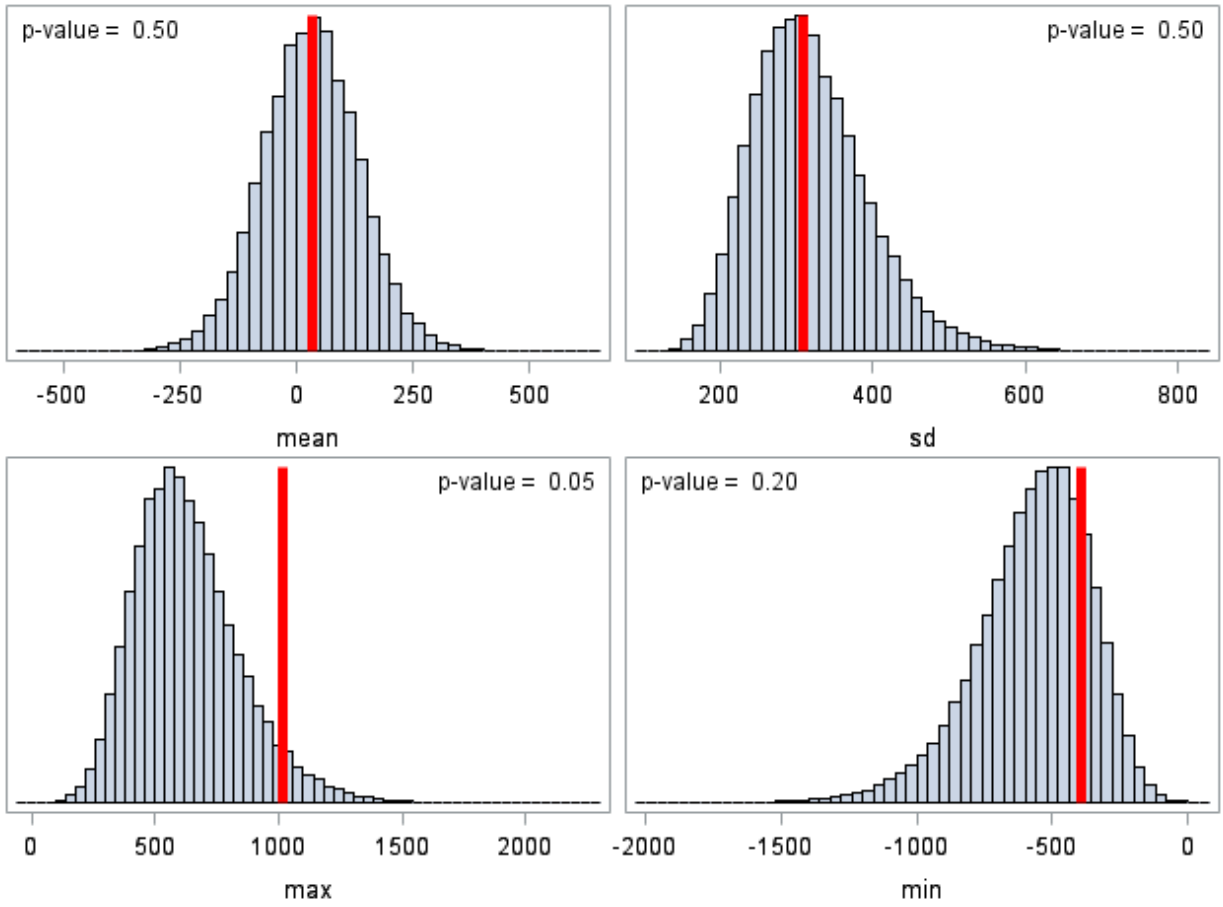

Figure 15: *Posterior Predictive Distributions for General Education Model*

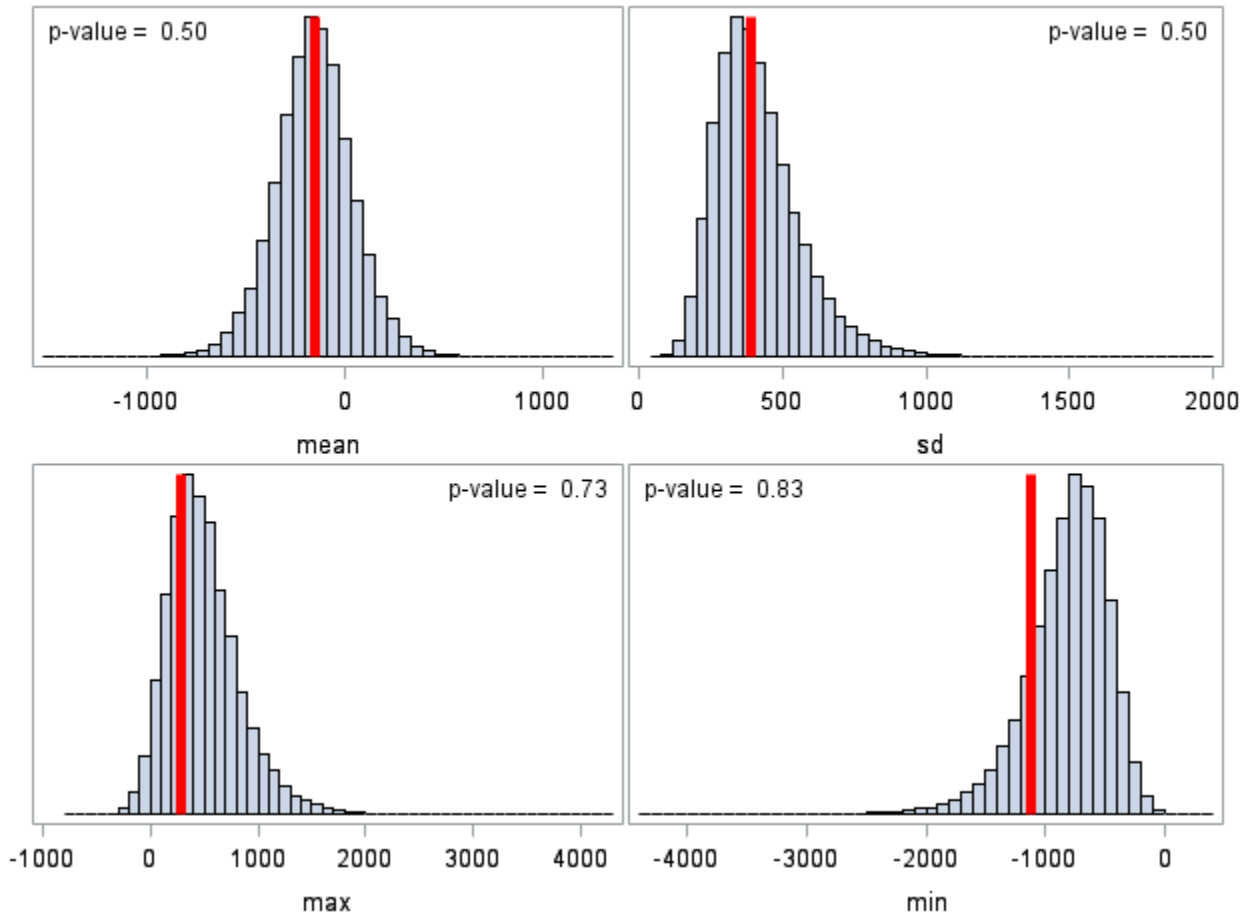

Figure 16: *Posterior Predictive Distributions for Special Education*

Bayesians avoid the classical "reject/fail to reject" algorithm for interpreting *p*-values, but for the sake of tradition I wanted to see Bayesian *p*-values between .95 and .05. I focused on the minimum and maximum values rather than mean and standard deviations as replicating them is a tough task. I was pleased then with the overall fit of my model for schoolchildren in special education. But the small *p*-values for some sample statistics for the obtained posterior distributions in the case in general education was a bit alarming. It is unclear, however, why they appeared inconsistent with data. It is most likely a problem with my model specifications. The implication: my model for students in special education nicely fits the data; my model for general education lacks such a desirable fit: its fit is uncomfortably tolerable.

Discussion

Effects of Student Grouping Methods in General Education

The results of my analysis suggest it is more probable than not that nothing of interest happened to studied youth in general education – at least, in regards to reading outcomes, as measured by annual gain scores on state reading tests. The null hypothesis is the most credible one given my choice of prior and sample evidence. It states that the differential effect of the two grouping methods on annual reading gains is negligible for all practical purposes. But a posterior predictive check found the model was not quite consistent with the data, and this suggests further elaborations on the model are needed to be certain.

My conclusions about the high probability of the null, of course, rest on a small sample. The large widths of the posteriors and Bayesian 95% HDIs show all conclusions are uncertain. Researchers who wish to gain more certainty, however, can use this study to improve their future parameter estimates. They can use the posterior quantities reported in this study to set new priors and gather a new sample, then they can compute a new posterior yielding more certain estimates. Given a steady flow of incoming data, the math guarantees Bayesian posterior distributions will generally converge. This is one of the bonuses of using Bayesian inference.

As expected, prior choice matters. Sample sizes were too tiny to overwhelm the prior's influence on the posterior. Instead, the posteriors represented a balance of sample information and prior judgments. Credulous scientists, for example, will not be persuaded from the data the null is true. Arguably, this analysis simulates the behavior of real scientists. When the evidence is pithy, scientists do not reach consensus. Nonetheless, it was reassuring to see convergence in posteriors towards the null. Jeffreys prior can be interpreted as the group's 'impartial' scientist, and its results would best approximate classical analysis.

It is more probable than not, given this analysis, that something of substantive interest happened to reading gain scores in special education. Scientists, using Jeffreys prior, induce natural proportioning probably outperformed clustering. One cannot, of course, generalize inferences outside of the study, but so what? One of the main purposes behind intervention research is to detect effects, and prompt further research into the matter. If we can isolate the mechanism producing the effect then we may be able to amplify their effects (House, 1991). The implication: more research is needed to clarify and understand why natural proportioning may have produced higher results than clustering.

It is difficult to reliably compute parameter estimates from small samples with certainty. Bayesians can sometimes do this if they have a lot of background knowledge at their disposal to inform their priors. But such was not the case here. Notwithstanding my analysis did yield some interesting results. I induced, for example, natural proportions *probably* worked better than clustering. Moreover, even if the requisite sampling conditions had been met for the standard classical analysis procedures (they were not), I still would have preferred Bayesian analysis because it supplied more information about unknown parameters.

As expected, classical *p*-values did in fact find mean differences to be insignificant, even for schoolchildren in special education. There was too much uncertainty, given such tiny sample, to eliminate the null hypothesis, even if there really was a small effect. Yet, by using Bayesian analysis, I refined my judgment about the credibility of the null hypothesis. In this case, I found it to be very unlikely while still being a live option. But, again, in this study I was stuck with Bayesian analysis for good or worse. The logic of classical inference was inadmissible here because I used judgment samples and propensity score matching (see Chapter Two).

Interpretation of Findings

If you know one child with Autism, then you know one child with Autism (Hacking, 2009). One cannot presume, for example, that if this inclusionary classroom placement worked well for a cohort of students with Autism in the past then the same will be true this year. It probably will be, of course, but only trial and error can demonstrate it for sure. And sometimes too many contingencies are thrown into the mix to make any confident predictions about the 'fit' between schoolchild and classroom, and the amount of unanticipated interaction effects probably accrues when more and more students are clustered into the same place.

Recall, McLeskey and Waldron's (2000) main argument for clustering depended on the supposition school administrators (or their informants) could predict – or, at least, forecast - which classroom had superior support systems. Otherwise, there is no reason to cluster students in such classrooms. But, given the chaos-effect, such forecasting may prove perilous despite much planning and oversight. Success here is always partially a matter of luck. Some people, of course, will be better at anticipating future events than others. I conjecture McLeskey and Waldron had 'the gift,' and this then explains why I could not replicate their success with the clustering method (McLeskey & Waldron, 2000).

One advantage of the above interpretation of my study findings is that it makes testable predictions. My interpretation entails, among other things, that there should be more variability in learning outcomes across schools using clustering rather than natural proportioning. Its effect will sometimes be greater than natural proportioning, sometimes lesser: it all depends on the accuracy of the forecast. To refute this conjecture, researchers could use a more elaborate HCT design involving multiple schools, and a multilevel analog of a Bayesian *t*-test to estimate overall variability between schools taking into account grouping method.

Limitations

This study has several limitations. One limitation was the utilization of a pre-existing data set. The data set had a relatively small sample size. It was, therefore, impossible to make precise point and interval estimates. I also could only glean information about students who took state reading tests. The data set also never specified what percentage of the day students with disabilities actually spent in inclusionary classrooms, and even more worrisome, from an analytic perspective, was I did not know how many students were clustered into the same classroom. If there was a massive amount of clustering, then the statistical model should have taken this into account. But, despite all these data set limitations, the sample did provide useful details about the null hypothesis' plausibility, and functions as a nice first approximation.

A threat to my study's interval validity is the omitted covariate. This is always a threat in education research, and nothing can infallibly shield against it. Even one omitted variable can shipwreck our best propensity score matching efforts. The inclusion of good covariates, after all, makes this threat less and less probable. Readers then are urged to double-check my covariates, and draw their own conclusions about its adequacy for our purposes. In regards to my study's external validity, one should not extrapolate effects beyond the 60 cases examined in this study. I traded off my external validity to gain higher internal validity.

Yet another limitation of this study was the lack of a conventional fidelity check. A fidelity check would have ensured that studied schools really implemented the requisite student-grouping method. In some ways, the credibility of my findings is interlinked with the credibility of Hoppey, Black, and Michelson's (2015) findings. Their study substitutes as my 'fidelity check' as they visited the school sites, observed classrooms, and interviewed key participants. Perhaps, this is an example of the necessary '*qual-quant* dialogue' among researchers.

Implications

My study has implications for both practitioners and researchers. This study reinforces the conviction that placement decisions matter. One can enact inclusion using natural proportions or clustering methods, but natural proportioning outperformed clustering in the case of special education students – at least, in regards to reading skill acquisition. This suggests the natural proportions method should be the default method. In other words, use natural proportioning rather than clustering when both methods are options and the final decision makers have no strong rationale for using clustering rather than natural proportions. An adequate rationale in this case, for example, might be natural proportions did not work out well last year.

Scholars should generate theories explaining what difference natural ratios might make on learning outcomes. This study evidences it can indeed have a bigger effect than clustering. Theorists should then try to identify under what conditions this may or may not be true. It is conceivable, for example, a lower percentage of students in special education in a classroom makes it easier (less stressful) on inexperienced instructional faculty, and so under such conditions produces bigger effects.

This study then warrants future investigation. Future researchers may try to replicate my findings with other outcomes of interest, such as math or writing. Researchers should also entertain the possibility there may be important effects on non-academic outcomes of interest too. I wager it would be worthwhile to replicate this study at a secondary level as well, and investigate whether these findings hold across the different disability categories. Moreover, it might also be worthwhile to include students with disabilities on alternate diploma options. Finally, future studies can yield more certain estimates of effect sizes, assuming they used bigger samples or more Bayesian analysis on new samples of similar sizes.

Final Remarks

Bayesian analysis provided a useful alternative to classical analysis in this study. Bayesian analysis lacks some of the luster of the objectivity of classical *p*-values. Everyone using a classical *p*-value, for example, will reach the same conclusion given the same sample. But Bayesian analysis compensates for this loss of objectivity with its own version of objectivity. It gives scientists with different priors a formal method for reaching consensus in their posteriors on matters of importance.

Some methodologists are confident everyone will be doing Bayesian analysis in the foreseeable future (Kruschke, 2011). I am skeptical about this forecast, as they have been saying this forever. But with the availability of new software programs, like PROC MCMC (SAS 9.4), may make a lesser version of their predictions finally come true: Bayesian inference may join and come alongside classical inference, as a specialized tool in the researcher's toolkit. This possibility is realizable for the first time ever as researchers without specialist training in statistics can now do Bayesian inference using computer programs, like SAS software. It is impossible then to say for sure, but there may be a coming surge of interest in Bayesian methods.

CHAPTER FIVE

DISCUSSION

In this chapter, I do my best to 'tie' everything in my dissertation together, and elaborate on why previous chapters, taken together, make a contribution to the knowledge base regarding the problem occupying this dissertation.

Summary of Dissertation

Two opinions prevail among statisticians about priors: they are assets, they are liabilities. Each opinion, of course, is correct. Some priors enlighten posteriors; others distort them. Either way, they will influence posteriors especially when sample sizes are small. Bayesians are risk-takers, who must learn to trust their priors within set limits.

My dissertation investigated why priors belong in a romantic science of education, and how we can maximize their rewards, minimize their risks. Classical statisticians play it safe, and omit priors. Classical inference is needier than Bayesian, however, and the sampling conditions it requires are scarce in education science. My argument in Chapter Two entailed, that Bayesian methods, compared to classical, were more adaptable to conditions in education.

Bayesian methods offer alternatives to classical procedures. They use judgment samples and historical controls. They cope with probability samples and random assignment too. All that matters to Bayesians is the visible data. I am not at all suggesting by this that classical inferences are inferior to Bayesian ones. Instead, I am arguing that it is wiser to reserve classical inferences to studies with the appropriate sampling conditions to preserve their logic. Absent these conditions, Bayesian inference almost becomes a practical necessity in education research.

When Bayesian methods are applied in research, prior selection is usually controversial. Perhaps, the prior is obvious because of an established consensus, but not likely. I wager prior setting in education will usually require some hard thinking. Consequently, in Chapter Three, I queried about what methods researchers can do to set defensible priors in research. Specifically, when conducting *t*-tests to infer causation in small pilot studies.

Imagine a researcher sets up a minuscule pilot study to test an intervention. She wishes to use statistically significant findings to persuade funders to sponsor a massive follow-up study. She is lucky, however, and has at her disposal a probability sample, and was able to randomly assign all sample members to control and intervention groups. She runs the experiment, and then decides to use a parametric independent means *t*-test to evaluate her findings. Classical inference is a live option, but her sample size is small and she anticipates statistical power will be an issue. One possibility to overcome this limitation is to go Bayesian instead. But she needs to select an appropriate convenience prior to conduct the Bayesian version of the *t*-test.

I then argued in Chapter Three it does not take much knowledge at all to set up a default prior for a Bayesian *t*-test. She can, for example, use only her knowledge of possible effect sizes to specify a prior. This single prior represents a balanced mix of distinct priors for each possible effect size of interest, and then it defaults to whatever prior in the mix is most consistent with the sample data in the analysis phase. This could be a respectable convenience prior.

I used Monte Carlo simulation methods to assess the performance of the Bayesian version of the *t*-tests compared with the usual classical version under the same conditions. I discovered Bayesian *t*-tests had ample power. Moreover, they outperformed their classical cousins in regards to their control of Type I and Type II error. These findings suggest Bayesian *t*-tests can be appropriate for significance testing in education.

Convenience priors, like mixed priors, minimize the risk Bayesians will use mischievous priors to reach unwarranted (albeit, desirable) conclusions. They are, for example, free of any favoritism. Sensitivity analysis, moreover, can establish how dependent the posterior is on the selected prior. If multiple priors produced the same (or approximate) posterior, one can infer the posterior was not too dependent on the selected prior. This convergence, thus, reassures readers Bayesian analysis is not bias run amok.

Finally, in Chapter Four, I illustrated some of the important mechanics of Bayesian inference with a true story. I was asked by school district personnel to evaluate whether a school district-level initiative really had a positive and significant impact on state test scores for a subpopulation of interest. They gave me a small sample of test scores from students placed either in control and intervention conditions – or, at least, placements somewhat approximating these condition-types. But it was not a probability sample, and I also had to use propensity score matching to try to balance groups on extraneous variables.

I decided to conduct a *t*-test. Given sampling conditions, I decided to use the Bayesian version of this statistical test to preserve the logical integrity of my statistical inferences. Even if requisite conditions for classical version of the *t*-test had been met, however, I still would have preferred the Bayesian version of it to improve my statistical power. Power was an issue in this evaluation study because of a limited small sample size.

I felt comfortable setting priors. I had access to a team of qualitative researchers who had visited all the relevant school sites. They had visited the sites on several occasions, and had interviewed teachers and principals about the new initiative. They also used qualitative analysis techniques to enhance their judgments about the schools. This possibility made a Bayesian *t*-test an even more attractive option for me as I had informants with qualitative knowledge.

Together all three chapters in this dissertation build a cumulative case for using Bayesian methods more frequently in special education. There is a wealth of clinical expertise in special education, and researchers can use their favorite qualitative method to elicit this information. The prior is the natural vehicle for utilizing this rich qualitative knowledge in quantitative analysis. Bayesian statistics is computer intensive, however, and this made it infeasible in the past. But today, as my dissertation demonstrates, we can use computers to solve the otherwise intractable posterior integrals and proceed in Special Education with Bayesian inference.

Tying Up Loose Ends

Before closing out this concluding chapter, I wish to make a couple of observations about important issues that never arose when considering the contents of each chapter individually. These issues include (a) Subjective and Objective Bayesian takes on posterior probabilities, and the nature of epistemic-type probability, (b) a plausible account of how assessing epistemic-type probability is possible for mere mortals, and (c) how I can build a future research agenda out of my dissertation findings. To facilitate the rest of my discussion in this chapter, I divided my comments into three distinct parts. Each part is devoted to one of the above issues.

Part I

Two Versions of Bayesian Statistical Inference

Credence-type probability comes in two flavors, *epistemic* and *psychological*. Epistemic-type probability is prescriptive. Given our scientific evidence, for example, general relativity theory is probable. This is an example of epistemic-type probability. It carries normative weight, and surely people err if they disagree with it. Psychological probability is descriptive. A fan says it is probable her sports team will win the game despite much counter-evidence. The type of probability is autobiographical, and no one else is obliged to adopt it.

Subjective Bayesians claim psychological-type probabilities should dominate priors and posteriors. The only constraints on psychology-type probability, of course, are the axioms of the probability calculus. Outside of these minimal constraints, they will permit researchers to set up priors as they see fit. The only caveat is they recommend researchers avoid prior probabilities of 1s or 0s. These prior probabilities, after all, cannot be modified by evidence.

Objective Bayesians, in contrast, are maximalists. They hold probabilities fit for science are all about epistemic-type probability – not psychological-type. This implies priors should, in addition to their compliance to the probability axioms, carry some normative force. They want their priors, in other words, to represent the opinions of the wise rather than fools. But since it is impossible to operationally define epistemic probability, they must try to discern its presence in their priors by using their best judgment.

I confess my dissertation silently adopted the Objective Bayesian stance. Throughout my dissertation, for example, I took it for granted posteriors and priors could be reasonable or not, and I engaged in the search for adequate convenience priors. Subjective Bayesians, on the other hand, would wonder why I was being so fussy. They think we should just pick a prior and move on. All they care about is whether one's prior is existentially true to oneself.

Objective Bayesians, like me, however, want more than 'authenticity' from our priors. We want epistemic-type probability too. Specifically, we want our posteriors to model – and we do concede they only model – judgments of an ideal rational agent about the epistemic-type probability of propositions given samples. All models, of course, are imperfect. But they can still inform us about what ideal rational agents might or might not induce in similar circumstances. Priors then are fancy model components (Gelman & Shalizi, 2013), and we should care about them because they help us to gain posteriors 'modeling' the opinions of 'the wise.'

Without operational definitions of epistemic probability, of course, Objective Bayesians cannot guarantee their posteriors encapsulate wisdom. In any given application, posteriors, for all they can prove, may contain only prejudice. Without operational definitions, they must trust their own cognitive faculties to evaluate its presence in posteriors. But classical statisticians must do the same. They, as I argue below, must evaluate the epistemic probability of their sampling distributions. The following thought experiment supports this bold claim.

Thought Experiment

Imagine a statistician wishes to induce causation within the classical system, and only has access to a small convenience sample (N=6). Without getting lost in the details, she sets up an elementary version of the null hypothesis: there is no intervention effect. This null, obviously, is non-parametric. It, unlike the more familiar parametric version of the null hypothesis, refers to no invisible parameters. To test this austere null, it suffices to randomly assign the 6 participants into two subgroups of 3 (i.e., control and intervention), run the experiment, and then evaluate a test-statistic of interest. Perhaps, in this case, the absolute mean difference between control and intervention group can suffice as her test-statistic.

The logic of the classical significance test in this case is elegant. There are 20 possible grouping arrangements, and she randomly selected one of these before running the experiment. If the null is really true, however, it should not matter, even after the experiment is over, which grouping arrangement she actually used to compute the test-statistic. All participants should, according to the null, be interchangeable. To test whether this is the case, she can compute the test statistic under each possible grouping arrangement, and count up what proportion yielded test statistics weirder than the obtained one. If it is only 5% (or whatever significance level she picked before observing the results) she can reject the null.

On the surface, it seems the requisite sampling distribution here is fully determinable by the math. The experimenter placed 3 out of 6 people (A,B,C,D,E,F) into the intervention group. If she listed these possible combinations out one-by-one (such as ABC, ABE, and so on) she will eventually count up with 20 such combinations in total, $\binom{6}{3} = 20$. This is a truth of math, and it is why she carved up the sampling distribution space into 20 possibilities. Notwithstanding, she needs more then this math to justify her final sampling distribution selection.

The math underdetermines the requisite sampling distribution. Other people could reduce the total count of 20 possibilities to 19, for example, by replacing two of her possibilities with one singular possibility. Example, we may replace both her "ABC" and "DEF" events with the single event "ABC or DEF." This new sampling distribution then has only 19 events in it in total. But, of course, there is no reason to stop gerrymandering at 19. Someone can keep on doing this until they obtain sampling distribution producing a likable *p*-value.

We judge people partitioning the sampling distribution into anything other than 20 events as making a cognitive mistake. But they surely are not erring in their math. The math is neutral in regards to the issue of gerrymandering events. We must instead appeal to epistemic-probability to counter these illicit moves. We need to say, in other words, that the wise person will carve up sampling distribution spaces this way rather than that, and to do otherwise is foolish. This argument is similar to what Objective Bayesians must argue in the case of priors.

Classical statisticians then, like their Objective Bayesian colleagues, must make appeals to epistemic-type probability. But what warrant do any of us mortals have to even try to assess epistemic-type probability? It is an entity beyond the comfortable realm of operation definitions. But the statistical enterprise seems to presume we can validly assess it, even if only imperfectly. The success of inferential statistics shows this presumption is plausible.

Summary

The Bayesian family encompasses both Subjective and Objective Bayesians. Subjective Bayesians make sense of posteriors with psychological-type probabilities; Objective Bayesians with epistemic-type probability. Subjective Bayesians are minimalist, and they do not expect too much from their posteriors. Objective Bayesians, in contrast, want their posteriors to encapsulate epistemic-type probability. But assessing posteriors resemblance to epistemic-type probabilities naturally presents a difficulty for Objective Bayesians as they are intangible.

Objective Bayesians, however, share this difficulty with classical statisticians. They too must assess epistemic probability. They need to sensibly carve up the sampling distribution space in ways overlapping with epistemic-type probability. It may be – and often is – an easier task to carve up a sampling distribution than a prior, but this difference is one of degree rather than kind. It is somewhat unfair then to single out priors, and treat them like bias magnets.

The success of the inferential statistics enterprise, including the inferential victories of Objective Bayesians and classical statisticians, clearly show that we mortals do in fact have the astonishing capacity to assess epistemic-type probability. My dissertation is a demonstration of this. But without an account of how we are able to do it, the whole issue remains murky. Anyone who is distraught by this mystery, of course, can become a Subjective Bayesian, and no longer need to worry about it. But this is giving up, and it would be a bit hasty.

In the next section of my dissertation, I will with much trepidation enter into the highly contested philosophical terrain of the problem of induction. This is because the problem of assessing epistemic-type probability is really at the core of the modern riddle of induction. I deem this adventure to be a necessary evil. It is a pathway to understanding why statisticians can indeed (albeit, fallibly) assess epistemic-type probability.

Part II

Assessing Epistemic Normativity in Inductions

Imagine a primordial group of hunters encounter a couple of mammoths for the first time. They are all soundly asleep. They entertain two possible hypotheses about them, 'All mammoths in the world sleep' and 'These mammoths perpetually sleep until death.' They all deem the latter posit absurd, the former plausible. They make such bold judgments notwithstanding the fact each is compatible with the evidence. How did they (and their living decedents) detect epistemic-type probability in inductions apart from evidence? This is a grand riddle. It took some extraordinary philosophical ingenuity on the part of some great philosophers, like David Hume, just to recognize the riddle existed at all.

There are several ways of formulating the riddle of induction, each with its advantages and disadvantages. In his celebrated book, *Fact, Fiction, and Forecast*, Nelson Goodman (1955/1979) gave us the modern version of the problem of induction. The available evidence underdetermines the 'natural' induction for the occasion, yet despite this unavoidable ambiguity, we humans can usually select the winner in our ordinary daily life and science.

On Goodman's analysis, we do so because habit guides our inductive meta-cognition, not logic – a conclusion that, despite Goodman's protests, imperils the status of induction as a rational tool of inference.

The philosophical move I would like to make in response was anticipated in a short response to Goodman written by the philosopher Donald Davidson in the 1960s. In *Emeroses by Another Name*, Davidson (1966/2006) sets forth an intriguing counterexample to Goodman's analysis showing, among other things, why habit cannot be the right answer. But I want to build on Davidson's insights here too and offer a clue to the solution to this grand riddle.

The New Riddle of Induction

In Goodman's nomenclature, projections (or inductions) are merely general descriptions posited as bridges for travelling from determined cases to undetermined cases. Projections are confirmed by their positive instances, disconfirmed by their negative instances, and neither confirmed nor disconfirmed by their underdetermined cases. Goodman identifies several preconditions for projectibility. A projection occurs when (a) negative instances are absent, (b) *undetermined* cases are present, and (c) positive instances are counted as lawlike instantiations. This last remark needs elaboration.

Regardless of a projection's truth or scientific importance, Goodman (1979) argues positive instantiations of it should seem "lawlike" – i.e. capable of sustaining counterfactuals and subjunctives. To illustrate his point, consider a hypothesis stating everyone now in a given room has three sisters. Suppose we talk to a woman in this room and find out she has three sisters. She is a positive instance of the sibling hypothesis, but - absent a hunch this should continue to be so - we still have no reason to think the next person we encounter in the room will have three sisters too. Karl Hempel's notable Black Raven Paradox results when this "lawlike" quality of virtuous projections is left out of our reckonings (see Goodman 1979, p. 73).

Goodman contends Hume's analysis of induction survives all critical scrutiny. As a good Humean, Goodman thinks our meta-inductive device detects the epistemic normativity of cogent inductions by sorting observed positive instances of general statements into "lawlike" and "non-lawlike." It does this without help from Reason. Hume (1739/1999) aptly puts it, "That the sun will not rise tomorrow is no less intelligible a proposition, and implies no more contradiction, than the affirmation, that it will rise" (p. 108). Our assessments of epistemic-type probability, likewise, are not (and cannot be) products of Reason.

Goodman's skepticism towards induction is a little bit deeper than Hume's, if that is possible. Hume concedes we humans anticipate what will happen in a collision of billiard balls. He presumes we passively learned to "see" such necessary connections in the regularities of experience. In contrast, Goodman conceives of us as active learners. We do not sit idle until external prompts cause us to induce. Our minds are too proactive for that. Trying to make sense out of regularities in experience, we wildly project necessary connections everywhere in nature and then try to sort them out afterwards. Goodman only wonders why we feel only a certain class of possible inductions fitting the evidence hold epistemic normativity.

His proposal to this riddle: We fixate on only projections couched in customary terms. To make his case, Goodman begins with the premise ontology is relative to some descriptions of things. Different descriptions of things offer incommensurate ontologies, and each such ontology implies different law-like connections. But, of course, it is natural for one description of things to become dominate in a community of speakers. Projections framed in speech too far outside of these customary ontological molds then seem funny sounding (i.e. they appear to lack any epistemic warrant to these speakers).

In Goodman's analysis, what counts as a passable induction today is only a matter of contingency. Inductions dressed in unfamiliar terms sound strange to us. But nothing prescribes in advance what will sound strange to speakers. Languages are relatively stable, but languages evolve over the long haul. Moreover, humans may be psychologically "hard-wired" to think and talk conservatively about things, but this was not an inevitable outcome of evolution. Perhaps, given a different evolutionary history, humans would be more liberal in their linguistic habits. So, Goodman's conclusion about our meta-inductive palate radically alters the "official" story about induction as the *logic* of discovery in science (Bacon, 1620/1961).

Goodman's Argument

To test whether people really do show a conservative bias when judging statements to be lawlike (and so suitable for projection) Goodman devises a clever thought experiment. He starts with a general statement fit for projection because its positive instances could be lawlike:

H_1 All emeralds are *green.*

He then exchanges this statement's familiar color predicate with an unfamiliar one in order to study what happens to the lawlike status of its positive instances. He substitutes into $H_1$the color predicate 'grue' (Grue means 'green' before time *t,* otherwise 'blue'); briefly:

H_2 All emeralds are *grue.*

The statement, 'All emeralds are grue,' is not a manifest contradiction. Nonetheless, Goodman (correctly) notes people will (correctly) reject it as unfit for induction.

But a close inspections of H_1 and H_2 shows their positive instances come from the same stock of evidence – namely, green emeralds before time *t*. Reason only tells people that H_1 and H_2 cannot both be simultaneously affirmed. Nonetheless, given a green emerald before time *t,* people will say that only H_1 is supportable by the evidence. They make such a bold meta-inductive judgment because they think emerald-observations can be lawlike instantiations of H_1, but not H_2. If, contrary to all expectations, all the green-looking emeralds suddenly did 'change' color after time *t* (i.e. they were grue colored) people would attribute this to a fluke or a spontaneous miracle – not a result of something lawlike, like grue coloration.

But if neither evidence nor Reason explains why people surmise positive instances of H_1 but not H_2 to be lawlike then what is informing people's meta-inductive judgments in this matter? Goodman concludes people do not like 'grue' because it is unfamiliar to them. As Goodman explains:

> Plainly 'green,' a veteran of earlier and many more projections than "grue," has the more impressive biography. The predicate 'green' is much better *entrenched* than the predicate "grue" (p. 94).

In other words, Goodman thinks the tossup between 'green' and 'grue' is ultimately decided upon grounds of green's 'entrenchment,' or its repeated usage in color descriptions in familiar discourse about emeralds. The unfamiliarity of 'grue,' therefore, makes people distrust it over green.

To see why Goodman thinks people rely upon custom to inform their meta-inductive judgments, suppose a familiar predicate is substituted back into H_2; briefly:

H_3 All emeralds are *red.*

The substitution of 'grue' with 'red' creates a hypothesis with a different empirical significance than H_1, but people still will judge positive instances to be – in a meta-inductive sense – lawlike. I mean that in the absence of counterexamples, people would take statements to the effect that this emerald is red to be confirming instances of H_3. This seems to lend credence to Goodman's analysis that 'grue' is unsuitable for lawlike statements because it is an unfamiliar term.

Ultimately, Goodman's test depends upon the assumption that Reason is indifferent to the choice of 'green' and 'grue.' But many may wonder if Reason really is so indifferent. Perhaps, 'green' is more suitable than "grue" because of Ockham's razor - I think not. Inspection of H_1 and H_2 indicates they have the same simplicity. My fictional objector who appeals to Ockham's razor against the 'grue' predicate fails to see this because she gives too much attention to the way it has been defined in the argument. But the same argument can be presented differently to speakers whose native language employs grue rather than green. Briefly:

> **Green** means grue if examined before time *t*, otherwise bleen - where 'bleen' means blue after time *t*, otherwise grue.

Thus, it is only a bug in the English presentation of the argument that made hypotheses using 'green' or 'red' appear simple whereas the hypothesis using 'grue' complex. But, aside from accidents of language, hypotheses using 'grue' are as qualitatively simple as hypotheses using 'green' – at least, as far as Reason discerns. These facts also seem to add weight to Goodman's contention that the only germane difference between the two hypotheses is the familiarity of their color predicates to English speakers.

Consequences of Goodman's Analysis

To recap, Goodman (correctly) posits people can usually assess if an induction lacks epistemic-type force regardless of its compatibility with evidence (If people were not good at it, evolution would have eliminated the human species a long time ago). He also (again correctly) posits that only lawlike statements are capable of projection. He then uses these two insights to conduct a thought experiment designed to show custom rather than Reason guides our meta-inductive judgments. On the basis of this experiment, he argues we deem 'grue' unsuitable for projection only because it is less familiar than 'green.'

If Goodman's conclusions are warranted then important implications follow for our understanding of induction and related topics. For example, it would shed light on the elusive criterion for laws. It also completes Hume's project of dethroning Reason and putting custom in its place. As Goodman elegantly expresses: "If I am at all correct, then, the roots of inductive validity are to be found in our use of language" and when Reason conceives of a limitless number of hypotheses for the same data "the line between valid and invalid predictions (or inductions or projections) is drawn" in the linguistic sands of how the world "has been described and anticipated in words" (p. 121). Again, these conclusions, despite Goodman's best protests to the contrary, imperil the use of induction as fit for the scientific enterprise.

Davidson's Masterstroke

Previously, I consider Goodman's new riddle of induction and his own analysis of its skeptical implications towards induction. But Davidson offers us a way to salvage induction from such skepticism and, thus, restore induction's status as a rational tool of discovery in science. One recalls from the first section Goodman's thought experiment pivots on the unsuitability of unfamiliar terms for lawlike statements. Davidson, however, gives a counterexample to Goodman's general claim. I reproduce, explain, and defend it from possible rejoinders below.

Davidson responds to Goodman's claim with a counterexample (see Davidson, 2006, p. 120): Consider the following hypothesis:

H_4 All emeroses are *gred*

The predicate 'gred' is akin to the predicate 'grue,' except it states the color turns red rather than blue at time *t*. The object emerose is a little bit like these odd color predicates. It is any object that is an emerald, if examined before time *t*, otherwise a rose. The non-existence of emeroses in this world is entirely beside the point.

This hypothesis surely qualifies as a little peculiar. It is a much odder hypothesis than 'All emeralds are grue.' It is stranger because it contains not one but two foreign predicates – one for the antecedent the other for the consequent. It should be, thus, doomed for two reasons rather than one – at least, if Goodman's analysis of the situation is at all correct here. Notwithstanding, Davidson claims people intuit gred is projectable over emeroses. If true, one must conclude, contra Goodman, something else besides customary word choice informs people's meta-inductive judgments. The point of Davidson's counterexample then is to expose a fatal flaw in Goodman's analysis.

Of course, Goodman could try to get around Davidson's counterexample by saying H_4 should be judged not lawlike (In fact, this is exactly how Goodman does try to get around Davidson's counterexample in his writings, see Goodman 1979, p. 106). But Davidson anticipates such a dodge and blocks it. Here is Davidson's argument in his own words:

> Recently Goodman has claimed that [H_4] is not [lawlike]. Here I consider whether he is right. Let us pretend the following are true and lawlike:
>
> [H_1] All emeralds are green
>
> [H_5] All roses are red
>
> Then [H_4] is true, and we have good reason to believe it (p. 120).

Davidson's masterstroke is terse and dense. Perhaps the best way to elucidate his writing here is to suspend all disbelief for sake of argument and posit emeroses actually exist.

Emeroses are green emeralds that seem to us, English speakers, to magically 'transform' into red roses at time *t*. Naturally, the color of these entities changes from green into red at time *t*. But this description matches the definition of the color gred. So, by definition, emeroses are gred. Moreover, when we grant H_1 and H_5 we rightly concede the color green is projectable over emeralds and red is likewise projectable over roses. So, consistency demands, the projectability of gred over emeroses before time *t* since they really are green emeralds at this time and red roses afterwards. If this does not qualify them as gred nothing else will.

The flaw in Goodman's test is the assumption that the unit of empirical significance is isolated bits of language, called, "predicates." Davidson explains, "Goodman's test for deciding whether a statement is lawlike depends primarily on how well behaved its predicates are, taken one by one" (p. 120). In other words, Goodman thinks he is testing the behavior of predicates, controlling for everything else. This assumption entitles him to make conclusions about which predicates are suitable for lawlike statements. The problem with his procedure is it ignores the effects of semantic holism.

The Hermeneutical Circle

Logical empiricists once fondly dreamed that an unbridgeable chasm between speculative metaphysics and the natural sciences existed. To them, science was a methodology rather than a knowledge base. Their trademark thesis was that metaphysical statements were worse than false – they were literal non-sense. They defended it with the so-called analytic-synthetic distinction – sometimes called "Hume's fork" in honor of its greatest defender among history's intellectual elite.

Hume's fork stated that the meaning of all intelligible statements comes from a synergy between their logical and factual components. If a sentence's truth value depended only on its logical components, however, it was analytic (otherwise, synthetic). They also thrust foreword their famous verification principle of meaning as a test for determining whether a passable sentence was analytic or synthetic (Ayers, 1937/1952). If a sentence could be confirmed under all circumstances it was found to be analytic; if it could be confirmed in only some, synthetic. Thus, the analytic-synthetic distinction, supplemented with the verification criterion of meaning, was science's bulwark against metaphysical intrusion.

In his classic article, *Two Dogmas of Empiricism,* Quine forever blurs the lines between science and metaphysics. To achieve his objective, Quine destabilized the ill-founded doctrine of the analytic-synthetic distinction and its corollaries. He replaced them with his robust doctrine of semantic holism. Quine's sanitized empiricism entails, among other things, that it is impossible to draw a principled line between analytic and synthetic statements. Quine boldly claims, "It is folly to seek a boundary between synthetic statements, which hold contingently on experience, and analytic statements, which hold come what may. Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system" (p. 43).

Quine (1951) compares the totality of human knowledge to a field of force; it's boundary condition being experience. Conflicts with experience occasion change in truth evaluations of statements. But there are multiple ways to redistribute truth values across all the statements in the system to accommodate recalcitrant experience, and so preserve our favorite claim's credibility. This latitude entails, among other things, that in principle one has the freedom to preserve any claim about reality in our science. In a memorable quote, Quine (1951) even suggests the ancient posits of Greco-Roman style gods change weather patterns can be made compatible with all the modern empirical evidence if we make enough changes to our system of science. The warrant for atom-posits in meteorology instead of god-posits is pragmatic – not empiricist. Atom-posits can forecast the weather patterns better. In a bolder move, Quine dubiously declares all beliefs, including laws of logic, revisable.

Quine then explains our ability to assess epistemic-normativity in inductions as a by-product of pragmatism. We eliminate inductions from the competition when they conflict with accepted evidence, of course, but we also eliminate them not because we logically must but because their acceptance would irritate us too much. Interconnections among beliefs in the web cause any change near the interior to cascade throughout the web, and naturally this upsets us. We far prefer to preserve our system of beliefs than change it. It is one thing to make minor tweaks in our cognitive schema, and it is another thing all together to replace one cognitive scheme with another: one is a pleasant experience, the other painful. We, thus, desire to modify statements near the periphery and keep the interior intact. This natural desire to minimize pain and maximize pleasures explains why we sometimes discard conflicting evidence as moot and other times accept it, and so modify beliefs accordingly.

The problem with 'grue' on Quine's analysis is not that it is unfamiliar to us, but that it does not belong in emerald discourse. Quine posits humans reject 'grue emeralds' because of a naturally strong inclination to make as little revision in their systems of beliefs as possible. As Quine puts it, "A recalcitrant experience can, I have urged, be accommodated by a variety of reevaluations in various quarters of the total system; but… our natural tendency is to disturb the system as little as possible…" (p. 43). Quine's language of 'natural tendency' is a little vague, but it is a tendency born out of pragmatism rather than logical realities. We humans behave this way to minimize cognitive dissonance – not because logic demands it.

I do not accept Quine's account, but I think it provides the vital clue about what is wrong with Goodman's analysis of his own thought experiment. Goodman thought he was testing how well-behaved predicates are when substituted into single statements encapsulating atomic empirical significance. He mistakenly thought his test involved simply replacing one color predicate, green, with another, grue, in a sentence, and drew a mistaken inference from this premise – i.e., familiarity of terms mattered. But in reality, he was actually testing how well predicates behaved when substituted into new domains of discourse.

This fatal oversight on Goodman's part is what, I think, actually invalidates Goodman's interpretation of his experimental findings. And Davidson figured this out first. His suspicion that semantic holism rather than our unfamiliarity with grue was the true culprit behind our rejection of 'All emeralds are grue" led Davidson to invent his clever counterexample, "All emeroses are gred." This is because people have an aversion to making massive changes in their web of beliefs so they would not be comfortable with the hypothesis all emeralds are grue. But given a less sacred domain of discourse for 'grue'-talk, such as emeroses, people should have no problem with the funny color predicate.

The effect of semantic holism on our meta-inductive judgments is a special instance of the Hermeneutical Circle. The Platonic Socrates pointed out to Meno long ago that scientists cannot undergo inquiry without background knowledge. Otherwise, they will not know what they are looking for or when they found it. But, of course, background knowledge does not make inquiry redundant - inquiry improves background knowledge. So, there is a virtuous rather than vicious circle: inquiry without background knowledge is empty; inquiry without background knowledge is blind. Induction, like inquiry in general, is a hermeneutical exercise.

Davidson's Quinian criticism shows the problem in Goodman's analysis, but it does not move us one inch beyond pragmatism. This is desirable state of affair for pragmatists, and they can be perfectly contented with it. But the Quinian account of induction leaves something to be desired for realists, like me. Realists want the sense of epistemic-type probability to come from something more substantive than brute pragmatism. I think it is possible, however, to improve upon Quine's account of induction with virtue epistemology (Aristotle, 1985).

Virtue epistemology has no established definition, but I take it to mean something similar to Plantinga's (1993) bold thesis that knowledge (i.e. warranted belief) depends on external favors (i.e. factors beyond the mere content of beliefs). Specifically, knowledge production for mortals entails our cognitive faculties are, among other things, (a) functioning properly, (b) situated in a suitable epistemic environment, and (b) built with a competent design plan aimed at reliably outputting true beliefs. Virtue epistemologist then adds to this thesis the proviso that knowledge production requires the acquisition of certain moral and intellectual virtues, such as a desire for the truth. And, so, Quine's 'natural instinct' becomes an intellectual virtue rather than a 'pragmatic calculus.' This theory, thus, preserves induction's epistemic normativity. But, alas, a full exposition of my proposal takes us beyond the scope of this dissertation.

Summary

Objective Bayesians hold their priors to encapsulate epistemic-type probability. But the formal math fails to distinguish psychological-type from epistemic-type probability, and this makes selecting appropriate priors with epistemic-type probability difficult. But this problem manifests itself when classical statisticians carve up their sampling distributions, and it is really a special instance of the more general problem of the new riddle of induction.

Much more philosophical investigation needs to be done into the problem of induction, of course, but I think we can discard Goodman's proposal that the solution to this problem can be found with Humean-type appeals to custom, and Davidson's counterexample of gred emeroses show why the argument for Goodman's analysis collapses. It fails to properly take into account the effect of semantic holism. Semantic holism, however, ruins Goodman's analysis, but not pragmatism. It can take on a pragmatic or realist flavor. It all depends on whether we attribute our desire to preserve the web of beliefs to pragmatism (Quine, Davidson) or to confidence in the general accuracy of beliefs produced by our cognitive faculties (virtue epistemology). But I will save the debate between pragmatism and realism in education science for another day.

The operation of a hermeneutic circle between background knowledge of emeralds and samples of emeralds then has ample explanatory power to explain the grue problem. Experience of emeralds then is much richer than positivist usually to think. As McDowell (1979) writes:

> But the notion of the world, or how things are, which is appropriate in this context is a metaphysical notion, not a scientific one: world views richer than that of science are not scientific, but not on that account unscientific (a term of opprobrium for answers other than those of science to science's questions) (p. 19).

If McDowell and others are right, there is more cognitive capital in experiences of emeralds than Goodman credits. Hence, there is nothing – except undefended metaphysical scruples – that could cause one to delimit the guide of induction to the realm of custom.

Part III

My Future Research Agenda

I began this dissertation with several hunches. These included (a) education researchers are wise to use priors in special education inquiry, (b) they can elicit priors from content experts, and (c) qualitative researchers, construed as content experts with qualitative knowledge rather than methodologists, may make good informants. My dissertation findings support these three hunches, and so support a future research agenda I elaborate on below.

Implication for Philosophers

It is an unfortunate legacy of the past dominance of behaviorism over special education science that most researchers now associate statistics with positivism. The truth is much more interesting than that. Positivism is actually at odds with much of what passes for parametric inferential statistics in special education. Inferential statistics can also be parceled into several philosophies of inquiry. All equally opposed to positivism.

The Bayesian version of statistical inference is consistent with a romantic (and so non-positivistic) science of special education. What that means exactly is a job for philosophers to figure out, but at minimum it suggests that education researchers need teleological explanations. I even wager it is impossible to fully account for human disability without invoking teleology somewhere in the science. This makes qualitative knowledge indispensible.

Qualitative knowledge about special education is spread out across multiple stakeholders: teachers, principles, educational leaders, students, parents, and so on. Bayesians can use their favorite qualitative methods to mine their informants for qualitative knowledge. They can then fuel their statistical inferences with these qualitative treasures. Bayesian procedures, of course, depend for their validity on Bayesian epistemology.

The omission of Bayesian epistemology, and statistical philosophy in general, from texts on the philosophies of inquiry in education should be corrected in future texts. My study of the history of educational research has led me to believe that a great deal of educational research is conducted to demonstrate a method – doing things right - rather than to contribute to a complex area of scholarship. To curb this unfortunate tendency, researchers who plan to use Bayesian statistics should be given some exposure to Bayesian epistemology in their coursework. It would be a worthwhile project then to write a text on philosophies of inquiry encompassing statistical philosophy, including Bayesian epistemology.

Implications for Methodologist

My dissertation has implications for methodologists in education too. Methodologists are not method-technicians, but method-inventors. Their mission encompasses two responsibilities. One is to prepare and equip the next generation of education scholars to build up the knowledge base with defensible research. The other responsibility is to offer criticism of established methods, and recommend improvements or innovations. I tailor my comments to methodologists then around these two responsibilities.

In regards to the first responsibility, my dissertation suggests there is a growing need for methodologists to prepare and equip graduate students in education to harness Bayesian methods and interpret Bayesian analysis. Many graduate students in education do not have strong math backgrounds, and traditional Bayesian textbooks and Bayesian software often assume users had previous exposure to, at minimum, probability theory, calculus, and linear algebra. One task of methodologists then is to translate Bayesian procedures into a consumable format for education researchers. This is most likely to be accomplished using statistical software programs, such as SAS 9.4 (Proc MCMC), to derive posterior integrals rather than calculus.

In regards to the second responsibility, my dissertation suggests continued investigation into the possibility of using mixed priors to inform education research. The mixed prior, like empirical Bayes, is a little bit unorthodox in the Bayesian system. It violates the controversial likelihood principle, which stipulates the all sample information of interest should only come from the likelihood function. The mixed prior then cheats because it illicitly pulls sample data to inform both the likelihood function and the prior.

The mixed prior worked in my simulation study, and I am not worried too much about its conflict with the likelihood principle. The philosophy inspiring that principle has already been duly criticized by prominent Bayesian methodologists (Gelman & Shalizi, 2013). Gelman and Shalizi argue that the prior is really best construed as a model component. Bayesians should then test their prior to see if it fits the data, as appropriate (i.e. prior predictive checking).

One possible avenue of future research then is to investigate using mixed priors to handle cases where multiple experts disagree. One could then elicit priors from each expert, and mix them into a single prior using some weighting scheme. The advantage of this is that data can sort out which of these expert elicited priors is most consistent with the data. I also suspect that the inclusion of multiple experts in the prior selection process increases the odds a reasonable prior will be produced in analysis.

To test this last claim, methodologists could set up some psychology-type experiments. They could simulate survey data about a factious school program with open-ended items, ask several people to examine them, and guess whether the program was effective or not. They could then examine how accurate informants are under different conditions, and estimate how well Bayesian *t*-tests with mixed priors representing their collective opinions would fare in terms of their classical frequency properties using simulation methods.

Implications for Researchers

Building on my dissertation findings, I want to investigate how best to use Bayesian and classical inference in concert to build the special education knowledge base. I make a distinction between research for other content experts and research for outsiders. Usually, it is easier to sell our choice of prior to fellow experts. In this case, there is an agreed upon knowledge base. But the omission of priors in research for outsiders makes sense. It does not force outsiders to try to evaluate the expert's choice of prior. I see then a suitable division of labor for Bayesian and classical inference in a science of education.

The labor of Bayesian inference would be to persuade insiders of the merits of one's work. A pilot study is often conducted for the benefit of insiders for example. Bayesian analysis, thus, is a highly suitable choice in this circumstance. Other insiders can evaluate a researcher's choice of priors in these cases, and they can determine whether the findings are significant or not. After all, funding agencies usually rely on inside experts they recruit to make decisions about which proposals to fund and reject.

The labor of classical inference would be to persuade outsiders. This is especially true in a democracy. People without command of content literature are often not competent evaluators of an expert's choice of prior, and so the omission of priors in research intended to sway public policy makes sense. It seems to align with our democratic aspirations.

This division of labor, of course, is only for practical guidance. Exceptions to it are easy to imagine. The sampling conditions for classical inference may not be available for example. In these cases, researchers can proceed with Bayesian analysis at both the preliminary and final stages of investigation. I am confident Bayesian analysis could build consensus among experts in the long run.

Concluding Remarks

Qualitative knowledge is about types of things, and their excellences. There are many ways we humans produce qualitative knowledge about the world, and much of it is derived from non-formal sources (e.g., narratives, clinical experiences, intuitions). The purpose of my dissertation was to investigate the possibility of improving quantitative analysis with qualitative knowledge in special education science using Bayesian statistics. Three publishable articles formed the substance of this dissertation. In the first article, I argued that Bayesian statistics is generally more applicable than classical statistics. Specifically, I argued that the necessary sampling conditions for classical inferences are much too rare in special education science to sustain the current popularity of classical analysis. In the second article, I investigated the frequency properties of Bayesian *t*-tests using mixture of subjective priors as a default. I discovered they had remarkable power and adequate Type I error control under conditions typically encountered in special education. In the last article, I offered a practical demonstration of how to overcome statistical challenges with Bayesian methods. Together these three articles explored new ways that researchers can marshal the qualitative knowledge of their informants and so build up a romantic science of special education.

REFERENCES

Anastasiou, D., & Kauffman, J. M. (2012). Disability as cultural difference: Implications for special education. *Remedial and Special Education, 33*(3), 139-146.

Aristotle. (1985). *Nicomachean Ethics.* (I. Terence, Trans.) Indianapolis: Hacket.

Ayers, A. J. (1937/1952). *Language, truth, and logic.* New York, NY: Dover.

Bacon, F. (1620/1961). Novum Organum. In W. Kaufmann, *Philosophic classics: Bacon to Kant* (pp. 9-25). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Pyschological Methods, 18*(2), 252-264.

Barton, J., Burrello, L., & Kleinhammer-Tramill, J. (2012, April). *What does a district's commitment to "inclusive practice" mean and what is it impact?* Paper presented at the 2012 Annual Conference of the American Education Research Association, Vancouver, BC.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized controlled trails. *New England Journal of Medicine, 342*, 1878-1886.

Bhaskar, R. (1979). *The possibility of naturalism: A philosophical critique of the contemporary human sciences.* Atlantic Highlands, NJ: Humanities Press Inc.

Board of Ed. of Hendrick Hudson Central School Dist. v. Rowley., 458 U.S. 176 (1982)

Bolstad, W. M. (2007). *Introduction to Bayesian statistics.* Hoboken, N.J: John Wiley.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Buckley, J. (2004). Simple Bayesian Inference for Qualitative Political Research. *Political Analysis, 12*(4), 386-399.

Burrello, L. C., Sailor, W., & Kleinhammer-Tramill, J. K. (2013). *Unifying educational systems: leadership and Policy Perspectives.* New York, NY: Routledge Press.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data anlysis.* New York, NY: CRC Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York, NY: Lawrence Erlbaum Associates.

Crotty, M.J. (1998). *The Foundations of Social Research: Meaning and Perspective in the Research Process.*Oaks, CA: Sage.

Danforth, S. (2006). From epistemology to democracy: Pragmatism and the reorientation of disability research. *Remedial and Special Education*, 27(6), p. 337-345.

Davidson, D. (2006). Mental Events (Appendix: An Emerose by Another Name). In E. Lepore, & K. Ludwig (eds.), *The Essential Davidson.* New York, NY: Oxford University Press.

Denzin, N. K., & Lincoln, Y. S. (2008). *Strategies of Qualitative Inquiry* (3rd ed.). Thousand Oaks, CA: Sage.

Duke Evidence Based Practice Center. (2009, September 18). Use of Bayesian Statistics in Randomized Clinical Trials: A CMS case study. Rockville, MD: author.

Edgington, S. E. & Onghena, P. (2007). *Randomization tests* (4th ed.). New York, NY: Chapman & Hall.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33*, 503-513.

Fisher, R. A. (1930). Inverse probability. *Proceedings of Cambridge Philosophical Society, 26*, 528-535.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*, 8-38.

Gilhool, T. K. (1989). The right to an effective education: From Brown to PL 94-142 and beyond. In D. Lipsky, & A. Gartner, *Beyond seperate education: Quality education for all* (pp. 243-253). Baltimore, MD: Paul H. Brookes.

Gill, J. (2002). *Bayesian methods: A scoial and Behavioral sciences approach* . Washington, DC: Chapman & Hall/CRC.

Gill, J., & Walker, L. (2005). Elicited priors for Bayesian model specification in political science research. *The Journal of Politics, 67*, 841-872.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical Methods in Education and Pyschology (3rd).* Needham Heights, MA: Allyn & Bacon.

Goodman, N. (1979). *Fact, fiction, and forecast* (4th ed.). Cambridge, MA: Harvard University Press.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (pp. 105-117). Thousand Oaks, CA: SageHacking, I. (2001). *An introduction to probability and inductive logic.* New York, NY: Cambridge University Press.

Hacking, I. (2006). *The emergence of probability: A philosophical study of early ideas about probability and induction and statistical inference* (2nd ed.). New York, NY: Cambridge University Press.

Hacking, I. (2009). How we have been learning to talk about autism: The role of stories. *Metaphilosophy, 40*(1-2), 499-516.

Hand, D. J. (2008). *Statistics.* New York, BY: Sterling Publishing Co., Inc.

Harrison, G. W. (2011). Randomization and its discontents. *Journal of African Economies, 20*, 626-652. doi:10.1093/jae/ejr030

Hatch, J. A. (2002). *Doing qualitative research in education settings.* Chicago, IL: University of Chicago Press.

Hicks, T., & Knollman, G. (2014). Secondary analyses of National Longitudinal Transition Study-2: A statistical review. *Career Development and Transition for Exceptional Individuals*, doi: 10.1177/2165143414528240.

Hoff, P. (2009). *A first course in Bayesian statistical methods.* New York: Springer.

Hoppey, Black, and Mickelson. (2015.) *The evolution of inclusive practice in two elementary schools: Developing teacher purpose, instructional capacity, and data informed practice*. Manuscript submitted for publication. Tampa, FL: University of South Florida.

House, E. (1991). Realism and Research. *Educational Researcher, 20*(6), 2-9.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago, IL: Open Court.

Hume, D. (1739/1999). *An Enquiry concerning Human Understanding.* (T. L. Beauchamp, Ed.) New York, NY: Oxford University Press.

Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004)

Iversen, G. R. (1984). *Bayesian Statistical Inference (Series: Quantiative Applications in the Social Sciences).* Beverely Hills, CA: Sage University Paper.

Jaynes, E. T. (1968). Prior Probabilities. *IEEE Transactions of System Science and Cybernetics, 3*, 227-241.

Jones, P., Carr, J. F., & Fauske, F. (2011). *Leading for inclusion.* New York, NY: Teachers College Press.

Jorgensen, C. (2005). The least dangerous assumption: A challenge to create a new paradigm. *Disability Solutions, 6*(3), 1-15.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of American Statistical Association, 91*, 1343-1370.

Kauffman. (2011). *Toward a science of education: The battle between rogue and real science.* Verona, WI: Full Court Press.

Kauffman, J. M., & Badar, J. (2014). Instruction, not Inclusion, should be the central issue in special education: An alternative view from the USA. *Journal of international Special Needs Education, 17*, 13-20.

Kolmogorov, A. N. (1933/1956). *Foundations of the Probability Theory.* (N. Morrison, Trans.) New York, NY: Chelsea Publishing Company.

Krathwohl, D. R. (1998). *Methods of educational and social science research.* New York, NY: Longman.

Kruschke, J. K. (2011). *Doing Bayesian data analysis a tutorial with R and BUGS.* Amsterdam: Academic Press.

Kuhn, T. (1962). *The structure of scientific revolutions* (2nd Edition ed.). Chicago, IL: University of Chicago Press.

Labree, D. F. (2011). The lure of statistics for educational researchers. *Theory of Education, 61*(6), 621-632.

Lagemann, E. C. (2000). *An elusive science: The troubling history of education research.* Chicago, IL: The University of Chicago Press.

Lane, K. L., Wehby, J. H., Little, M. A., & Cooley, C. (2005). Students educated in self-contained classrooms and self-contained schools: Part II. How do they progress over time. *behavior Disorders, 30*, 363-374.

Lanehart, R. E., Rodriguez de Gil, P., Kim, E. S., Bellara, A. P., Kromrey, J. D., & Reginald, S. L. (2012). Propensity score analysis and assessment of propensity score approaches using SAS procedures. *SAS Global Forum 2012.* SAS Institude, Inc.

Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics.* New York, NY: Springer.

Lewis, D. (1980). A Subjectivist's Guide to Objective Chance. In C. J. Richard, *Studies in Inductive Logic and Probability* (pp. 263-293). Berkeley: University of California Press.

Lynch. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods and Research, 32*(3), 301-335.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* New York, NY: Springer.

Marks, S. (2011). Special education: More about social justice, less about caring. *Phi Delta Kappan, 93*(1), 80.

McDowell, J. (1978). Are Moral Requirements Hypothetical Imperatives? *Proceedings of the Aristotelian Society, Supplementary Volumes*, *52*(1), 13-29.

McGrayne, S. B. (2012). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down russian submarines, & emerged trumphant from two centuries of controversy.* CN: Yale University Press.

McLeskey, J., & Waldron, N. L. (2000). *Inclusive schools in action: Making differenes ordinary.* Alexandria, VA: Association for Supervision and Curriculum Development.

McLesky, J., & Waldron, N. L. (2011). Educational programs for elementary students with learning disabilities: Can they both be effective and inclusive? *Learning Disabiliteis Research and Practice*, 48-57.

Mohr, L. B. (1990). *Understanding significance testing.* New York, NY: Sage.

Morgan, D.L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1* 48-76. doi: 10.1177/2345678906292462

Morgan, P., Frisco, M. L., Farkas, G., & Hibel, j. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education, 43*(4), 236-254.

Muthen, B. (2013). BSEM measurement invariance analysis. Mplus Web Notes: No. 17. January 11, 2013. Retrived at www.statmodel.com/BSEM.shtml.

Nagel, T. (2012). *Mind and cosmos: Why the materialist neo-Darwinian conceptio nof nature is almost certainly false.* New York: Oxford University Press.

Odom, S.L., Brantlinger, E., Gersten, R., Thompson, B., & Harris, K.R. (2005). Research in Special Education: Scientific Methods and Evidence-Based Practice. *Exceptional Children. 71*, 137-148.

Paul, J. L. (2005). *Introduction to the Philosophies of Research and Criticism in Education and the Social Sciences.* Upper Saddle River, NJ: Pearson.

Phillips, D. (1987). Validity in qualitative research: Why the worry about warrant will not wane. *Education and Urban Soceity*, 20, p. 9-24.

Phillips, D. C., & Burbules, N. (2000). *Postpositivism and educational research.* Lanhan, MD: Rowman & Littlefield.

Plantinga, A. (1993). *Warramt and Proper Function.* New York, NY: Oxford University Press.

Pollard, W. E. (1986). *Bayesian statistics for evaluation research: An introduction.* Beverly Hills, CA: SAGE.

Pugach, M. (2001). The stories we choose to tell: Fulfilling the promise of qualitative research in special education. *Exceptional Children, 67*(4), 439-453.

Quine, W. (1969). *Ontological Relativity and Other Essays.* New York, NY: Columbia University Press.

Ramsey, F. P. (1926). Truth and Probability. In Rmasey, 1931, *The Foundations of Mathematics and Other Logical Essays.* Ch. VII, p. 156-198, edited by R.B. Braithwaite, London: Kegan, Trubner & Co.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods (2nd).* Thousand Oaks, CA: Sage Publications.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects: Using experimental and observational designs.* Washington, D.C.: American Educational Research Association.

Sider, T. (2011). *Writing the Boook of the World.* New York, NY: Oxford University Press.

Silver, N. (2012). *THe signal and the noise: Why so many predictiosn fail - but some don't.* New York: The Penguine Press.

Smith, D. J. (2003). *In search of better angels.* Thousand Oaks, CA: Corwin Press, Inc.

Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills? A propensity score weighting analysis. *Journal of School Psychology, 51*(2), 243-260.

Toson, A., Burrello, L. C., & Knollman, G. (2012). *Justice for all: The capability approach and inclusive leadership.* London: International Journal of Inclusive Education.

U.S. Department of Education, N. C. (2014). *Digest of Education Statistics, 2013.* Washington, D.C.: NCES.

Von Mises, R. (1936). *Warsheinlichkeit, Statistik und Wahrheit.* Vienna: Springer-Verlag.

Ware, W. B., Ferron, J. M., & Miller, B. M. (2013). *Introductory statistics: A conceptual approach using R.* New York: Routledge.

Williams, M. (2001). *Problems of Knowledge: A critical introduction to epistemology.* New York, NY: Oxford University Press.

Worrall, J. (2007). Why there's no cause to randomize. *British Journal of the Philosophy of Science, 58*, 451-488.

Yell, M. L. (2006). *The law and special education* (2nd ed.). Upper Saddle River, NY: Pearson.

Yu, C. H. (2006). *Philosophical Foundations of Quantiative Research Methodology.* Lanham, MD: American University Press.

Zigmond, N., Kloo, A., & Volonino, V. (2009). What, where, how? Special education in the climate of full inclusion. *Exceptionality, 17*, 189-204.

APPENDIX A

SAS PROGRAMMING CODE FOR SIMULATION STUDY

```
ptions ps = 500 ls = 160;
filename junk dummy; *Make sure HTML is Turned Off;
proc printto log=junk print=junk;
      run;
*+--------------------------------------+
  STEP 1: DEFINE SIMULATION CONDITIONS
 +--------------------------------------+;
*+-----------------------------------------------+
*To Manually Enter Condition use following code;
       DATA conditions;
            conditions=999;
      Delta=3; *Delta: 0~U(-.1,.1), 1~U(.2,.4), 2~U(.5,.8), 3~U(.8,1);
      Heterogeneity=0; *Variance Ratio 0=(1:1),1=(1:2),2=(1:1.5),3=(1:3);
      Shape=1;* Normalcy: 1=(Sk=0,Ku=0), 2=(Sk=1,Ku=3), 3=(Sk=1.5,Ku=5),
4=(Sk=2,Ku=6), 5=(Sk=0,Ku=25);
      Group_Ratio=0; *Group ratios 0=(1:1), 1=(1:2), 2=(1:3);
      Imbalance_bias = 2; *1=Larger Control Group, 2=Larger Treatment Group -
note: moot if there is balance;
      Big_N=12; *12,24,36,48,60 or any multiple of 12;
            n_reps = 3;
      run;

      *To randomize conditions use following code;
      /*DATA conditions;
            conditions=999;
      Delta=3;
      Heterogeneity=3;
      Shape=5;
      Group_Ratio=2;
      Imbalance_bias = 2;
      Big_N=60;
            Y=RAND("Uniform"); *Delta: T~U(-.1,.1), S~U(.2,.4), M~U(.5,.8),
L~U(.8,1);
                  If Y < 0.25 then Delta=0;
                  If (0.25 =< Y AND Y <0.5) then Delta=1;
                  if (0.5 =< Y AND Y <0.75) then Delta=2;
            X=Rand("Uniform"); *Variance Ratio
0=(1:1),1=(1:2),2=(1:1.5),3=(1:3);
                  If X < .25 then Heterogeneity=0;
                  if (0.25 =< X AND X <0.5) then Heterogeneity=1;
            if (0.5 =< X AND X <0.75) then Heterogeneity=2;
            Z=RAND("Uniform"); * Normalcy: 1=(Sk=0,Ku=0), 2=(Sk=1,Ku=3),
3=(Sk=1.5,Ku=5), 4=(Sk=2,Ku=6), 5=(Sk=0,Ku=25);
                  If Z < 0.2 then Shape=1;
                  If (0.2 =< Z AND Z <0.4) then Shape=2;
                  if (0.4 =< Z AND Z <0.6) then Shape=3;
```

```
                  if (0.6 =< Z AND Z <0.8) then Shape=4;
            A=Rand("Uniform"); *Group ratios 0=(1:1), 1=(1:2), 2=(1:3);
                  If A < (1/3) then Group_Ratio=0;
                  if ((1/3) =< A AND A <(2/3)) then Group_Ratio=1;
            *Imbalance_bias only matters if Group Ratio is set to 1 or 2.
          When Group Ratio is set to 0 Imbalance_bias becomes moot;
            B=Rand("Uniform"); *1=Larger Control Group, 2=Larger Treatment
Group;
                  if B < .5 then Imbalance_bias = 1;
            C=RAND("Uniform"); *Big_N: 12,24,36,48,60;
                  If C < 0.2 then Big_N=12;
                  If (0.2 =< C AND C <0.4) then Big_N=24;
                  if (0.4 =< C AND C <0.6) then Big_N=36;
                  if (0.6 =< C AND C <0.8) then Big_N=48;
            n_reps = 1;
            Keep Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N
Rater n_reps conditions;
            run; */
*+-------------------------------------------------------------+
      STEP 2: Setup sampling Procedures
 +-------------------------------------------------------------+;
 Data _null_;
      Set conditions;
*------------------------*
  Controls Group Balance
*------------------------*;
    Group_1=1;
 if Imbalance_bias = 1 then do; *Larger Control Group;
      if (Imbalance_bias = 1 AND Group_Ratio=0) then do;
            Group_2=1; *Balanced (Ratio 1:1); end;
      if (Imbalance_bias = 1 AND Group_Ratio=1) then do;
            Group_2=2; *Imbalanced (Ratio 1:2); end;
      if (Imbalance_bias = 1 AND Group_Ratio=2) then do;
            Group_2=3; *Imbalanced (Ratio 1:3); end;
            little_nT = Big_N/(Group_1+Group_2);
      little_nC = little_nT*Group_2; end;

if Imbalance_bias = 2 then do; *Larger Treatment Group;
      if (Imbalance_bias = 2 AND Group_Ratio=0) then do;
            Group_2=1; *Balanced (Ratio 1:1); end;
      if (Imbalance_bias = 2 AND Group_Ratio=1) then do;
            Group_2=2; *Imbalanced (Ratio 1:2); end;
      if (Imbalance_bias = 2 AND Group_Ratio=2) then do;
            Group_2=3; *Imbalanced (Ratio 1:3); end;
            little_nC = Big_N/(Group_1+Group_2);
      little_nT = little_nC*Group_2; end;
*--------------------------------------*
  Output Commands throughout Simulation
*--------------------------------------*;
      call symput('little_nC',little_nC);
      call symput('little_nT',little_nT);
      call symput('Conditions',Conditions);
      run;

*+----------------------------+
      STEP 3: Simulate Samples
```

```
 +-----------------------------+;

DATA Samples;
      SET conditions;

      Do Replicate = 1 to n_reps;

      *+-----------------------------------------------+
            Population Parameters for Treatment Condition
      +-----------------------------------------------+;

      *For conceptual clarity, Delta in this study is defined as the
       magnitude of the difference between treatment and control populations
      divided by the control group's sigma;

      if Delta = 0 then do; *Controls location of Treatment Population;
            Mu=50; end;
      if Delta = 1 then do;
            Mu=53; end;
      if Delta = 2 then do;
            Mu= Rand("Uniform");
            Mu=56; end;
      if Delta = 3 then do;
            Mu= Rand("Uniform");
            Mu=59; end;

      truth = Mu-50;

      if Heterogeneity = 0 then do; *Controls Scale of Treatment Population;
            Sigma= 10; end;
      if Heterogeneity = 1 then do;
            Sigma= 15; end;
      if Heterogeneity = 2 then do;
            Sigma= 20; end;
      if Heterogeneity = 3 then do;
            Sigma= 30; end;

      *+--------------------+
        Simulate Samples
      +--------------------+;

      DO ID = 1 TO &little_nC; *Control Condition;
         X=0;
       Y = RAND("Normal",50,10); OUTPUT;END; *~N(Mu,Sigma);

      DO ID = (&little_nC+1) to (&little_nT+&little_nC-1); *Treatment
Condition;
            X=1;
            Z= RAND("Normal");*~N(0,1);
*-------------------------------------------*
    Fleishman Transformations to nonnormality
*-------------------------------------------*;
               * The following give sk= 0, kr= 0;
      if shape = 1 then do;
      b=1;
      c=0;
```

```
      d=0;
      end;
           * The following give sk= 1.00, kr= 3.00;
      if shape = 2 then do;
      b= .83221632289426;
      c= .12839670935047;
      d= .04803205907079;
      end;
           * The following give sk= 1.50, kr= 5.00;
      if shape = 3 then do;
      b= 0.78340767306328;
      c= 0.18620283598117;
      d= 0.05704195885033;
      end;
           * The following give sk= 2.00, kr= 6.00;
      if shape = 4 then do;
      b= 0.82632385761082;
      c= 0.31374908500462;
      d= 0.02270660525731;
      end;
           * The following give sk= 0.00, kr= 25.00;
      if shape = 5 then do;
      b= -1.5666815059210;
      c=  0.00000000000000;
      d=  0.34817270324943;
      end;
      Z = b*Z + c*Z**2 + d*Z**3 - c;
      Y = Mu + (Sigma*Z); OUTPUT;END;*converts to raw metric;

      output; End;
*------------------------------*
  Cleanup Simulation Data Set
*------------------------------*;
      DROP Z b c d ;
      run;
*+--------------------------------------+
      STEP 4: CONDUCT STATISTICAL ANALYSIS
 +--------------------------------------+;
*---------------------*
  Classical t-test
*---------------------*;
      proc ttest data=Samples;
           by replicate Delta Shape Heterogeneity Group_Ratio Imbalance_bias
           Big_N truth n_reps;
           title 't-tests';
           class X;
           var Y;
        ods output conflimits=stats;
           run;
*-----------------------------------------------------*
  Bayesian t-test (w/ Jeffreys Priors for all parameters)
*-----------------------------------------------------*;
proc mcmc data=samples thin=10 nmc=100000 nbi=10000
           monitor=(mudif);
           by replicate Delta Shape Heterogeneity Group_Ratio Imbalance_bias
Big_N truth n_reps;
```

```
            parms muT 50 muC 50;
            parms sig2T 100;
            parms sig2C 100;
            prior muT ~ general(1); *convenience prior;
            prior muC ~ general(1); *convenience prior;
            prior sig2T ~ general(-log(sig2T), lower=0); *convenience prior;
            prior sig2C ~ general(-log(sig2C), lower=0); *convenience prior;
            mudif = muT - muC;
            if X = 1 then do;
            mu = muT;
            s2 = sig2T;
            end;
            else do;
            mu = muC;
            s2 = sig2C;
            end;
            model y ~ normal(mu, var=s2);
            ods output postsummaries=pointthree postintervals=intervalthree;
            run;
*------------------------------------------------------------------*
  Bayesian t-test with Mixed Prior for mean of intervention group
*------------------------------------------------------------------*;
proc mcmc data=samples thin=10 nmc=100000 nbi=10000
            monitor=(mudif) init=mode;
            by replicate Delta Shape Heterogeneity Group_Ratio Imbalance_bias
Big_N truth n_reps;
            parms muT 55 muC;
            parms sig2T 100;
            parms sig2C;
            lp = logpdf('normalmix', mut, 4, 0.25, 0.25, 0.25, 0.25, 50, 53,
56, 59, .333, .333, .333, .333);
            prior muT ~ general(lp); *Mixture prior;
            prior muC ~ n(mean= 50, var=1); *informed prior;
            prior sig2T ~ general(-log(sig2T), lower=0); *convenience prior;
            prior sig2C ~ igamma(shape=19, scale=2000); *informed prior;
            mudif = muT - muC;
            if X = 1 then do;
            mu = muT;
            s2 = sig2T;
            end;
            else do;
            mu = muC;
            s2 = sig2C;
            end;
            model y ~ normal(mu, var=s2);
            ods output postsummaries=pointfour postintervals=intervalfour;
            run;
*------------------------------------------*
  Bayesian t-test with Partial Prior
*------------------------------------------*;
proc mcmc data=samples thin=10 nmc=100000 nbi=10000
            monitor=(mudif) init=mode;
            by replicate Delta Shape Heterogeneity Group_Ratio Imbalance_bias
Big_N truth n_reps;
            parms muT 50 muC;
            parms sig2T 100;
```

```
            parms sig2C;
            prior muT ~ general(1); *convenience prior;
            prior muC ~ n(mean= 50, var=1); *informed prior;
            prior sig2T ~ general(-log(sig2T), lower=0); *convenience prior;
            prior sig2C ~ igamma(shape=19, scale=2000); *informed prior;
            mudif = muT - muC;
            if X = 1 then do;
            mu = muT;
            s2 = sig2T;
            end;
            else do;
            mu = muC;
            s2 = sig2C;
            end;
            model y ~ normal(mu, var=s2);
            ods output postsummaries=pointfive postintervals=intervalfive;
            run;
*+------------------------+
      STEP 5: MANAGE OUTPUT
 +------------------------+;
*-------------------------------------*
  Classical t-test Output Management
*-------------------------------------*;
      data stats;
      set stats;
      where Class='Diff (1-2)';
      data one;
            set stats;
            where Method='Pooled';
            mean=mean*(-1);
        UpperCL=LowerCLMean*(-1);
        LowerCL=UpperCLMean*(-1);
            bias = (mean - truth);
      RMSE = (mean - truth)**2;
            CI_width = UpperCL - LowerCL;
            CI_coverage = 0;
            if LowerCL =< truth and UpperCL >= truth then CI_coverage = 1;
            CIreject = 0; *Significence Test using Intervals;
            if LowerCL > 0 or UpperCL < 0 then CIreject = 1;
            run;
      proc means noprint data = one;
            by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N
n_reps;
            var bias RMSE CI_width CI_coverage CIreject CIreject;
            output out = onefinal mean = ;
            run;
      data onefinal;
            set onefinal;
            RMSE = SQRT(RMSE);
            Method=1;
            run;
      data two;
            set stats;
            where Method='Satterthwaite';
            mean=mean*(-1);
        UpperCL=LowerCLMean*(-1);
```

```
        LowerCL=UpperCLMean*(-1);
            bias = (mean - truth);
      RMSE = (mean - truth)**2;
            CI_width = UpperCL - LowerCL;
            CI_coverage = 0;
            if LowerCL =< truth and UpperCL >= truth then CI_coverage = 1;
            CIreject = 0; *Significence Test using Intervals;
            if LowerCL > 0 or UpperCL < 0 then CIreject = 1;
            run;
      proc means noprint data = two;
            by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N
n_reps;
            var bias RMSE CI_width CI_coverage CIreject CIreject;
            output out = twofinal mean = ;
            run;
      data twofinal;
            set twofinal;
            RMSE = SQRT(RMSE);
            Method=2;
            run;
*------------------------------------------------------------*
  Bayesian T-test with Jeffreys Priors Output Management
*------------------------------------------------------------*;
 data pointthree;
      set pointthree;
    bias = p50 - truth;
    RMSE = (p50 - truth)**2;
      run;
proc means noprint data = pointthree;
 by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N n_reps;
 var bias RMSE;
 output out = pointestimatesthree mean = ;
 run;
data intervalthree;
      set intervalthree;
      CI_width = HPDUpper - HPDLower;
      CI_coverage = 0;
      if HPDLower =< truth and HPDUpper >= truth then CI_coverage = 1;
      CIreject = 0; *Significence Test using Intervals;
      if HPDLower > 0 or HPDUpper < 0 then CIreject = 1;
      run;
proc means noprint data = intervalthree;
      by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N n_reps;
      var CI_width CI_coverage CIreject;
      output out = intervalestimatesthree mean = ;
      run;
data threefinal;
      merge pointestimatesthree intervalestimatesthree;
      RMSE =SQRT(RMSE);
    Method=3;
      run;
*------------------------------------------------------------*
  Bayesian T-test with Mixed Priors Output Management
*------------------------------------------------------------*;
 data pointfour;
      set pointfour;
```

```
    bias = p50 - truth;
    RMSE = (p50 - truth)**2;
      run;
proc means noprint data = pointfour;
 by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N n_reps;
 var bias RMSE;
 output out = pointestimatesfour mean = ;
 run;
data intervalfour;
      set intervalfour;
      CI_width = HPDUpper - HPDLower;
      CI_coverage = 0;
      if HPDLower =< truth and HPDUpper >= truth then CI_coverage = 1;
      CIreject = 0; *Significence Test using Intervals;
      if HPDLower > 0 or HPDUpper < 0 then CIreject = 1;
      run;
proc means noprint data = intervalfour;
      by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N n_reps;
      var CI_width CI_coverage CIreject;
      output out = intervalestimatesfour mean = ;
      run;
data fourfinal;
      merge pointestimatesfour intervalestimatesfour;
      RMSE =SQRT(RMSE);
    Method=4;
      run;
*-------------------------------------------------------------------*
  Bayesian T-test with Partial Prior Output Management
*-------------------------------------------------------------------*;
 data pointfive;
      set pointfive;
      where parameter='mudif';
    bias = p50 - truth;
    RMSE = (p50 - truth)**2;
      run;
proc means noprint data = pointfive;
 by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N n_reps;
 var bias RMSE;
 output out = pointestimatesfour mean = ;
 run;
data intervalfive;
      set intervalfive;
      CI_width = HPDUpper - HPDLower;
      CI_coverage = 0;
      if HPDLower =< truth and HPDUpper >= truth then CI_coverage = 1;
      CIreject = 0; *Significence Test using Intervals;
      if HPDLower > 0 or HPDUpper < 0 then CIreject = 1;
      run;
proc means noprint data = intervalfive;
      by Delta Shape Heterogeneity Group_Ratio Imbalance_bias Big_N n_reps;
      var CI_width CI_coverage CIreject;
      output out = intervalestimatesfour mean = ;
      run;
data fivefinal;
      merge pointestimatesfour intervalestimatesfour;
      RMSE =SQRT(RMSE);
```

```
    Method=5;
      run;
*---------------------*
  Merge Output Files
*---------------------*;
Data Results;
      Set onefinal twofinal threefinal fourfinal fivefinal;
*----------------------------*
  STEP 6: ISSUE FINAL REPORT
*----------------------------*;
proc printto log=junk print=print; *"C:\Users\CCM\Documents\Tyler's
Stuff\SAS\Test.txt"; *print or 'c:\results.txt';
      run;
PROC FORMAT;
      Value methodfmt
            1='Pooled'
          2='Satterthwaite'
            3='Convenience'
            4='Mixture'
            5='Partial';
      run;
Data Results;
      Set Results;
      Format method methodfmt.;
      Conditions=&Conditions;
proc print data = results;
      var Conditions method Delta Shape Heterogeneity Group_Ratio
Imbalance_bias Big_N n_reps bias RMSE CI_width CI_coverage CIreject;
      title ' Simulation Results';
```