SIMULATION BASED SPECIFICATIONS FOR EVALUATING HIGH STAKES EDUCATIONAL TEST RESULTS FROM A BAYESIAN EPISTEMOLOGICAL PERSPECTIVE: PHILOSOPHICAL FOUNDATIONS, RESPONSE SURFACE INVESTIGATIONS, AND PRAGMATIC APPLICATIONS OF RESULTING PSYCHOMETRIC MODESTY

by

DAVID ARTHUR FLUHARTY

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2007

MAJOR: EDUCATIONAL EVALUATION AND RESEARCH

Approved by:

_____

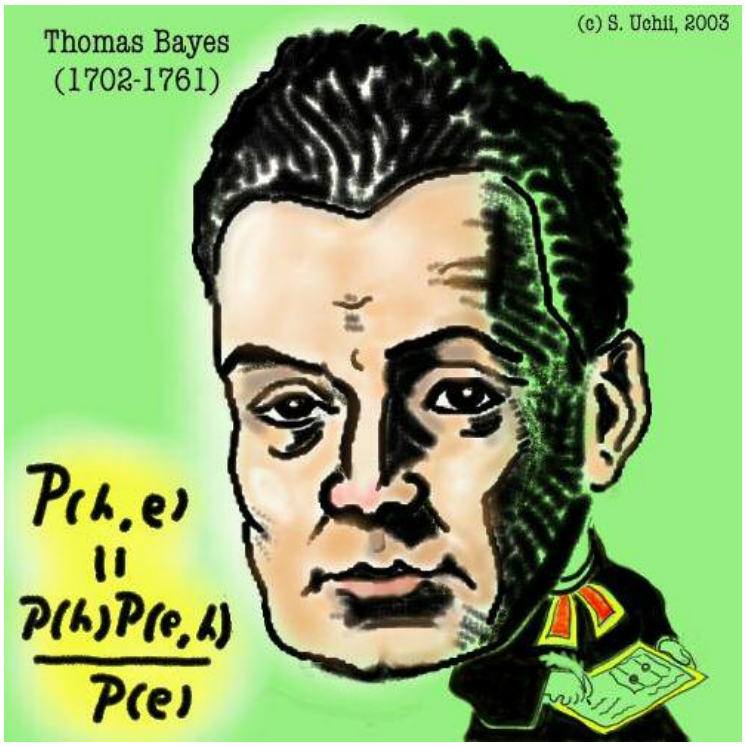
Advisor Date

_____

_____

_____

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346



Reproduced by Permission
Of Professor Soshichi Uchii

All models are wrong;

some models are useful.

G.E.P. Box, FRS

If the misery of our poor be caused

not by the laws of nature, but

by our institutions,

great is our sin.

Charles Darwin, FRS
Voyage of the Beagle
Quoted by Jay Gould

DEDICATION

This dissertation is dedicated to our daughter

Margaret Rose Elaine Fluharty—Reiter (1986-1995)

Who spoke of becoming a teacher

PREFACE

This is what might be called a teaching dissertation. This was pointed out to me by my advisor, Dr. Shlomo Sawilowsky. While most dissertations focus on the topic, a teaching dissertation also reviews foundational ideas which, for maturing disciplines such as educational statistics and measurement, are generally taken for granted. In addition, the author of a teaching dissertation places present research in context of those foundations. Such dissertations are rare outside of Education, and even within Education many dissertations are not teaching dissertations.

In this dissertation I go back to the foundations. Not only is a history of simulation presented, but the work discusses theories of probability, philosophy of science, and epistemology. It is indeed rare that a work which employs computer intensive methods goes back to discuss Plato or refers to Tina Turner. The primary student of this teaching dissertation is the author. It enabled me not only to explore a focused research question of deep interest, but also to gain a much better understanding of foundations that make the research possible.

I once heard it said that one medieval conception of the university was as a machine to perfect the students. Although this work has certainly not perfected me, it certainly has changed the way I view the world and how to do research. Thus, it is a record of this change. I hope you find it useful.

ACKNOWLEDGMENTS

This section is to thank just a few of the people without whose help this dissertation would not have been completed. No work of "original research" is truly original, that is, without foundation—and this is particularly true of this work which explores perennial ideas about epistemology. In addition to the individuals who wrote the works cited, every statistics, philosophy, theology, education, science, and social science professor, teacher, or trainer I have had—either in a university course or industrial setting—has probably in some way contributed to the thinking that has gone into this dissertation.

First among these I would like to thank my major advisor, Dr. Shlomo Sawilowsky. He not only made specific suggestions which improved my learning process (for example, to simulate off of an actual set of test scores), but assisted in much more fundamental ways. First, he steered me away from a topic that is still of great interest (the potential educational impact of teaching basic Statistical Design of Experiments at a High School level) because this would have been more appropriate for an Ed.D. than a Ph.D. Second, he supported my decision—unsurprising to those who know me well—to pick a topic about which I knew little (application of Bayesianism) but about which I wanted to learn and possibly pursue a research programme[1] . Third, he was persistent in encouraging me to work on the project despite several major career changes.

[1] Graduate students reading this should realize that working on part of one's advisor's research program will lead to much quicker completion!

Finally, as I am employed 300 miles from Wayne State, he went out of his was to assist in administrative matters.

I am grateful to the other members of my dissertation committee:

- Dr. Gail Fahoome had helped me work through some of the basic approaches used in this study during her Monte Carlo simulation course. In addition, she agreed to take join the committee at a late date.
- Dr. Boris Shulkin who joined the committee at a late date without the "benefit" of having had me as a student.
- Dr. TaChen Liang, my cognate advisor from the Mathematics Department, whose questions during my proposal defense prompted an initial study of WinBUGS Markov Chain Monte Carlo software. In addition, I am grateful for the facility I gained in working with probability distributions during one of his courses.
- In addition, I would like to acknowledge two original members of my committee who have left the university, Dr. Lori Rothenberg (now at SAS Institute) and Dr. Weimo Zhu (now at the University of Illinois at Urbana-Champaign).

There are several other Wayne State Faculty members to whom special thanks are due: First, to Dr. Susan Vineberg of the Philosophy Department, who graciously welcomed me to her seminar in Bayesian Confirmation Theory. Dr. Leon Wilson in the Department of Sociology gave me new perspectives on the use of advanced statistical methods in the social sciences. Within the

School of Education, three professors from whom I took more than one course influenced my thinking about the field of statistics, philosophy, and curriculum: Dr. Donald Marcotte, and the late Dr. Arthur Brown and particularly Dr. Leonard Kaplan who provided me with a curricular perspective on all research I might do relating to education—not to mention providing me with understanding which has greatly enhanced my teaching. Finally, I am indebted to Dr. Gerard Teachman, from whom I took my first course Philosophy of Education in 1995. The intellectual excitement of this course helped convince me to pursue my degree.

In addition to the individuals above, I owe a debt to every teacher of statistics I have had. At the University of Chicago Dr. Charles R. Nelson and the late Dr. Harry Roberts were particularly influential. As an undergraduate at Wheeling College (now Wheeling Jesuit University) I took my first statistics course from Fr. Joseph Kolb, S. J. The late Dr. John Gasiorwski, my undergraduate mentor, stressed the integration of philosophy and the social sciences. In industry I have been fortunate to have taken seminars and to benefit from conversations with Dr. J. Stewart Hunter and the late Dr. W. Edwards Deming.

A number of administrative personnel at Wayne State University helped me greatly—and repeatedly: Mr. Paul Johnson and Ms. Sharon Seller-Clark in the College of Education, Ms. Cindy Sokol of the Graduate School, and Ms. Pamela Brasgalla, formerly of the Ph.D. Office. Mr. Johnson was particularly

helpful in doing a number of tasks which would have required me to return to Detroit had he not done them.

In addition to those who helped with the research and administration, several organizations provided funding. First was a generous Graduate Professional Scholarship from Wayne State University. My former employer, Aloca Fujikura Ltd., provided substantial tuition reimbursement (consistent with their commitment to 'employability')' In addition, a number of credits were transferred from the Oakland University program in Applied Statistics which Ford funded when I was an employee at Ford.

Friends, colleagues, and managers in the quality, statistics, and automotive field have been very encouraging: Special thanks to Dr. Gerry Darnell and Suhial Horan of Remy Inc., and Dr. William B. Smith and Dr. Martha Aliaga of the American Statistical Association.

I would also like to express my gratitude to my parents, Ralph Fluharty and Grace Elaine Fluharty. They were of the generation of working class parents who believed in sacrificing for their children so their children could have a life better than their own. Although my parents certainly did not understand the ins and outs of academia, they were so convinced of the value of an education that it was always clear to me that I would go to college—even if they had no clear idea what one studies in college. Indeed, the journey that ended with this dissertation started with them. Parents are a child's first teachers.

In conclusion, I want to thank my wife, Mary Reiter. While other husbands might have been fixing something around the condominium, doing

something to directly advance their careers, or just spending time with their wives, she gave me constant encouragement while I pursued a dream.

The dataset (Appendix A) was provided by C. Schram, then of the Michigan Department of Education. The study was conducted with Design-Expert ® software Version 6 by Stat-Ease and Minitab® Statistical Software Version 13, Minitab, Inc. The staffs of both companies were extremely responsive and helpful. Numerous graphs are produced with Design-Expert ® software by Stat-Ease, Inc. (Minneapolis, MN). Portions of the input and output contained in this dissertation are printed with permission of Minitab Inc. All material remains the exclusive property and copyright of Minitab Inc. All rights reserved. Figure 3b, which is also reproduced on page iii, is reproduced with the permission of Professor Soshichi Uchill. Figure 4 is reproduced with the permission of Dr. Stephen Stigler.

The errors in this work are, of course, my own. (Note: Before doing a dissertation I always thought this was a perfunctory statement. Having worked on this paper for a number of years—and seeing what I would have done differently—I now understand why each researcher makes this statement.)

TABLE OF CONTENTS

LIST OF TABLES

TABLE **PAGE**

CHAPTER 1

INTRODUCTION

Background

This study is motivated by the actions of the governments in the United States at the federal and state levels that have attached increasingly high stakes to standardized educational tests. Just as the profound consequences that resulted from innovation of physicists at the beginning of the last century prompted a deep philosophical consideration of the meaning and methods of science, so to the profound consequences for the lives of thousands of individuals from the innovations by government officials should prompt a deep philosophical investigation of the meaning and methods of standardized testing. This work is part of this investigation and ultimately a call for psychometric modesty—that is, a call for government officials, the educational administrators they employ, and the general public not to expect or demand more of the tests than is psychometrically justified.

To this end this dissertation explores to what extent two influential, yet controversial, post war (post World War II) intellectual positions can be mutually illuminating and whether this illumination can be used to improve educational policy. The specific positions are as follows:

- There should be high stakes educational tests, and
- Bayesian approaches can be used to improve measurement, more precisely, that a simple Bayesian approach can improve educational evaluation.

To explore these propositions, an ancillary hypothesis that Classical Measurement Theory is in some sense true, or at least usable, is assumed. The exploration makes extensive use of Monte Carlo Simulation, Response Surface Methodology, and, to a lesser degree, Logistic Regression. The statistical tools are employed in a 'classical' rather than Bayesian mode. The specifically Bayesian mechanism is a simple form of Bayes's Theorem used more by individuals working in the Philosophy of Science than in Bayesian Statistics. It is hoped that this relatively gentle approach to Bayesianism will, if the results are published, encourage educators and educational administrators to consider Bayesian approaches, if not for policy implementation, at least to gain a deeper understanding of the limits of high stakes testing and the potential of Bayesian methodology. In a sense this is to stress an emphasis found at the roots of modern Bayesianism for an early and major work of Bayesian statistics published in the *Psychological Review* was aimed at researchers concerned with practical matters connected to this domain: "The 1963 paper by Edwards, Lindman, and Savage introduces Bayesian inference to practically minded audiences" (Kotz and Johnson, 1993).

Evaluation[2] has become a dominant theme in the public debate about education. Standardized tests, particularly high stakes tests, are becoming an increasingly important means of evaluating performance and implementing accountability. This was becoming apparent at the end of the last century. Phelps (1998) indicated that the American public *wants* standardized tests. Standardized tests can have high stakes for students, teachers, administrators, schools, and districts, and politicians. In 1998 Joftus and Whitney indicated that "Fifteen states currently prevent students who do not meet academic standards form graduating or being promoted" (p. 28). Manzo (1998) indicated the implications of high stakes tests for schools, "Illinois issues warnings to those [schools] that don't have at least half of all students scoring at grade level on state tests. Texas grants its highest rating only to those schools with at least 90 percent of students passing each subject area on state tests" (p. 26). Shadham (1998) indicated that states tests can have direct financial impact, and that in September 1998 the Pennsylvania legislature distributed "$10 million to 994 schools that scored exceptionally well on statewide tests" (p. 13). There is a long standing debate about the impact of the tests on education. However, Shadham (1998) also noted the debates rage as to whether high stakes tests used for

---

[2] The emphasis on educational evaluation can be viewed as part of a larger American intellectual trend, which might be termed 'Measurableism,' that is, the reduction of the evaluation of most fields important human endeavor, whether business, government, or education—and decisions upon which these evaluations are based—to simple, and at times simplistic, measurement. But these measurements often hide value judgments. The fact that the Bayesian epistemological framework brings the judgmental element into explicit

accountability are rigorous enough and whether improvements in test scores correlate with other examinations, such as other state exams or National Assessment of Educational Progress (p. 13). Blair (1998) cited others who wonder if the improvement in test scores represents an improved education:

> Some research shows that such incentive programs can boost test scores. In North Carolina, for example, 84 percent of students in grades K-8 met state standards in 1998, compared with 57 percent in 1997, following implementation of the state's incentive plan, officials said.. . .But some critics believe the high test scores [can produced results that] can be deceiving. Teachers and Principal may end up so preoccupied with tests that 'they may not provide the kind of time it takes to give students a comprehensive education,' said John I. Wilson, the executive director of the North Carolina Association of Educators. (p. 5)

More recently, the Center for Educational Policy (2006) has indicated that 22 states with 65% of US students and 71% of minority students have test requirements for graduation or advancement. This is expected to rise to 25 states with 75% of all students and 81% of minority students in 2012 (pp. 1-2). In addition, in 2006 ten additional states which did not require tests for graduation or advancement attach high stakes to certain tests (see Figures 1 and 2).

---

consideration is one of the author's motivations for using this framework. The intellectual predecessor of "measurableism" is Bridgman's Operationalism.

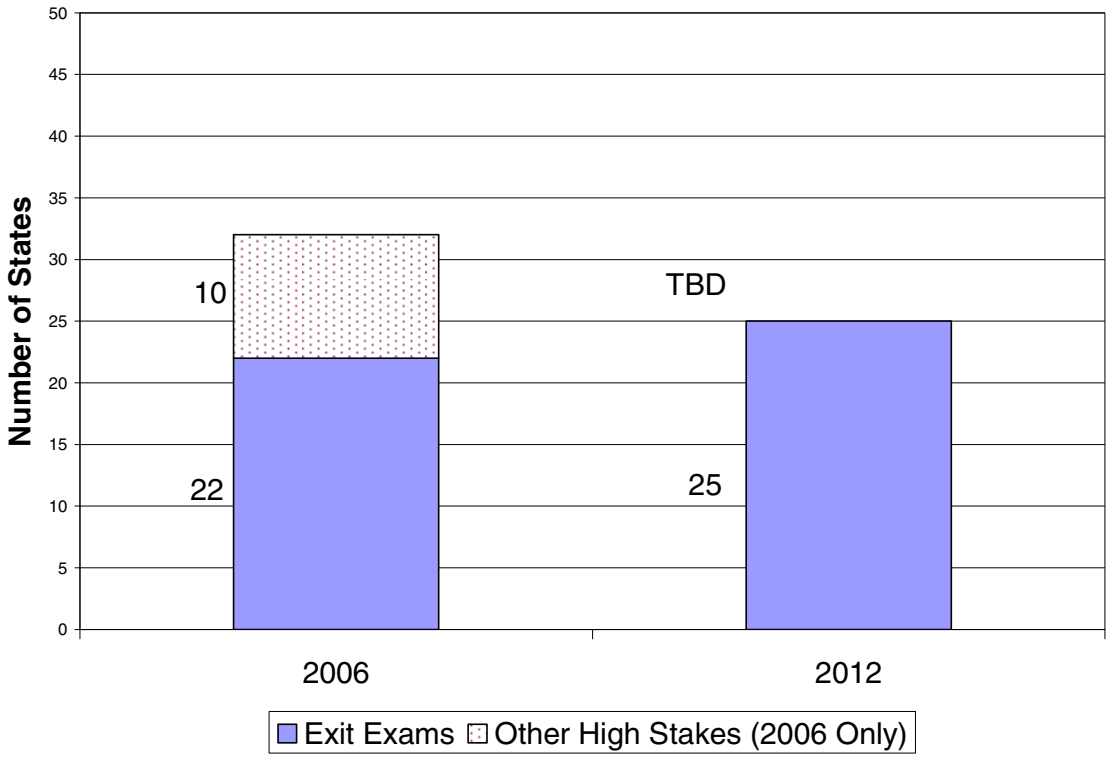


Figure 1. States with high school graduation or advancement tests in 2006 and 2012, or other high stakes (2006 only). Center On Educational Policy, 2006, pp. 1-2.

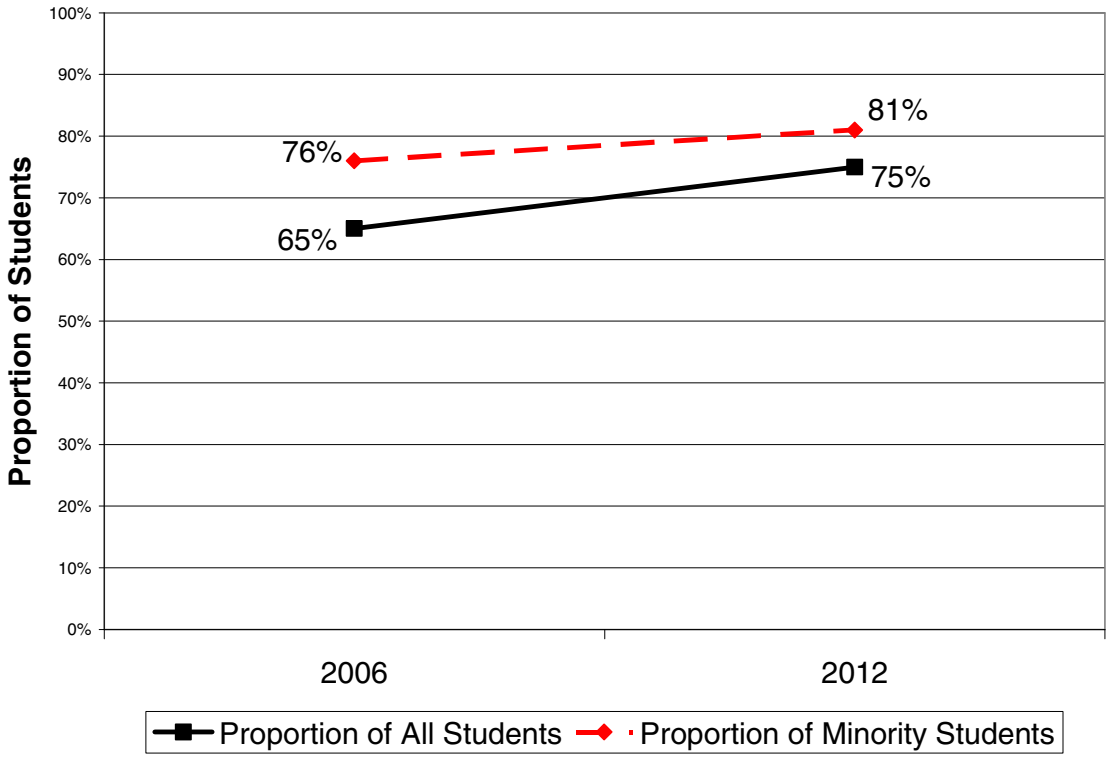


Figure 2. Proportion of US students in states with graduation or advancement in 2006 and 2012, Center On Educational Policy, 2006, pp. 1-2.

An example of a state with high stakes tests that are not graduation requirements is Michigan, in which state law not only required students to obtain a certain score on its Michigan Educational Assessment Program (MEAP) High School Test to graduate with an endorsed diploma, but also had the following intension:

> The new endorsement will be easier for colleges, universities and employers to use because the transcripts are part of the admission process and [are] often asked for by employers. Schools are now required to indicate whether or not the student earns the endorsement, the level of the endorsement and the test score on the transcript. (Michigan Department of Education, 1998)

Now, at the beginning of the 21st century, there are a number of books exploring the costs and benefits of such high stakes testing (Dwyer (2005), Thomas (2005), Firestone, Schorr, and Monfils (2004), Phelps (2004). In addition, among the most controversial aspects of the No Child Left Behind Act is the emphasis on testing. Finally, it is worth noting that a major concern is differential impact of high stakes tests on minority students, including increasing minority drop out rates. This is ironic given that a major motivation for the introduction of a major high stakes test, the SAT, was the desire to be more democratic in admissions (Lemann, 1999).

Purpose

As participants in a living discipline, the professional educational evaluators—as well as teachers and administrators--must constantly evaluate new methodologies, approaches, tools and frameworks, including those relating to the evaluation of high stakes tests. The focus of this study is to explore potential impact of one data driven intellectual technology, Bayesianism, to the evaluation of high stakes educational tests.

A Bayesian epistemological framework is closely tied to Bayesian statistical methodologies. These contrast with those of the dominant Frequentist school of statistics that has profoundly influenced educational evaluation for decades. One of the motivations for the use of statistically rigorous methodologies in educational evaluation is to put evaluation on a scientific basis. Some Bayesians, such as Howson and Urbach (1993) claimed preeminence in

scientific rigor, stating, "[The] Bayesian approach is the only one capable of representing faithfully the basic principles of scientific reasoning (p. 2)." The current study will explore a less pretentious possibility: Under what circumstances might a Bayesian epistemological framework improve educational evaluation? Specifically, are there circumstances where a Bayesian approach results in more accurate classification of examinees as meeting educational standards, a major purpose for which high stakes tests are used. Indeed, the correct classification can be seen as the very definition of validity. To this end, a Response Surface model based on data from Monte Carlo simulations is used to determine the conditions under which a Bayesian approach might result in more accurate classification than complete reliance on the Observed Score of a high stakes test alone. The creation of these conditions can be used as a *specification* for a system that would use a Bayesian epistemological framework to evaluate the results of the high stakes educational tests.

In exploring this educational evaluation question, this study will address one of the key areas for which the Bayesians claim much more capability than the those using Frequentist approaches, "[H]ow can one be certain, in any particular case, that one has selected the correct cause of an event out of the huge, indeed infinite, number of possible causes" (Howson and Urbach, 1993, p. 3). In this study, the cause of interest is whether or not the examinee truly meets a state mandated criterion. In the terms of Classical Measurement Theory, this is equivalent to having a True Score above the cut score.

Definition of Terms

Bayesian: Of or relating to Bayesian statistics or the Bayesian epistemological framework. One who practices Bayesian statistics or adheres to the framework.

Bayesian Statistics: A school of statistical though which (a) makes use of methods involving applications of Bayes's Theorem (in various forms), and (b) defines probability as a degree of belief. Note: Some would say that Bayesians recognized the use of Bayes's Theorem as the *only* way to update probability.

Bayesian Epistemological Framework: The view that Bayesian statistical methodologies provide a warranted method of inference, that is, for revising one's degree of belief in the truth of an hypothesis. See Bayes's Theorem.

Bayes's Theorem: A mathematical theorem derivable from the definition of conditional probability. The Theorem is named after the Rev. Thomas Bayes (1702-1761), whose posthumous work of 1763 is credited with the theorem's introduction, although the modern form is attributable to Laplace. Note: as a mathematical theorem, it is uncontroversial for Frequentists as well as Bayesians. It is the interpretation of the variables which is controversial.

Belief: An individual position on the truth or falsehood (or degree of truth) of a proposition.

Classical Measurement Theory (also Classical Test Theory): A theory of (educational) measurement that focuses on the test as a whole. Its cornerstone is that an Observed Score on a test is composed of two components, a True Score plus Error (symbolically O = T + E). Its results are sample dependent.

Cut Score: In educational testing, the score at or above which a test taker "passes" the test. The Cut Score is a Social Construct.

Decision Theory/Decision Theoretic: The areas of statistics and operations research dealing with rational decisions under uncertainty rather than truth or hypothesis testing.

Data Probabilities: See Probabilities, Bayesian.

(Statistical) Design of Experiments (DoE): A procedure to select combinations of factors to provide a great deal of information from a relatively small number of experiments.

Educational Evaluation: The process of determining if a student has met specific educational objectives.

Epistemology and Epistemological Framework: The branch of philosophy which deals with knowledge and justification. An epistemological framework is an approach to being confident that which is believed to be is knowledge is true, or at least useful.

Event: Something which happens, including (as a physicist might say) the event that an object exists.

Frequentist Statistics: The school of statistics which views probability of event as the limit of that event's relative frequency as the number of probability experiments of interest (which might produce that event) approaches infinity.

Item Response Theory (IRT): A theory of (educational) measurement that focuses on the item, rather than the test as a whole. It employs statistical models (with stringent assumptions) to develop Item Response Functions which provide

estimates of the difficulty of the items and the ability of students on the same scale (also called Modern Test Theory by its practitioners).

Knowledge: Undefeated, justified (warranted) true belief. Note: Many Bayesians dispense with the concept of knowledge and let belief do the work of knowledge.

Monte Carlo Simulation: A computer intensive statistical technique to simulate stochastic processes. It is particularly useful when the probability distributions of factors involved in the process are not susceptible to closed form solutions.

Observed Score: The score an examinee receives on a particular administration of a test. The reported score of a test.

Operationalism: An approach to Philosophy of Science initially suggested by Nobel Lauret P. W. Bridgman (Kasser, 2006) which seeks clarity by defining measurements in terms of strict operations. This idea was influential in physics at the turn of the 20th century. Presently used for some psychological factors, for example, a person is diagnosed as depressed if they have a certain score on a test of depression. In the Industrial Quality field, Deming insisted on operational definitions, and indicated it had three elements: A measurement system (which is in a state of statistical control), a criterion, and a decision.

Philosophy of Science: A branch of philosophy (closely related to Epistemology and sometimes closely connected with the History, Sociology, and Psychology of Science) that explores the degree to which science has a unique epistemic status for discovering knowledge.

Pragmatism: The American School of philosophy exemplified by practical applications of ideas and instrumentalism—that the meaning of an idea is in its effect.

Probabilities, Bayesian: In Bayesian Statistics, one combines a prior probability with a data probability through Bayes's Theorem to obtain a posterior probability:

- Prior Probability: The degree of belief that certain proposition is true prior to knowing certain evidence. This is sometimes referred to as simply "the prior."
- Data Probabilities: The probabilities that certain evidence would (not) occur given the proposition of interest is true (false). This is a likelihood of obtaining an experimental result given the truth (or falsehood) of the hypothesis under study.
- Posterior Probability: The degree of belief that a certain proposition is true after certain evidence is known. This is sometimes referred to as simply "the posterior."

Note: For Bayesians all probabilities are viewed as degrees of belief.

Reliability: The psychometric property of a test (instrument) which assesses the degree to which it consistently measures the same characteristic.

Science: A way of gaining knowledge which involves both discovery and justification.

Social Construct: An unobservable object, created by a group of humans, frequently embodying the group's value judgments. Meeting an educational

standard is an example of a social construct.

True Score: The score an examinee would receive on a test if there were no random error. In Frequentist terms, the long run average score of an examinee if he/she could take the test over and over without any impact due to memory, fatigue, learning, additional study or other factors. True Scores are unobserved objects.

Unobserved Object: In the Philosophy of Science, an object which cannot be observed directly (examples include quark, atom, happiness, depression, and True Score) which serves an important explanatory function in a theory. An investigator may hold that the observed object has a real, independent existence or merely use it as a useful place holder—or inference ticket—in the explanation.

Validity: The psychometric property of a test (instrument) that assesses the degree to which it measures what it purports to measure.

Objectives of the Study

The study has the following objectives:

1. To describe the Bayesian Epistemological Framework and its relationship to the Philosophy of Science.
2. To evaluate the potential effectiveness of the Bayesian Epistemological Framework in a specific application, the evaluation of a High Stakes Educational Test Result.
3. To help determine if the further development of the Research Questions of the study might constitute a fruitful programme of research.
4. To demonstrate to educational researchers the value (in terms of efficiency

and ease of analysis) of combining Monte Carlo Simulation with Design of Experiments. (Note: This combination is used in industry to assist in solving complex engineering problems).

5. To promote research programmes in education that have three elements: (a) a strong *philosophical* foundation, (b) a means, or *procedure* of application to achieve the philosophical ends, and (c) pragmatic suggestions for promoting the procedure.
6. To promote Frequentist/Bayesian interchange by using predominately Frequentist techniques to demonstrate the value of a Bayesian approach.
7. To promote interchange between Bayesian Statisticians and Philosophers of Science working with Bayesian concepts by using a form of Bayes's Theorem favored by philosophers to explore an application of Bayesian statistics.
8. To be part of the response to the urgent need for philosophical reflection on the meaning of high stakes test scores brought about by the actions of government officials who have invested educational tests with consequences that profound implications for both individuals and society.
9. To lay the groundwork for one or more refereed publications which will lead to increased understanding of Bayesian approaches as well as the meaning of high stakes test results.
10. To popularize a term, "Psychometric Modesty."
11. To lay the groundwork for one or more popular publications for which will lead to increased understanding of Bayesian approaches as well as the meaning of high stakes test results.

12. Make clear the fact that educational administrators must choose between Operationalism and admitting the existence (or at least the use) of an unobserved object, a True Score and if they choose Operationalism, realize what the costs may be.

The Research Questions

The research questions are as follows:

Philosophical: What is the Bayesian Epistemological Framework?

General: Might evaluating high stakes educational tests from a Bayesian epistemological framework result in more accurate classification of examinees than reliance on observed scores alone?

Specific: What characteristics of the prior probability of meeting Michigan State Standards on the Mathematics portion of the Michigan Educational Assessment Program (MEAP) High School Test might result in a higher proportion of the students being classified as endorsable at Level I (Endorsed, Exceeds Michigan Standards) or Level II (Endorsed Met Michigan Standards) when their True Score would indicate they should be so endorsed and fewer students being classified as endorsable if their True Score is indicates they should not be so endorsed?

Human Subjects

No human subjects were used in the empirical portion of the study. It is a Monte Carlo simulation based on the distribution of actual test scores on for Grade 11 first time test takers on the Mathematics portion of the Michigan MEAP High School Test (data in Appendix A). These data are in the public domain. The names of students or their schools or cities were be part of the dataset.

Justification

The justification for this study can be viewed from two perspectives:

From the perspective of the producer of the study, the justification is to provide an opportunity for the author learn to about the *basic* tools Bayesian statistics and, more importantly, the underlying Philosophy of Science of this school of thought. It also provides the opportunity to determine if the area of applying Bayesianism to educational evaluation is likely to be a fruitful research agenda. Finally, it provided an opportunity to survey the history of statistics leading to the Bayesian approach.

From the perspective of possible consumers of the papers that might result from this programme of research, including educational evaluators and individuals interested in applying new ideas in an applied field, the following justification is offered. The application of any new approach, whether it be in educational evaluation or another field of endeavor, should have three elements:

1. A solid *philosophical* foundation. This is needed because any methodology without such a (possibly implicit) foundation risks ad-hocism and thus limited value. A function of professional researchers is to make such philosophical foundation explicit.
2. A *procedure* for application. No matter how solid the philosophical justification for applying an idea, without a specific procedure to apply the idea there can be no application. Such procedures should have specifications, that is, criteria for evaluating their inputs and outputs. This paper focus on the specifications for an input.
3. A *pragmatic* facilitation of that application. For example, the procedure must not only exist, it must be relatively simple for the practitioner to apply.

Another justification for this paper is to explore the combination of Monte Carlo and Design of Experiment techniques in educational research. Over the past decade, this has become somewhat common in engineering applications.

In addition, some justification must be provided for the fact that this study does not develop a specific procedure for producing prior probabilities that an examinee meets the standard which the high stakes tests seeks to measure. The development of a methodology, or criteria, on which to judge such production methods is logically prior to developing those methods. Without criteria, there is no possibility of evaluating the methodology. Thus this study seeks to specify the characteristics of the distribution of prior probabilities which such methods—to be developed at a later stage of this research programme—must if they are to be considered for adoption.

This final point is a justification for the simple application of Bayesianism used in this paper, that is, employing a form of Bayes's Theorem mathematically equivalent to that contained in elementary texts that discuss the theorem. As most application of Bayesian statistical tools by educational evaluators within the next decade would be at an elementary level, it is reasonable to have the application fairly basic. This is not only because Bayesianism is new to education, but it is new to most 'non-Bayesians.' This can be seen in the debate contained in the August 1997 issue of the *AmericanStatistician* (51)2, regarding the wisdom of teaching of Bayesian statistics at an introductory level.

Assumptions and Limitations

1. The study assumes that educational evaluators are interested in the True Score of an examinee rather than the Observed Score. More precisely, that anyone interested in educational testing would be more interested in the knowing if a student actually meets an educational standard than they achieved an Observed Score above a Cut Score which is defined as the level of meeting the standard. In other words, they would be more interested in the True Score if they understood the difference between it and the Observed Score. There are alternatives, for example, that the educational evaluators will take a strict Operationalist position. Thus, they would make achieving an Observed Score a necessary and sufficient condition to be classified as proficient. However, to do so is to reject both Classical Measurement Theory and Item Response Theory.

2. The study assumes that professional educational evaluators strive to be less arbitrary, and that to be less arbitrary is to be more scientific.
3. In large part the study assumes—or more precisely uses—Classical Measurement Theory, i.e., $O = T + E$, that is, that a examinee's Observed Score is composed of their True Score plus an Error term which is independently and identically distributed normal for all examinees taking a test a specific administration and that $E \sim N(0, \sigma_{meas})$, where σ_{meas} is the standard error of measurement.
4. As only one administration of the Mathematics portion of the Michigan MEAP High School Test (Appendix A) is used as a basis for the Monte Carlo simulations, the results are sample specific and are not immediately generalizable to any other high stakes test or any standardized test.
5. The study does not work out a specific procedure for implementing an application of the Bayesian epistemological framework in educational evaluation. Rather, it explores the specifications (in terms of attributes of the prior probabilities) that such systems must meet.
6. The study involves no human educational evaluators. The specific mechanism for producing "Priors" is beyond the scope of the paper.
7. The philosophical portion of the study does not provide a complete justification for a Bayesian epistemological framework or its application in educational evaluation. Rather, it provides the justification that the Bayesian framework is reasonable.

8. It assumes, as has been frequently done in the history of statistics, that the simultaneous holding of different conceptions of probability is justified in developing pragmatic methods to solve practical problems.

CHAPTER 2

PHILSOPHICAL FOUNDATIONS AND LITERATURE REVIEW

This literature review consists of three major sections. The first is concerned with philosophy. It is a contention of the research programme of which this dissertation is a part is that a solid philosophical foundation is needed for any methodology of evaluation. Moreover, if one is to be an active participant in a system that imposes the rewards and punishments of high stakes tests on individuals and organizations, ethics demands that one is—in some sense of the term—justified in imposing the consequences. Part I, Philosophical Foundations, contains (a) an exploration of number of ideas including the importance of epistemology to this project, (b) various concepts of probability and the history of their development, (c) a brief general discussion of the Philosophy of Science, and (d) the derivation of Bayes's Theorem and relationship of Bayesianism to induction, confirmation, and the Philosophy of Science. Part II contains a discussion of Bayesian Statistics and a brief review of educational literature which uses a Bayesian approach. Part III provides a brief outline of Classical Measurement Theory. The short description other statistical methodologies employed, specifically Monte Carlo Simulation, and Design of Experiments, and Logistic Regression are presented in Chapter 3.

Part I. Philosophical Foundations

Epistemological Fundamentals

One of the founders of modern Bayesianism, Jeffrey (1980) began a classic paper with the phrase "The central problem of epistemology is often taken to be that of explaining how we can know what we do . . ." (p. 225). Starting an article with this phrase is indicative of the importance Bayesians have placed on this branch of philosophy. One might even say that some Bayesians view their methods as a process for doing epistemology.

Audi (1998), a philosopher, indicated that epistemology, "broadly concerned, is the theory of knowledge and justification" (p. 47). For Audi, this broad conception must encompass a list of concepts: "belief, causation, certainty, coherence, explanation, fallibility, illusion, inference, introspection, intuition, meaning, memory, reasoning, relativity, reliability, and truth" (Audi, 1998, p. 9). Fetzer and Almeder (1993) provided the following, more succinct, definition, "Epistemology (the theory of knowledge). The study of the conditions of knowledge and of efforts to resolve the problem of criteria" (p.47). In exploring the importance of epistemology to the rational person, Audi (1998) stated the following:

> Knowledge and justification are not only interesting in their own right as central epistemological topics, they also represent positive values in the life of every reasonable person. For all of us, there is much we want to know. We also care whether we are justified in what we believe—and whether others are justified in what hey tell us. The study of epistemology can help in this quest*, even if it often does so indirectly* [italics added]*.* Well-developed concepts of knowledge and justification can play the role of ideas in human life: positively, we can try to achieve knowledge and

> justification in relation to subjects that concern us; negatively, we can refrain from forming beliefs we think lack justification, and we can avoid claiming knowledge we think we can at best hypothesize. If we learn enough about knowledge and justification conceived philosophically, we can better search for them in matters that concern us and can better avoid dangerous pitfalls that come from confusing mere impressions with justification or mere opinion with knowledge. (pp. 9-10)

But what is this thing called knowledge, so central to epistemology? Audi (1998) provided the following formulation of this question:

> Knowledge arises in experience. It emerges from reflection. It develops through inference. It exhibits a distinctive structure. The same holds for justified belief. But what exactly is knowledge? . . . knowing is at least believing. But clearly it is much more. A false belief is not knowledge. A belief based on a lucky guess is not knowledge either, even if it is true. . . .What is not true is not known. (p. 216)

Audi (1998) indicated that Plato "has sometimes been loosely interpreted as taking knowledge to be justified true belief" (p. 217). Fetzer and Almeder (1993) in a similar vein, indicated that the standard conception of knowledge is that "knowledge is warranted, true belief" (p.26). Audi (1998) indicated that some have stated that knowledge is distinguished from mere belief when it is "undefeated justified true belief" (p. 217).

Belief is important to knowledge. Beliefs are propositions about experience (either our own or another's). Sources of belief are its grounds. Audi (1998) stated, "We have seen what at least some of the appropriate kinds of ground are: most basically, perceptual, memorial, introspective, and rational, but also testimonial and *inferential* [italics added]" (p. 244). Beliefs are the basis of propositions. However, Audi (1998), admitting the difficulty of "a straightforward analysis of knowledge which is *both* illuminating and clearly correct", suggested

that “We might say that knowledge is true belief based on the right way and on the right kind of ground” (p. 243). Audi (1998) also stated “Knowledge is often partial . . .. Still, once we get such an epistemic handle on something we can usually learn more about it” (p. 18). This task of ‘getting an epistemic handle,’ that is, doing something pragmatic to learn, is related to the Bayesian programme. There are two approaches to getting this handle: Deduction and induction.

Among others, Reichenbach (1995) contrasted the *analytical* nature of deduction with the *synthetic* nature of induction. Logical deduction is truth preserving and thus certain in the sense that conclusions must be “wrapped up in the premises (Reichenbach, 1995, p. 111).” Thus statements that are deductively logical are analytic or empty. This is in contrast to synthetic statements which add to our knowledge. Reichenbach (1995) continued, “All the synthetic statements which experience presents to us, however, are subject to doubt and *cannot provide us with absolute certain knowledge* [emphasis added]” (p. 112). Thus, there is a quandary: Belief and knowledge are closely related, and knowledge is far superior to mere belief, it may be often unattainable.

Jeffrey (1980) described the quandary of epistemology as follows

> [P]hilosophers . . . set themselves the problem of explain how we can get along, knowing as little as we do. For knowledge is sure, and there seems to be little we can be sure of outside of logic and mathematics and truths related immediately to experience . . . the rest, including most of the theses of science, are slippery or insubstantial or somehow inaccessible to us. Outside the realm of what we are sure of lies the puzzling region of *probable knowledge*—puzzling in part because of he sense of the noun seems to be canceled by that of the adjective. (p. 227)

Fetzer and Almeder (1993) stated the problem in simpler terms, “[T]he

fundamental problem confronting epistemology is that at any specific time we have no way to distinguishing warranted beliefs that are true from warranted beliefs that are false" (p. 32). Jeffery (1980) proposed to solve the problem by declaring it a non-problem:

> The obvious move is to deny that the notion of knowledge has the importance generally attributed to it, and to try to make the concept of belief do the work that philosophers have generally assigned the grander concept. I shall argue that is the right move. (p. 227)

In this Jeffrey (1980) indicated his debt to Ramsey (1980) in that "The kind of measurement of belief with which probability is concerned is . . . a measurement of belief *qua* basis of action" (p. 227). The statements of Jeffrey (1980) and Ramsey (1980) raise the questions of the role of belief and probability in the production of knowledge. This is the topic of the next section.

The Ideas of Probability

The Ideas of Probability: Historical Development

An appreciation for the rich history of the development of probability and statistics can be gained from reading David (1998), Hacking (1975, 2001), Porter (1988), Daston (1988) and Kruger, Daston and Heidelberger (1987). One of the salient points of this history is that the ideas of probability as taught today in a typical elementary statistics course are fairly new. Describing older conceptions, Hacking (1975) called probability Janus-faced. The reference to a frequent representation of the Roman god Janus as having two faces indicates that probability had two aspects. "Janus, a dual-faced god, presided over all that is

double-edged in life. His image was found on city gates, which looked both inwards and outward, and he was invoked at the start of each new day and year when people faced both backwards and forwards in time" (Cotterell and Storm, 2006, p. 56). Daston (1988) provided more than two aspects. Daston (1988) made other points of import to the three faceted research programme (philosophical, procedural, pragmatic) of this paper is a part. *First, that frequently in the history of probability different concepts of probability were held simultaneously to solve practical problems*[3]. Second, that a key element of the Enlightenment (roughly 1680 to 1790) programme was to reduce the thought process of the most enlightened to a simpler, more procedural form, which could be used by all. From the perspective of this historical context, the programme of the modern Bayesians, who take their name from one who lived in the Enlightenment, may seem all the more appropriate.

The centrality of probability to statistics was described by Savage (1972), a founder of modern Bayesian statistics, "It is unanimously agreed that statistics depends somehow on probability" (p. 2). He also stated, "Considering the confusion about the foundations of statistics, it is surprising and certainly gratifying, to find that almost everyone is agreed on what the purely mathematical properties of probability are" (Savage, 1972, p. 2). The source of this agreement, however, is, fairly recent according to Daston (1988):

---

[3] The methodology of Chapter 3 simultaneously holds Classical Measurement Theory (with its concept of a True Score and Observed Scores distributed normally around the true score) and Bayesianism (which views probability as a personalist degree of belief).

> Although the famous correspondence between Blaise Pascal and Pierre Fermat first cast the calculus of probabilities in mathematical form in 1654, many mathematicians would argue that the theory achieved full status as a branch of mathematics only in 1933 with the publication of A. N. Kolomogorv's *Grundbegiffe der Wahrscheinlichkeitsrechnung* [*Foundations of the Theory of Probability*]. . .Although philosophers, probabilists, and statisticians have since vigorously debated the relative merits of subjectivist (or Bayesian), Frequentist, and logical interpretations as a means of applying probability theory to actual situations, all accept the formal integrity of the axiomatic system as their departure point. (p. 3)

> The history of probability theory is intertwined with notions of belief and opinion, as well as with games of chance and statistics. As Hacking has pointed out, in the late medieval and early renaissance period probability was a characteristic of opinion and not knowledge; the latter was of course, certain. If we take "belief" as a synonym for opinion then on the traditional view probability is a mark of belief and not of knowledge. (p. 9)

One would hope that if probability is progressive, as other sciences (particularly mathematics) are believed to be, that the historical development of probability would lead us to one definition of probability. Kyburg and Smokler (1980) stated "It has been said (facetiously) that there is no problem about probability: it is simply a non-negative additive set function, whose maximum value is unity" (p. 4). But, as these authors stated, this statement is facetious. Thus it could be asserted by Savage (1972) that "Virtually all controversy therefore centers on the question of interpreting the generally accepted axiomatic concepts of probability, that is, of determining the *extramathematical* [italics added] properties of probability" (p. 2). Savage indicated that the controversy however was—and it must be added remains —great:

> [A]s to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. There must be dozens of different interpretations of probability defended by living authorities, and some authorities hold that several different interpretations may be useful,

> that is, *that the concept of probability may have different meaningful senses in different contexts* [emphasis added]. (p. 2)

The Ideas of Probability: Six Models of Probability

One could characterize the various approaches to probability with six models, as presented below.

1. Formal probability.

Formal Probability involves the axioms of the probability calculus. These are listed in all elementary statistical texts, for example, Berry (1996). Another formulation, which uses the notation P(a) = "the probability of event a," is as follows:

1. $0 \leq P(a) \leq 1.0$ for all a in A, where A is a probability space

2. $\sum_{a \epsilon A} P(a) = 1$

3. P(a and b) = P(a) * P(b) if events a and b are independent

2. Empirical Probability and Relative Frequency

Empirical probability is the most familiar of these interrelated models to students in a course in elementary statistics. It is the simple proportion with which an event occurs. More formally, the [limit of the] long run relative frequency of an event occurring as the number of opportunities for that event approaches infinity provides the relative frequency interpretation of probability. Kyburg and Smokler (1980) credited Venn's *The Logic of Chance* of 1886 with the first formal formulation of this approach. Von Mises *Probability, Statistics and Truth* (1951) contains another exposition. Probability according to the

empirical/relative frequency notions is sometimes called objective probability, which distinguishes it from the subjective probability.[4]

3. *Mechanical Probability*.

Mechanical Probability is akin to the relative frequency interpretation but is more deductive than inductive. An example is that of throwing in a perfectly consistent 'fair' manner a perfectly balanced six sided die, with the faces (sides) numbered 1 to 6. The probability of the die landing with a 1 face up on any given throw is 1/6. This is also the probability for obtaining the each possible result (1,2,3,4,5,6). The conceptual difference between Mechanical and Relative Frequency interpretations has to do with the application of the systems the laws with which Mechanical Probability can be deduced. One would expect a 1 to come up 1/6 of the time when throwing the perfectly thrown perfect die an almost infinite number of times. Thus, there is sometimes equivocation.

4. Logical Probability.

Kyburg and Smokler (1980) used the following description of Logical Probability, "probability is an undefinable logical relationship between one set of propositions and another" (p. 11). Kyburg and Smokler (1980) went on to say the following:

> [T]he extreme version of this alternative is to take probability as representing a logical relation between a proposition and a body of knowledge, between one statement and another statement (or a set of statements) representing evidence. Such a view was first formulated

[4] The author of this dissertation has wondered, in a humorous vein, if the truth of the long run relative frequency explanation of probability would imply that the end of the universe would be delayed if the proportion of heads flipped over the eons had not equaled exactly half the number of trials.

> explicitly by Keynes [in 1921 in *A Treatise on Probability*] and has been defended by Carnap [in 1962 in *The Logical Foundations of Probability*], Hintikka [in 1965 in "On a Combined System of Inductive Logic"], and Kyburg [in 1974 *The Logical Foundations of Statistical Inference*], among others. The essential characteristic of this view is this: given a statement, and given a set of statements constituting evidence or a body of knowledge, there is one and only one degree of probability which the statement may have, relative to the given evidence. (p. 5)

5. *Subjective or Psychological Probability*.

This is the probability a person assigns to an event. Few restrictions are placed upon this probability. For example, a weatherperson might say that 'there is a 60% chance of rain tomorrow.' Subjective probabilities need not be 'reasonable.' For example, a die-hard sports fan might say of their 'I am 110% certain of victory in tomorrow's game.'

6. Personal Probability.

This "view holds that probability measures the confidence that a particular individual has in the truth of a particular proposition, for example, the proposition that it will rain tomorrow" (Savage, 1972, p. 3). A personal probability distinguished from a subjective probability in that relevant probabilities must obey the probability calculus and principle of coherence. For example, while a person might have the following subjective (psychological) probabilities of tomorrow's weather:

P(rain) = 0.70

P (snow) = 0.01

P (sunshine) = 0.39

The sum of these probabilities is 1.10, which is more than 1.0 and violates the probability space. Thus, it is not coherent. "The notion of *coherence* of a body of beliefs was first introduced by Ramsey in 1926" (Kyburg and Smokler, 1980, p. 13). Kyburg and Smokler stated, "Having a distribution of degrees of belief which obey the conventional rules of the probability calculus of probabilities is a logically necessary and sufficient condition of not having book made against one" (p. 8). Here they refer to the so-called "Dutch Book" argument in which a person who has an incoherent set of beliefs regarding a set of events can be guaranteed to come out a net loser in a series of bets regardless of what the outcome occurs. Thus, while probabilities of different weather conditions can add to more than 1.0000 so long as they are subjective/psychological probabilities, these would not be permitted as personalistic probabilities—and it is personalistic probabilities that a Bayesian would use. Two points are to be made. Personalistic probability does not specify which probability(ies) must change to be coherent. For example, it would be acceptable for the person realizing the violation of the probability calculus to say "My probability for snow tomorrow in Detroit, Michigan is 1.0," even if the it is July 14 and today's temperature has ranged between 87 and 103. Second, the literature sometimes uses the word "subjective" to discuss all degrees of belief. This leads to confusion.

Philosophy of Science

The primary question of the Philosophy of Science can be expressed as follows: what is it about science that justifies its epistemological claim relative to other forms of attempts to provide explanations for the world (Kasser, 2006a)? In other words, what is it about science that gives science its unique epistemic claim as a special way to discover knowledge?

General Ideas

It is a contention of this paper that Bayes's Theorem provides a useful link between the Philosophy of Science (discussed in this section) and the theory of probability (discussed in the prior section). However, an understanding of the Philosophy of Science is logically prior to any exploration of the position of Bayesians. In addition, as it is an assumption of this study that educational evaluators strive to be scientific—or more accurately, that given a choice, they would prefer methodologies of educational evaluation that have a more solid scientific basis to those with less scientific basis. Thus, an exploration of what it means to be scientific is in order.

Present ideas of science are rooted in empiricism. However, this view of science—which means theories are susceptible to confirmation by observation or experimentation or that they are susceptible to falsification, particularly by statistical methodologies—is rather new in human thought. Reichenbach (1995) described the perspective of the classical writer as follows:

> To Plato, however, the concept of empirical knowledge would have appeared an absurdity. When he identified knowledge with mathematical knowledge, he wanted to say that observations should play no part in knowledge . . . Arguments from probabilities are impostors, so we learn from one of Socrates' disciples in the dialogue *Phaedo.* Plato wanted certainty, not the inductive reliability which modern physics regards as its only attainable goal. (p. 106)

Reichenbach (1995) indicated, "Plato. . . regarded mathematics as the supreme form of all knowledge. His influence has greatly contributed to the widespread conception that unless knowledge is of a mathematical form it is not knowledge at all" (p. 106). Reichenbach (1995) first quoting Plato's Republic VII, indicated:

> "Whether a man gapes at the heavens or blinks on the ground, seeking to learn some particular of sense, I would deny that he can learn, for nothing of that sort is matter of science". . . [Reichenbach stated] Empirical Science could not be rejected more strongly in these words, which express the conviction that knowledge of nature does not require observation and is attainable through reason alone. (pp. 529-30)

Thomas Aquinas (1225-1274) stated in the *Suma Theologica* (excerpt in Westphal [1995]), "But the other three intellectual virtues, namely wisdom, science, and understanding, are *about necessary things* [italics added]. . ." (p. 11). Hacking (1975) stated the following about the development of science and probability:

> In scholastic epistemology opinion was probable when well attested. Then the world began to testify by its signs. So the probable sign is the sign through which the world gives testimony. Frequency and credibility are thus linked. When conventional and natural sign are finally distinguished, it is the latter that furnish 'internal' evidence. With these transformations, in hand, the dual concept of probability was possible. (p. 180)

Speaking of René Descartes' (1596-1650) profoundly influential *cogito ergo sum* ('I think, therefore I am'—Descartes' attempt to achieve sound basis from which to start reasoning) Reichenbach (1995) stated, "The interesting question is: how is it possible that a logical issue, the attainability of certainty, was dealt with by a maze of arguments composed of [logical] tricks and theology . . . " (p. 100). Reichenbach (1995) summed up the logical/philosophical vs. the empirical approaches as follows:

> The kind of philosophy which regards reason as a source of knowledge of the physical world has been called rationalism. This word and its adjective *rationalistic* must be carefully distinguished from the word *rational*. Scientific knowledge is attained by the use of rational methods, because it requires the use of reason in application to observational material. But it is not rationalistic. This predicate would apply not to scientific method, but a philosophical method which regards reason as a source of synthetic knowledge about the world and does not require observation for the verification of such knowledge. (p. 107)
>
> He [the scientist] would insist that observation cannot be omitted from empirical science and would leave to mathematics merely the function of establishing connections between the various results of empirical investigation. (p. 106)

Hacking (1975), commenting on the development of science, stated the following:

> Opinion was the staple of low science while knowledge was the goal of high science. Paracelsus was the 'Luther of the physicians', as Copernicus was the Luther of the astronomers. One consequence of their twin revolution was that knowledge and opinion, formerly disparate, entered the same league. Or rather, what happened was that a substantial part of the potential domain of knowledge, including astronomy and the investigation of motion, became part of the domain of opinion. In the writing of Hume, the term 'knowledge' is reserved for pure mathematics. This agrees with the scholastic conception of knowledge as demonstration from first principles. But Aquinas though one could demonstrate causes and thereby explain why things are as they are. For Hume, demonstration is a matter of the 'comparison of ideas'. This operation can be performed chiefly in the realm of mathematics. Cause,

> on the other hand, is relegated to the other scholastic category that Hume variable calls 'opinion' or 'probability'. Once the concept of internal evidence was established by 1669, the final transformation needed for the skeptical problem of induction was this transference of causality from knowledge to opinion. (p. 180)

Chalmers (1995) indicated that philosophers of science seek to be both descriptive and prescriptive. Thus, they seek not merely to accurately describe the process by which the scientific community arrives at what is accepted as scientific knowledge, but also to guide practicing scientists by suggesting procedures that are in some sense 'better' to follow. Mayo (1996) indicated that "Defenders of the Bayesian Way can and do argue that even if scientists are not conscious or unconscious Bayesians, reconstructing scientific inference in Bayesian terms is of value in solving key problems of philosophy of science" (p. 102). However, Mayo (1996) rejected the contention that Bayesians have anything of value to say about science either prescriptively or descriptively. In doing so Mayo (1996) employed an analogy comparing the Bayesian view of science to viewing the Mona Lisa as a possible result of a paint-by-numbers kit. Mayo's Da Vinci would protest "I assure you I did not create it by means of a paint-by-numbers algorithm. Your ability to do this in no ways shows that the paint-by-number method is a good way to produce new art" (p. 101). It is beyond the scope of this paper to settle this ongoing and vitriolic debate. *However, it does indicate that the philosophical justification for the use of any Bayesian approach does entail a discussion of the philosophy of science.*

Chalmers (1995) stated, "The laws and theories that make up scientific knowledge make general assertions . . . and such statements are called

universal statements" (p. 3). Such universal statements contrast with "*singular statements* [which] . . . refer to a particular occurrence or state of affairs at a particular place at a particular time" (p. 3). The process of science is the arrival at *new* universal statements.

Salamon (1966), among many others, has drawn an important distinction between deduction and induction, "Questions of deductive validity are generally referred to systems of formal logic, and they usually admit definite and precise answers. Questions of inductive correctness are far more frequently answered on an intuitive or common-sense basis" (p. 111). If all scientific theories could be deduced from *known* first principles, then the philosophy of science would be as absolute as a deductive logic. In other words, with the truth of the premise established, the conclusions must be definite, precise, and true. This is because deductive logic is 'truth preserving.' Fretzer and Admeder (1993) indicated that the price for this truth preserving quality is that conclusions reached by deduction are analytic, that is, they are contained (though hidden until demonstrated by syllogism) in their premises (p. 5). However, little progress can be made in science through mere deduction from established first principles. It is the function of science to establish these first principles. Fetzer and Admeder (1993) indicated that synthetic knowledge is necessary to provide information about the world (p. 6).[5] Such syntheses require induction. Thus Popper was able to state "The empirical basis of objective science has nothing "absolute" about it" (p. 63).

---

[5] However, Fetzer and Almeder (1993) also indicated that the status of the analytic/synthetic distinction is not completely resolved (p. 6).

Models of Science

There are perhaps as many approaches to science as there are philosophers of science and scientists. Twenty-five of the major approaches to or issues in the Philosophy of Science are listed below, and several will be briefly described in this section. Kasser (2006a) discusses most of the listed topics, and many of the brief descriptions below draw on his work. The section ends with concluding comments on the Philosophy of Science and the philosophical approach of Bayesianism. The list follows:

1. THE Scientific Method
2. Classical Approaches (Descartes, Newton,)
3. Classic Empiricism (Loch, Berkley, Hume) and "Hume's Guillotine"
4. Mill's Methods
5. Bridgman's Operationalism
6. Induction by Enumeration
7. Induction to the Best Explanation
8. Realism vs. Empiricism
9. The Problem of Demarcation (Being too permissive vs. being too restrictive with respect to what is and is not called science)
10. Naturalism and Instrumentalism
11. The Hypothetico-Deductive Method
12. Logical Positivism
13. Hemple's Covering Law
14. Popper's Falsificationism and Corroboration
15. Putnam's Historical Approach
16. Fisher and Neymann-Pearson Hypothesis Testing
17. Kuhn's Paradigms, Incommensurability, and Normal Science
18. Lakatos's Scientific Programmes
19. Quine's Holism and the Web of Belief
20. The Statistical Design of Experiments (DoE) of Box, Hunter and Hunter
21. Duhem's Problem and the Role of Ancillary Hypotheses
22. Van Fraasen Answers to Why Questions
23. The Strong Program of the Sociology of Knowledge
24. Laudan's Pessimistic Induction
25. Sociological Approaches, Radical Critiques, and Feyerabend's Anarchy

THE scientific method.

Many people learn about a method called “THE Scientific Method” in high school and college. The word THE is emphasized because it is frequently taught as THE way to do science and as well as the way science is done. The following nine step process, *draws* not only on the type of presentation one might receive in pre-college (and even college) education, but also on Hemple (1966). In addition a “pre-step,” numbered zero, is included. This is to start with a question of interest (at the pre-college level, this is often described as “choose a topic”). As discussed below, this step is actually profound, touching on the theory-ladenness of observation and the necessity of auxiliary hypothesis. The method below includes the inductive steps (1, 2 and 4 below), a review of literature (step 3, which could save many engineers a lot of work), plus prediction and testing (steps 5 and 6) plus replication (step 7) and use of the knowledge (step 8) and/or an iterative approach (steps 9):

0. Start to think about a question of interest
1. Observation
2. Analysis of the observed data
3. Literature review (a step frequently skipped in pre-college work and industry)
4. Formulation of a hypothesis
5. Test the hypothesis (experiment)
6. Accept or reject the hypothesis based on the experiment
7. Replication (ideally by another experimenter at another location using other equipment and possibly slightly different methodology)
8. Use the knowledge gained through science to control nature
9. Begin again at step 1 with a more precise hypothesis.

In general, when one learns THE scientific method at the pre-professional level the specifics of steps 5 and 6 are not well specified. The most popular

specification of this is to be found in statistical hypothesis testing.

Classical approaches.

Philosophy of Science has been part of science since the Greeks, including Plato and Aristotle. While the Scholastics looked at knowledge as something which could not be a product of induction, later scientists and thinkers such as Bacon, Galileo, Descartes, and Newton sought a method of discovery that would be self-justifying. Part of the Enlightenment programme was to reduce complex intellectual work to procedures. Later developments split discovery and justification into two phases, as in THE scientific method, above.

Classical empiricists.

The Classical Empiricists, such as Locke, Berkeley, and Hume pointed out the great costs we pay when we insist that all of our knowledge must come from observation. John Locke stated, for example, that we do not hear or see a dog, rather we only have ideas (sense impressions) of the dog and are incapable of understanding how we have these ideas since we do not directly experience their object. George Berkeley went several steps further, stressing that as all we have is the sense impressions there is no dog, just the idea of the dog put in our minds by God. He did require that the ideas in various minds were coordinated in law like fashion and that it was possible for us to make predictions based on the patterns of these ideas. Hume pointed out that we never see a cause and effect relationship (for example, we do not see a billiard ball strike another and cause it to move, we only see the first billiard ball touch the other and the second move) (Kasser, 2006a).

Operationalism.

An influential approach by a Noble prize-winning physicist, P. W. Bridgman, operationalism, "requires that scientific terms be defined in terms of operations of measurement and deduction" (Kasser, 2006). This approach had a great deal of influence on scientists early in the 20th century. Among these was a student of physics, W. Edwards Deming, who became *the* leading statistician of the post-war (post World War II) "Quality Movement." As a prerequisite for quality, Deming stressed the need for operational definitions which have three parts: a) a process of measurement in statistical control, b) a criterion, and c) a decision. For example, to determine if a train is on time, one might refer to a specific clock in the train station, the criteria would be that the train is completely stopped (no forward momentum) and the passenger doors are open at the specified location on the station platform at 6:14, and the decision would be that as the train was at rest with the doors open at 6:14 according to the clock in the train station, then it was on time.

Operationalism has also had a profound effect on certain approaches to psychology, where certain diagnoses (for example, depression) are defined as certain scores on certain tests (Kasser, 2006a). Classifying a student as meeting state standards if they achieve a certain score (or higher) on a standardized test on a given day is an example of operationalism.

Induction by enumeration.

Induction by enumeration is a “common sense” way of doing induction, and to some degree, science. The classic example of induction by enumeration as the sun always has rises in the morning, it will rise tomorrow morning. However, this falls short of science because it does not explain *how it is that* the sun has risen in the past and will rise tomorrow. Moreover, without a good theory for the continued rising of the sun, a case could be made that there is some mechanism with just so many ‘rises’ and today’s could have been the last. Chalmers (1995) summed up the “naive inductivist position” by saying “science is based on the principle of induction, which we can write: If a large number of As have been observed under a wide variety of conditions, and if all those observed As without exception possessed the property B, then all As have the property B” (p. 5). This lack of currency of this approach to science was summed up by Salmon’s (1966) statement, “Not since Francis Bacon has any empiricist regarded the logic of science as an algorithm that would yield all scientific truth” (p. 112-3).

Logical positivism.

Logical Positivists, are responsible for developing the Philosophy of Science as a (sub) discipline. They sought not so much to offer proscriptive advice as to how scientists should proceed but to provide rational reconstructions of a scientific theories, that is, to reconstructions which demonstrated the special epistemic place of science. Among the most important

members of this group are A. J. Ayer, Rudolf Carnap, and Carl Gustav Hempel. Developing in Vienna and Berlin, many of the Logical Positivists fled Nazi Germany for the United States. They were empiricists, stressing that all science had to relate to what could be observed. In fact, they wanted to drive every vestige of metaphysics out of science—saying that unscientific statements were meaningless. They stressed the linguistic and formal aspects of theories, and their "conception of how scientific theories work was so influential that it is generally called the 'received view of theories.' " (Kasser, 2006b, p 29). *Kasner's discussion of Logical Positivism below is particularly important when considering the methodology of the present dissertation:*

> [In the received view there a deductive system and] The statements of a deductive system are uninterrupted, they are purely syntactic. They exhibit themselves in logical relationships, nothing more. The deductive system gets interpreted when observational terms get explained observationally. All interpretation for the Positivists comes through observation. The logical structure of the theory then lets meaning flow around the system and many statements of the theory only receive a partial interpretation in observational terms, so they are not straightforwardly true or false. Statements involving theoretical terms are considered something like inference tickets not as descriptions of the world. (Lecture 25)

The name Logical Positivism derives from two philosophical trends. The first term logic, is related to the advances in logic which began about 1870. The second, Positivism, refers to Compt's Positivism. Compt saw humanity as passing through three stages, the religious, the symbolic, and the positive (Kasser, 2006a). The Logical Positivists were hostile to metaphysics, believing that "any cognitively meaningful statement must be analytic or it must be a claim about possible experience" (Kasser, 2006a).

They were interested in the relations among ideas, not in matters of fact. They saw the purpose of Philosophy as being "to clarify linguistic problems and exhibit the relationships between scientific statements and experience" (Kasser, 2006a). They focused on "sentence sized" rather than "word sized" terms. They stress the logical relationship among statements. Thus it is appropriate to note that sentences can be viewed from three perspectives:

- Formal, that is logical,
- Semantic, that is, to what do they refer, and
- Pragmatic, to what use are they put (Kesser, 2006a).

Hempel's covering law.

Kasser (2006a) considers Hempel's Covering Law the "centerpiece of logical positivism's philosophy of explanation" and that it dominated the field of Philosophy of Science for decades. He describes it as treating "explanation as the derivation of the explanandum [thing explained] from an argument containing at least one law of nature." Okasha (2002), says that Hempel's method for answering a why question regarding a phenomenon that actually occurs, such as why does sugar dissolve in water,

> [W]e must construct an argument whose conclusion is 'sugar dissolves in water' and whose premises tell us why this conclusion is true. The task of providing an account of scientific explanation then becomes the task of characterizing precisely the relation that must hold between a set of premises and a conclusion, in order for the former to count as an explanation of the latter. . . .
> Hempel's answer to the problem was three-fold. Firstly, the premises should entail the conclusion, i.e., the argument should be a *deductive* one. Secondly, the premises should all be true. Thirdly, the premises should consist of at least one general law. . . such as 'all metals conduct electricity. . . (p. 41)

A problem with the Covering Law is the direction of causality. For example, Okasha (2002) illustrates this with the example of a flagpole's shadow. The general laws (light travels and in straight line and the laws of geometry) combined with particular facts (the angle of elevation of the sun and the height of a flagpole) explain, under the covering law the length of a shadow. However, using the same general laws but exchanging the conclusion for one of the particular facts, the length of the shadow *explains* the height of the flagpole. While it is consistent with our common sense understanding that the height of the flagpole (along with laws of nature) *will* cause the shadow to be of a certain length, we reject the possibility that the length of the shadow *causes* the height of the pole—although the Covering Law, considered by some a high point of logical positivisms (which is concerned with *linguistic* relationships)—seems to accept either conclusion.

Hypothetico-deductive method.

The essence of the Hypothetico-Deductive Method is to insist that the theory logically entails empirical evidence. Salmon (1966) describes the Hypothetico-Deductive method as follows:

> From a hypothesis, in conjuction with statements of initial conditions whose truth is not presently being questioned, a prediction is deduced. Observation reveals that the prediction is true. We conclude that the hypothesis is confirmed by this outcome. The inference is, as certain nineteenth-century theorists insisted, an inverse of deduction. By interchanging the conclusion with one of the premises it can be transformed into a valid deduction. (p. 115)

Realism vs. empiricism.

The Logical Positivists were Empiricists, and empiricism has had a strong hold on the Philosophy of Science, particularly since Einstein (Kasser, 2006a). But there is a tension between seeking an explanation (which often requires positing unobservable objects, such as atoms) versus requiring some grounding in experience (a counter-example being the apparent demise of string theory in physics). Those who seek the 'real' explanation, even if it means positing unobservable objects (such as quarks or True Scores) are called Realists. Those who believe science must confine itself to that which is observable are Empiricists. Kasser (2006a) points out the importance of this essential tension in the Philosophy of Science—how to reach the aspiration of science for a good explanation of the natural world while insisting on some grounding in 'fact' (or data).

Lauden's pessimistic induction.

Kesser (2006a) summarized Lauden's Pessimistic Induction as follows: "Most successful scientific theories have turned out to be false, so we should expect that currently successful theories will turn out to be false." Among the later falsified theories are not only the wave theory of light (which almost every 19th century physicist thought true), but the most successful of all theories, Newton's mechanics (Kesser, 2006a).

The problem of demarcation.

A fundamental problem in nearly every approach to the Philosophy of Science is how to construct a demarcation so things that should be science are not excluded (for example, atomic theory) while excluding things that are generally considered non-science (for example, astrology).

Poppers falsification and corroboration.

A problem with the naive inductive and hypothetico-deductive methods is that there are a potentially an infinite number of causes consistent with an effect (see subsection on Duhen's Problem below). For example, the success of an experimental prediction that the sun will rise tomorrow morning could be used as confirmation of the theory that the sun revolves around the earth as well as that the earth revolves around the sun.

Although the evidence can support any one of a number of theories, Chalmers (1995) indicated, "The falsity of universal statements can be deduced from suitable singular statements. The falsificationist exploits this logical point to the full" (p. 39). Thus the position Popper and his falsificationsists followers is that science should produce highly *falsifiable* hypotheses, the bolder the better. He viewed bold hypotheses which have not been falsified by strenuous tests as corroborated, although tentatively. In addition, as Popper never admits an hypothesis can be supported, much less proven, he writes of the importance corroboration, his "term for theories or hypotheses that have

survived serious attempts to refute them. Because Popper insists that corroboration has nothing to do with confirmation, he claims we have no reason to think corroborated theories more likely to be true than untested ones (Kasser, 2006). This seems related to the term undefeated in the definition of knowledge as '*undefeated* warranted true belief.'

Fisher and Neymann-Pearson statistical hypothesis testing.

Closely related to Poppers Falsificationism is the Statistical Hypothesis Testing approach of Sir R. A. Fisher. Another closely related approach is that of J. Neymann and Egon Pearson. The uncomfortable union of these two approaches is what is generally taught in statistics courses:

1. Choose a null hypothesis (and an alternative hypothesis in the Neymann-Pearson approach)
2. Choose a significance level and test statistics. The significance level determines at what level of result is said to be statistically significant.
3. Run the (randomized) experiment
4. Calculate the value of the test statistic. If the test statistic is in the "critical region" reject the null hypothesis: (a) in the Fisherian paradigm, this lack of rejection implies nothing about what is to be held; (b) in the Neymann-Pearson paradigm, this rejection is in favor of the alternative hypothesis.

Statistical design of experiments (DoE).

Design of Experiments (DoE) which is a group of methods for varying more than one factor at a time deeply influenced by Sir R.A Fisher, is an influential approach to doing science among statisticians (who frequently assist scientist and engineers). A classic exposition is *Statistics for Experimenters* by Box[6], Hunter, and Hunter (1978), affectionately referred to as BH^2. The first

[6] George E. P. Box was Fisher's son-in-law.

chapter of BH^2 contains a discussion of the Philosophy of Science, including an illustration reminiscent of Plato's cave analogy. It recommends an iterative approach with each stage of experimentation suggesting a refined (or new) hypothesis, which explored through statistically planned experimentation. Statistically planned experimentation obtains a great deal of empirical information (that is, predictive equations) from a relatively small number of experimental runs. Also, Box's famous (at least among applied statisticians) quote is worth noting: "All models are wrong; some models are useful."

In practice, using DoE frequently begins with a screening experiment to find promising (probably influential) factors and goes on to build an Empirical Response Surface Model. These models can be linear, quadratic, or higher order. The methodology stresses empirical models and prediction rather than explanation. In this way it is similar to positivism.

Putnam's Historical Approach.

In the 1970s another approach to philosophical reference not only made the discussion on unobserved objects seem more reasonable, it counteracted some of the excessive concerns on Kuhn and Fierabend concerning incommensurability (see below). This is the Causal Chain-Historical approach of Putnam and others. What this approach does is link reference to a causal chain. Thus, the wave theory and particle theory of light are theories about the same 'stuff,' light, not incommensurable theories about different thing. This approach also marked the beginning of the end of Logical Positivism (Kassner, 2006a).

This was, among other reasons, because this approach made sense that there was a necessity of identity, as opposed to the Logical Positivist idea that all necessity was formal and logical. Thus, "the Causal-Historical approach provides promising resources for enriching semantic access to the world" (Kassner, 2006, Lecture 25).

Kuhn's paradigms.

The methods discussed in the sections above share, to a greater or lesser extent, the theme of the objectivity of science. Perhaps the most influential departure from--or more precisely, variation--on this theme is that of Thomas Kuhn's (1970) *The Structure of Scientific Revolutions*, which has been called by some the most influential book of the 20th century (Kasser, 2006a). Kuhn made a distinction between normal science and a paradigm shift. In normal science a disciplinary matrix exists in which scientists engage in puzzle solving within a framework. [Khun (1991) preferred the term disciplinary matrix to paradigm in his later work.] A paradigm shift, or scientific revolution, takes place when normal science can no longer solve puzzles. Eventually, a scientific revolution takes place and the new disciplinary matrix constitutes the basis for forthcoming normal science. This new disciplinary matrix is incommensurable with that existed before the 'revolution.' Khun's approach involves looking at the scientific enterprise as predicting "theories as structured wholes" (Chalmers, 1995, p76).

Note: The present work can be seen as an exercise in normal science, although it crosses several paradigms and has implications for a 'crisis' in the

practice of educational measurement, it uses existing scientific tools and theories.

Lakatos's Scientific Programmes.

The work of Lakatos (1970) shares with Kuhn (1995) a sociological rather than 'objective' descriptive view of science. Chalmers (1995) explained Lakatos' view of the scientific enterprise in terms of 'research programmes' which have a 'hard core' which remains unchallenged as long as the research programme does not become 'degenerative.' Until that time scientists work in the 'protective belt' around the 'hard core.' It is hypothesis in this belt which can be rejected, (almost) never the hard core—at least until the programme becomes degenerative. Lakatos offered little advice on how to determine if this degenerative phase has been entered.

Duhen problem (also called the Quine-Duhen thesis).

The discussion above points out that while science strives to discover universal statements, it is difficult to develop experiments which test (form a confirmationalist or falsificationist perspective) one and only one hypothesis. This is called the Duhen Problem (or the Quine-Duhen Thesis). Fetzer and Almeder (1993) described it as follows:

> The view that hypotheses, even in science, are never subject to empirical test one by one but only in sets. The results of observations and experiments, for example, typically depend upon various assumptions other that the truth or falsity of the hypothesis under investigation. These may concern *background knowledge . . . auxiliary hypothesis . . .*and *initial condition* [emphasis added]. (p. 42)

Fetzer and Almeder (1993) used the following example of the Duhen Problem: The bishops of Padua may not have been narrow-minded in their

refusal to look through Galileo's telescope. Rather they realized that implicit in the use of this unproved instrument was an auxiliary hypothesis. Thus, Galileo was not providing conclusive evidence about the movement of the heavens (p. 42). Kasser (2006a) discussed Quine metaphor of a "Web of Belief." Generally, one can change an element of the web as long as one is willing to make adjustments on another element. In addition, one often makes changes around the periphery of the web without changing the center.

Radical critiques.

As Chalmers (1995) pointed out—and as one gains from the insights of Khun, Quine, and Lakatos—all observations are theory laden. This has led to some fundamental criticisms of the scientific enterprise. Among the critics is Feierabend. The cause of his criticism—in the Aristotelian sense of final cause that is the ultimate reason for this criticism — is that people should be free from the tyranny of science. (It might be more accurate to be free from the tyranny of scientists or the users of science, e.g., eugenicists.) At its most extreme, some critics say that science is only sexist, racist mythology, or, less polemically, a way of knowing and possibly controlling what world which has no more intrinsic validity than literature or magic. These ideas were most dramatically presented during the "Science Wars" of the 1990s (Kasser, 2006a). At its simplest, the scientific programme could be summed up by George Box, "All models are wrong, some models are useful." Thus, pushing Box into a corner, one may see the 'theory' of gravity as a myth, but it is certainly a helpful myth when one is

trying to decide about how best to get off a 30 foot ladder. As discussed below, the idea of the Bayesians of probability as a degree of belief helps this quandary.

Concluding comments on the philosophy of science.

Salamon (1966) stated "Hume's problem of the justification of induction remains at the foundations of scientific inference to plague those who are interested in such foundational studies" (p. 132). Hume, rightly, pointed out that induction is never certain. Thus, one engaged in the scientific enterprise, which has as its purpose the development of universal 'truths' which are by nature synthetic, can not only never be as sure as when relying on deduction, buy also one can never be sure at all. Lauden's Pessimistic Induction points out that many theories which were accepted by almost all scientists at a point in time (Newton's physics or the wave theory of light) have eventually been found to be false. Lauden asks if given that this is true of what was at some point considered our best science (and one might add science is our best and most successful way about gaining 'knowledge' of the world), how can we ever claim to have knowledge? (Kasser, 2006a)

As discussed in the next section, the Bayesians have cut this Gordian Knot by seeing issues associated the knowledge of truth as non issues by replacing knowledge with 'degrees of belief.' One of these positions is summarized by the statement by Howson and Urbach (1993) that "[the] Bayesian approach is the only one capable of representing faithfully the basic principles of scientific reasoning" (p. 2).

Bayes's Theorem and the Philosophy of Science

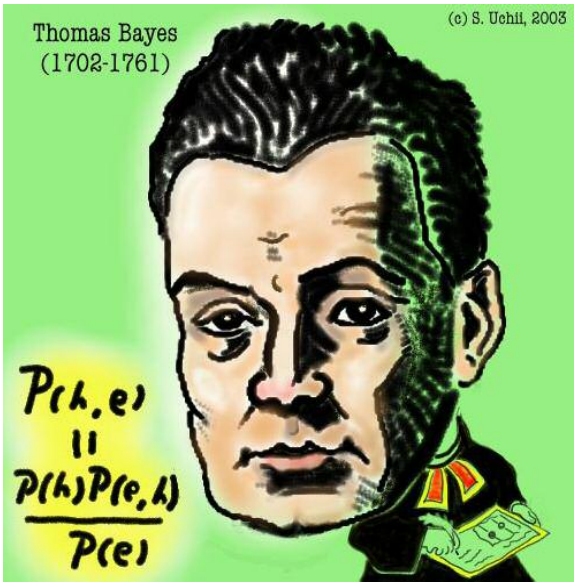


Figure 3a and 3b: Portraits of Thomas Bayes. 3a on the left is in the public domain. 3b is reproduced with permission from Professor Soshichi Uchii. It includes Bayes's Theorem Bayes holds an illustration from his 1763 article

Rev. Thomas Bayes, FRS (1702-1761) was an English Non-Conformist cleric whose posthumously [7] published work provided (1763) a proof of what has come to be known as Bayes's Theorem. Although Laplace's later formulation is perhaps more important and Stigler (1999) has pointed out it is possible that another person 'discovered' Bayes's theorem, it is from the Bayes that the both

the school of statistics and the school of Philosophy of Science takes their names.

Salmon (1966) says of Bayes's Theorem "As a theorem in the uninterrupted calculus of probability, it is entirely noncontroversial" (p 116). Before turning to the meaning of the theorem that generates controversy, the following derivation of the theorem from the calculus of probabilities is provided. In this exposition Bayes's Theorem is noncontroversial because it is based on a manipulation of the definition of conditional probability in Equation (1) below. The following notational conventions are used:

P(a) indicates the probability (however defined) of the "event a".

P(a $\cap$ b) indicates the probability of intersection of events a and b, that is, the probability that both "event a" and "event b" occur.

P(a | b) indicates the probability of "event a" occurring given that "event b" has occurred.

Note: As, in algebra, any letters can be used (consistently) rather than a and b, for example e and h. Thus equations (1a) and (1b) are both formulations of the definition of conditional probability.

$$P(h \mid e) = P(e \cap h) / P(e) \quad (1a) \qquad P(e \mid h) = P(e \cap h) / P(h) \quad (1b)$$

---

[7] On conjecture as to why Bayes did not publish this article during his lifetime is that it might be theologically suspect.

Multiplying both sides of 1a by P(e)

$$P(h \mid e)\, P(e) = P(e \cap h) \qquad (2a)$$

Multiplying both sides of 1b by P(h)

$$P(e \mid h)\, P(h) = P(e \cap h) \qquad (2b)$$

Note that the left hand side of 2a and the left hand side of 2b both equal $P(e \cap h)$, thus they are equal. Thus

$$P(h \mid e)\, P(e) = P(e \mid h)\, P(h) \qquad (3)$$

Dividing both sides of the equation by P(e)

$$P(h \mid e) = \frac{P(e \mid h)\ P(h)}{P(e)}, \quad P(e) \neq 0 \qquad (4)$$

Equation (4) is called the classic statement of Bayes's Theorem (Kasser, 2006). However, it is known that

$$P(e) = \sum_{i=1}^{n} P(e \mid h_i)\, P(h_i) \qquad (5)$$

where h_i, i = 1 to n are mutually excusive and exhaustive events and

$$\sum_{i=1}^{n} P(h_i) = 1$$

Substituting 5 into 4

$$P(h \mid e) = \frac{P(e \mid h)\ P(h)}{\sum P(e \mid h_i)\, P(h_i)} \qquad (6)$$

However, in a dichotomous case

$$P(h \mid e) = \frac{P(e \mid h)\ P(h)}{P(e \mid h)\ P(h) + P(e \mid \text{not } h)\ P(\text{not } h)} \tag{7}$$

Dividing top and bottom of the right hand side by P (e | h)

$$P(h \mid e) = \frac{P(h)}{P(h) + \frac{P(e \mid \text{not } h)}{P(e \mid h)}\ P(\text{not } h)} \tag{8}$$

Dividing top and bottom of the right hand side by P(not h)

$$P(h \mid e) = \frac{\frac{P(h)}{P(\text{not } h)}}{\frac{P(h)}{P(\text{not } h)} + \frac{P(e \mid \text{not } h)}{P(e \mid h)}} \tag{9}$$

Bayesians generally supply the following meanings in equations 10 through 15 to the terms of Bayes's Theorem:

P (h | e) = P (hypothesis is true | evidence) (10)

P (e | h) = P (evidence of experiment | hypothesis is true) (11)

Equation 11 is also called the likelihood of the evidence given the hypothesis is true. It is sometimes referred to as a "data probability."

P(h) = P (hypothesis is true prior to seeing experimental evidence) (12)

Equation 12 is called the prior probability.

P (e | not h) = P(evidence of experiment | hypothesis is not true) (13)

Equation 13 is also called the likelihood of the evidence given the hypothesis is not true. It is another data probability.

P(not h) = P(hypothesis is not true prior to having experimental evidence) = 1- P(h) (14)

Given the meanings in 10 through 14 above, Bayesians give the following *controversial* meaning to Bayes's Theorem (Equation 7 above):

$$P(\text{hypothesis is true} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{hypothesis is true}) * P(\text{hypothesis is true before we have evidence})}{P(\text{evidence} \mid \text{hypothesis is true}) * P(h) + P(\text{evidence} \mid \text{hypothesis is not true}) * (1 - P(h))} \quad (15)$$

Kasser (2006a) has provided a succinct summary of the two positions that define the Bayesian approach as well as describing its impact on both science and the Philosophy of Science:

> Bayesian conceptions of probabilistic reasoning have exploded onto the philosophical and scientific scene in recent decades. Such accounts combine a subjectivists interpretation of probability statements with the *demand* [emphasis added] that rational agents update their degrees of belief in accordance with *Bayes's Theorem* (which is itself an uncontroversial mathematical result [from the definition of conditional probability]. (Kasser, 2006)

Chalmers (1995) in a section entitled "The Retreat to Probability" criticized probabilistic approaches to science, and one would assume the Bayesians, as follows: "Their [the inductivists] technical programme has led to interesting advances within probability theory, but it has not yielded new insights into the

nature of science" (p. 19). However, Salmon (1966), who has been described as an empirical Bayesian by Fetzer and Almeder (1993, p11), summed up the importance of Bayes's Theorem as follows:

> Bayes's Theorem casts considerable light upon the logic of scientific inference. It provides a coherent schema in terms of which we can understand the roles of confirmation, falsification, corroboration, and plausibility. It yields a theory of scientific inference that unifies such apparently irreconcilable views as the standard hyothetico-deductive theory, Popper's deductivism, and Hanson's logic of discovery. (Salmon, 1966, pp. 120-1)

Kasser (2006) echoes some of the same philosophical aspects of Bayesianism as Salmon: "Bayesianism is a remarkable programme that promises to combine the positivists' demand for rules governing rational theory choice with a Kuhnian role for values and subjectivity (Kasser, 2006)." This was amplified by Howson and Urbach (1993), among the strongest proponents of Bayesianism, who say of an equivalent of Equation 9 above:

> From the point of view of inductive inference, this is one of the most important forms of Bayes's Theorem. For Since P(~h) = 1-P(h), it says that P(h|e) = f ((P(h), P(e|~h)/P(e|h)) where f is an increasing function of the prior probability P(h) of h and a decreasing function of the *likelihood ratio* P(e|~h)/P(e|h). In other words, for a given value of the likelihood ratio, the posterior probability of h increases with it prior, while for a given value of the prior, the posterior probability of h is the greater, the less probable e is relative to ~h than to h. (pp. 28-29)

Numerous books and more numerous articles have been written to carry on the debate between Bayesians and their opponents as to the reasonableness of applying scientific/epistemological meaning to Bayes's Theorem. Howson and Urbach (1993), Mayo (1996) and Earman (1992) and Kaplan (1996) are particularly worth reading. This paper takes the less extreme, more pragmatic,

position that a Bayesian approach:

- Is not unreasonable (has some support among the scientific community), and
- Is possibly useful in solving a practical problem of educational evaluation.

Again, Bayesianism is not universally accepted, either in statistics or philosophy. For example, Wayne State University, a Carnegie I institution, does not have a statistics course devoted to Bayesian statistics. There is, however, a seminar in the Department of Philosophy on Bayesian Confirmation Theory. Most introductory texts are Frequentist in nature, Berry (1996) being an exception. Hoeting (2005) stated that more and more statisticians are viewing Bayesian methods as just tool in the statistical tool box rather than a point of honor [the present authors paraphrase]. Among the reasons are the increasing advances in computing. Yen (2006), for example, indicated that Markov Chain Monte Carlo techniques, at the core of many applied uses, are "easily implemented using existing software packages or programming languages, e.g., WINBUGS (Spiegelhather, Thamas, & Best, 2000), S-Plus (MathSoft, 1995) or FORTRAN (Baker, 1998)," However, anyone with a passing acquaintance use of any of these programming languages knows the words "easily implemented" might be better stated as 'easily implemented by researchers with skill in statistical programming.' Guthrie (2006) during an introductory seminar on WINBUGS indicated that the program is far from intuitive and that people learn to use the program from other people who have learned to use the program. Thus, while there are web based applets and software macros (e.g., Albert,

1996) aimed at students who are learning statistics at an introductory level, introductory statistics are predominately Frequentist or Data Analytic[8].

Among philosophers, Kasser (2006) has stated:

> Predictably, a Bayesian backlash has also been gaining momentum in recent years. [Among the objections to Bayesianism are its] surprisingly subjective approach to probability assignments, as well as the Bayesian treatment of the problem of old evidence (it appears that we can never learn anything from evidence that is already in).

Glymour (1980) and Mayo (1997) are particularly strident in their opposition to Bayesianism as indicated by the literary allusions of these use in their titles. Glymour's major anti-Bayesian article is entitled "Why I Am Not A Bayesian." This could be a literary allusion to Bertrand Russel's *Why I am not a Christian.* The title of Mayo's 1997 article "Duhem's Problem, the Bayesian Way, and Error Statistics, or "What's Belief Got To Do With It?" reminds one (at least this one who like rock music) of Tina Turner's classic, "What's Love Got to Do With It" (Turner, et. al. 1984).

Speaking of human mental cognitive capabilities, Kasser (2006a) stated, "We do not have the processing power to meet Bayesian standards even in the fairly simple cases. Coherence requires logical omniscience, namely, that we know all the logical consequences of our beliefs, and that is unrealistic." (Kasser, 2006a, lecture 32). However, Kasser went on to point out that this is not a

---

[8] The 'data analytic' approach uses a great deal of Exploratory Data Analysis. However, even books with a large data analytic component (e.g., Moore and McCabe (1993) still have a significant hypothesis testing component. Authors generally do not go too far away what the market wants. A possible exception is Horel and Snee (2001).

problem unique to Bayesianism, for even non-Bayesian logical coherence is beyond human mental capabilities.

Other objections described by Kasser (2006a) are that scientists do not act like Bayesians (reporting priors and posteriors in published research), that Bayesians have a "somewhat brazen tolerance of subjective probabilities. . ." , and, perhaps the two most serious concerns: difficulties having to do with the nonexistence of a 'catch all hypothesis,' and the problem of 'old evidence' (Lecture 32).

It may be true that scientists have certainly not acted like Bayesians in the past, to some degree they have not acted like Frequentists who reject a null hypothesis: They are interested in establishing *their* hypothesis (see Howson and Urbach, 1993). Moreover, scientists from many disciplines are increasingly using Bayesian methods and collaborating with and/or employing statisticians who are Bayesians (FDA, 2006; Hoeting, Lecture 1, Colorado State University 2005).

The old evidence problem is explored in Earman's (1992) *Bayes or Bust*, Mayo (1997), and many other works. The problem can be seen in the classic formulation of Bayes's Theorem, equation (4). If we know something has happened, that is, it has actually happened,[9] and thus probability of its happing must be 1. If the probability of any event is 1, its conditional probability on *any* other event is also 1. Thus, "because P(E) is 1, P(E/H) is 1 "(Kasser, 2006,

---

[9] We are leaving aside the possibility that we are mistaken in our belief that the event has happened.

Lecture 32). This is a mathematical truth. If P(E/H) is 1, simple algebra (1 times P(H) in the numerator divided by 1 in the denominator) leaves us with the result that the posterior probability, P(H|E) is equal to the prior probability P(H). Thus looking at old evidence does not *and mathematically can not* change result in a posterior probability different from our prior probability—even if one is a "true" Bayesian who has a completely subjective probability and is committed to updating probabilities only using Bayes's Theorem.

Kasser (2006a) points out that our common sense understanding of scientific progress counts explaining old evidence positively. He provided the following example: The fact that Newton's theory entailed Kepler's laws of planetary motion, explained the tides, and provided a unified theory explaining a number of other phenomena was counted as evidence for the Newton's theory. He goes on to describe some of the ways Bayesians have responses to the problem of old evidence, including the ideas that one should perform the mental gymnastics of determining a prior excluding this old evidence, i.e., use a counter factual. Another approach (using the Kepler/Newton example) has been to say that one is relying not on the *old* evidence of Kepler's laws to confirm Newton's theory but on the *new* information (evidence) that Newton's theory entailed Kepler's laws (Kesser, 2006a). The present author proposes a somewhat kindred, but far simpler solution: As Bayes's Theorem contains variable to which humans supply meaning, merely define the variable "e" not as "evidence" (new or old) but as a "subjective experience of evidence." Thus one could become more and more convinced of a proportion each time one thinks about it, whether

the evidence one is considering during this thinking is new or old. *This may be descriptive of the way people update their beliefs when weighing the pro and con of everyday arguments over and over again.* Moreover this solution to the problem of old evidence is completely consistent with a truly subjective understanding of probability. It is possible to object that this could lead to runaway increases (or decreases) in one's degree of belief in the truth of a hypothesis on exceedingly weak evidence just by thinking about the evidence over and over again. The only response the present author has to this complaint is to say that it is absolutely true (just as it is true that Bayesians can tolerate 'bizarre' probabilities) and that, perhaps, a society may want to reserve the term 'science' for those who insist on updating of belief by follow certain (possibly Popperian) guidelines.

The 'catch all problem' is apparent in equation 9 when one realizes that there are frequently an infinite number of competing hypotheses to the hypothesis under investigation:

$$P(e|\text{ not } h) = \sum_{i=1}^{\infty} P(e \mid \text{not } h_i) \qquad (16)$$

This is not a problem for this dissertation because we have a dichotomous case (the student either meets or does not meet the standard). One might solve the problem by noting that we are dealing with subjective probabilities and the individual can constrain these to total less than one.

Part II. Bayesian Statistics and Bayesian Educational Research

After a discussion of Bayesianism from the point of view of Philosophy of Science and before reviewing educational literature which employs Bayesian methodologies, it is worthwhile to step back and describe Bayesian statistics, particularly as the statistical techniques used this paper might be considered Frequentist—or at least not exclusively Bayesian. An excellent description is to be found in the United States Food and Drug Administrations Draft "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials" (FDA, 2006). This paper is recommended to any readers interested in a good *description* not only of Bayesian statistics, but why it is being considered by a government regulatory agency for use in matters which are, literally, of life and death. The term description is stressed because Draft Guidance contains not a single equation, Greek letter, or computation! Several quotes from this document are provided below. Although some of the comments specifically discuss clinical trials, the logic of methodological choice is applicable in any field, including[10] education.

> What is Bayesian statistics?
> Bayesian statistics is a statistical theory and approach to data analysis that provides a coherent method for learning from evidence as it accumulates. Traditional (frequentist) statistical methods formally use prior information only in the design of a clinical trial. In the data analysis stage, prior information is considered only informally, as a complement to, but not part of the analysis. In contrast, the Bayesian approach uses a consistent, mathematically formal method called Bayes's Theorem for

---

[10] It is worth noting again that one of the classical papers on applied Bayesianism was in a Psychological journal. (Edwards, Lindman, & Savage, 1963)

> combining prior information with current information on a quantity of interest. (FDA, 2006, Section 3.1)
>
> Why use Bayesian statistics for medical devices?
> When good prior information on clinical use of a device exists, the Bayesian approach may enable FDA to reach the same decision on a device with a smaller-sized or shorter-duration pivotal trial. . .
>
> Bayesian methods may be controversial when the prior information is based mainly on personal opinion (often derived by elicitation methods). The methods are often not controversial when the prior information is based on empirical evidence such as prior clinical trials. (FDA, 2006, Section 3.2).
>
> Why are Bayesian methods more commonly used now?
> Bayesian analyses are often computationally intense. However, recent breakthroughs in computational algorithms and many-fold increases in computing speed have made it possible to carry out calculations for virtually any Bayesian analysis. These advances have resulted in a tremendous increase in the use of Bayesian methods over the last decade. See Malakoff (1999). The basic tool that enabled the advances is a method called Markov Chain Monte Carlo (MCMC). For a technical overview of MCMC methods, see Gamerman (1997). (FDA, 2006, Section 3.3)

One final point about Frequentist vs. Bayesian methods is in order. Many students taking perhaps one or two courses in statistics (much less statistics at a pre-college level or in a fast paced industrial seminar)—as well as engineers, scientists, and other researchers—think that statistics methods give us numbers about the truth of a proposition regarding a substantive subject being studied. Informed by the discussion of Philosophy of Science above (for example, the discussion of Popper) and/or a careful reading of the good statistics text books, one is faced with the inescapable conclusion that the probabilities in Frequentist hypothesis testing are about the reliability of methods *if the truth were know* and those in Bayesian statistics are about *beliefs about the truth*, and *neither is or*

can be about the truth of the subject-matter proposition being studied. It is just that Bayesians are more honest (some Frequentists might say pompoms) about this distinction.

Although increasingly used, Bayesian analysis is rather rare in educational research, with most applications in Item Response Theory (IRT) and Computer Adaptive Testing (CAT). A published literature review found no applications of Bayesian approaches in determining if an individual had demonstrated proficiency on a high stakes educational test. Elmore and Woehlke (1996) found *only three* papers employing Bayesian methods in a content analysis of 1,715 papers published from 1978 to 1995 in three educational research Journals (*American Educational Research Journal, Educational Researcher, and Review of Educational Research)*, however

> Results are consistent with those of other studies in that the most commonly used methods were ANOVA and ANCOVA, multiple regression, bivariate correlations, descriptive statistics, multivariate analysis, nonparametric statistics and t-tests. The major difference in current methodology is the increased in the use of qualitative methods. (Abstract).

Haig (1996) recommended that "statistical inference practices in both educational and psychological research should be directed away from traditional significance tests in favor of Bayesian inferential methods" (abstract).

One of the papers in the Kotz and Johnson's (1993) *Breakthroughs in statistics Volume I: Foundations and basic theory* was Edwards, Lindman, and Savage (1993). In addition, Pollard (1989) continued the theme of how research might be done with a Bayesian statistics. However, Nunnally and

Bernstein (1994) devoted perhaps no more than four of its 700 pages to Bayesian-related topics. These include a brief explanation of Bayes's Theorem and references to computer adapted testing and IRT estimation procedures.

In addition to a brief overview of the Bayesian updating, Hambleton, Swaminathan, and Rogers (1991, pp. 38-9) listed various IRT computer programs with Bayesian estimation techniques, ASCAL, BILOG, MULTILOG (pp 49-50), and indicate "Swaminathan and Gifford (1982, 1985, 1986) have developed Bayesian procedures for one-, two-, and three-parameter models in which the prior distributions are placed on item and ability parameters" (p. 43). However, the emphasis here seems to be on parameter estimation, with the ultimate interest in the items, not on use with consequences, for they stated, "This procedure eliminates the problems encountered in the joint maximum likelihood procedure, namely that of improper estimates for certain response patterns" (Hambleton, Swaminathan, and Rogers, 1991, p. 43).

Bock and Mislevy (1982) explored Bayesian approaches to adaptive Estimation of Ability in a Microcomputer Environment. Huynh (1998) suggested a Bayesian approach for partial credit scoring.

Two major 2006 publications, research handbooks sponsored or published by major educational associations, provide some up to date summaries of the use of Bayesian (or more precisely, Empirical Bayes[11], EB)

---

[11] Difference between Bayesian and empirical Bayes is that "In a fully Bayesian solution, he distributional form of the prior, as well as the hyperperameters, are specified *a priori*. . .By contrast, in EB [Empirical Bayes], only the distributional form is specified beforehand. The hyperparameters are not specified. Rather they are estimated simultaneously. . ." (Braun, p. 248).

approaches. Chapter 14 (of 46 Chapters) in the *Handbook of Complementary Methods in Educational Research* edited by Green, Cammilli, and Elmore is Henry Braun's "Empirical Bayes." Braun states in his conclusion:

> Indeed, Bayesian methods generally are in the ascendancy—it appears that there is a greater willingness to accept the notion of introducing subjective beliefs into an analysis through the mechanism of a prior distribution. This is due, in large measure, to the increase in accessible computing power and powerful software that make Bayesian modeling feasible—along with many successful examples of Bayesian date analysis ([Gelman, Carlin, Stern & Rubin, 2003]) (Braun p. 256).

While most of Braun's chapter is devoted to exposition of EB, he also described a number of applications in educational research. He describes several examples of hierarchical modeling, for example, developing equations ". . . to predict whether a student will achieve a fixed standard on an examination or whether he or she will graduate from high school by a certain age." (Braun, p 252). In a hierarchical model, "we want to estimate a set of such prediction equations, one for each school" (Braun, p 252). Braun cited Wong and Mason (1985), Raudenbush and Bryk (2002) in this vein. He also indicates that Zwick & Braun (1993) produced studies on "a closely related, but technically more challenging problem . . . estimating the probability distribution governing the time to the occurrence of some event" (Braun, 2006).

Braun (2006) discussed the use of EB in metaanalysis, particularly in studies of Validity Generalization, citing among other Hedges (1988) and Braun (1989). He also discussed value-added models judging teacher effectiveness, which "have been used in Tennessee since 1993 and have been adopted by a number of districts in other states (Braun, 2006)." These mixed models

separately estimate the district effect (a fixed effect) and the teacher effect (a random effect), and a set of equations over multiple years is called a layered model. Among the sources Braun (2006) suggests are Sanders, Saxton, & Horn (1997, Ballou, Sanders, and Wright (2004), Harville (1976), Kupermintz (2003), McCaffey, Lockwood, Koretz, Louis, and Hamilton (2004), and Braun (2005).

Several pages of Brennan's 700-plus page *Educational Measurement*, Fourth Edition, 2006),[12] have headings focusing on Bayesian methods. In Chapter 5, "Scaling and Norming," Kolen states "IRT proficiency can also be estimated using Bayesian methods. . ." and he describes the Bayesian *expected a posteriori* (EAP). He also points out that "Unlike maximum likelihood estimates, Bayesian EAP estimates exist for all response patterns. . . " (Kolen, 2006). On the other hand he cautioned:

> Similar to the Kelley regressed scores, the EAP is a biased estimate. Examinees whose θ is above the mean in the population, on average, have estimates that lower their θ. Examinees whose θ is below the mean in the population, on average, have estimates that are greater than their θ. A consequence of this bias is that in a given population the variability of the EAO estimates of proficiency typically is less than the variability of θ because of shrinkage toward the mean for the EAO estimates. In addition, as with the Kelley regressed scores, the EAO estimates depend on the distribution in the population. (Kolen, 2006)

In the Chapter 4 of Brennan (2006) "Item Response Theory" (Yen and Fitzpatrick, 2006) the authors stated:

> It is possible to estimate abilities using statistical procedures other than MLE. The most commonly used alternatives involve Bayesian methods, in which a *prior* distribution is assumed for the parameter being estimated.

[12] A volume jointly sponsored by the National Council on Measurement in Education and the American Council on Education

> The more confident the user is about the prior information, the smaller is the standard deviation of the prior distribution. The prior information comes from knowledge about examinees that is external to the test, such as concurrent performance on another test (in which case the estimate is an *empirical* Bayes estimate) or the user's belief or knowledge about how a similar population of examinees performed on the test previously. (Yen and Fitzpatrick, 2006, p138)

Yen and Fitzpatrick (2006) also stated "Empirical Bayes estimators have been successfully used in combination with IRT models to provide more accurate subscores on achievement tests (p.138). Among the works they recommend are Wainer et al., (2001) Yen, (1987), and Yen, Sykes, Ito, & Julian (1997). Yen and Fitzpatrick (2006) discussed Markov Chain Monte Carlo (MCMC) methods, which "are usually carried out using Bayesian models," for estimation in more complex unidimensional and multidimensional IRT models. Citing a number of studies, starting with Albert (1992), they conclude "Results to date for MCMC methods are promising, but very preliminary. Although easily implemented using software packages or programming languages. . . MCMC methodology is conceptually complex, and the time required for parameter estimation is very lengthy."

Based on the above, it is clear that Bayesian concepts, methodologies (e.g., EB), and tools (e.g., MCMC) are increasingly used in educational research, particularly in conjunction with Item Response Theory. However, the literature review found no paper which had the focus of the present work, use of a *basic* simulation and a *simple* form of Bayes's Theorem (see equation 9) to determine if students can be classified more accurately (when Classical Measurement Theory's True Score is the standard) by employing a Bayesian approach than would be possible using Observed Scores alone.

Part III. Classical Test Theory

As this paper relies heavily on Classical Test Theory, it is appropriate to describe its essential points. Classical Test Theory considers problem in educational testing, specifically how to formally address the common sense notion that the score a person receives on a given day with a given administration of a given form of a given test is not the only possible score they could have received on *that* test—must less that this score is the *only* possible indicator of how they should be classified according to the social construct the test purports to measure. High stakes tests attempt to measure where a student does or does not meet an educational standard. A simple parable illustrates the difficulty in relying on such tests.

Imagine a state has a requirement to pass a social studies test with a score of 20 or more out of a total of 40 possible points (one point per multiple choice questions). Imagine further that there are four students, each of whom "just meets" the state educational standard at 7:00 a.m. on the day of a social science high school graduation test be to administered at 8:00 a.m.. The first student's mother had decided in advance—and told him—she would drive him to school that day. His i-pod is broken and his mother is turned to NPR which has a story about the Supreme Court and the difficulties a bill is having in Congress. Two questions of 40 that appear on the test have to do with the functioning of the Supreme Court and congress and the student gets them correct only because of

the information they picked up from the radio program[13]. The second student has been told by her parents consistently throughout the semester that they believe she will do well, to relax, and remember if she just happens not to pass there will be other opportunities to take the test which she will be sure to pass. The third student, who is from time to time beaten by his alcoholic father, has been told by the father as he leaves home to take the test that if fails the father will break his neck—and the student believes the father is being literal. The fourth student broke up with her boyfriend Friday because she found that he was sleeping with her best friend. She felt sick Saturday. On Sunday night she took a home early pregnancy test and it was positive. She had only a few hours sleep all weekend. Our common sense intuition that although cognitively at 7:00 a.m. all four students were "just proficient" in social studies as defined by the state, it is unlikely that all of four students would obtain the same score of 20 on the test, the score of a person who is 'just proficient[14]'. The 'unobserved object' of a True Score in Classical Measurement Theory gives theoretical content and the potential for quantification to this intuition.

---

[13] This story is based on the present author's experience of realizing he had not taken a course in Political Philosophy and reading summary chapters of two paperback anthology of political philosophy books the night before the Political Science GRE, on which there were two questions he believes he got correct because of this cramming.

[14] The present author has was appointed to and participated in an advisory committee which recommended cut scores for an administration of the Michigan MEAP High School Mathematics Test. A modified Angoff process was used by the consultants that facilitated the group.

Nunnally and Bernstein (1994) discussed Classical Measurement Theory at length. They point out that the basic concept is:

$$O = T + E \quad (17)$$

where O is the observed Score, T is the True score, and E is an error component. The true score is the score the examinee would obtain if there were no extraordinary factors. Such factors could include cramming, text anxiety, personal concerns (positive or negative, for example, winning the lottery), language difficulty, etc. In a Frequentist sense, T is the average score the examinee (at a given level of development and study) would get in the "long run" by taking the test over and over (but with one test score not influencing another).

Nunnally and Bernstein (1994) indicate that the true score, in deviation units, can be estimated from the observed score and data on the test, "True deviation scores (t') are thus estimated as the product of the reliability coefficient and the obtained deviation score [x]

$$t' = r_{xx}x \quad \text{(Nunnally and Bernstein, 7-3, p. 259)} \quad (18)$$

Nunnally and Bernstein (1994) went on to point out that a confidence interval can be built around the True score

$$t' +/- 2\ \sigma_{meas} \quad \text{(Nunnally and Bernstein,, p. 240, 260)} \quad (19)$$

where $\sigma_{meas} = \sigma_x \sqrt{(1 - r_{xx})}$ (Nunnally and Bernstein, 6-34, p. 239) (20)

A data set for one administration of the Mathematics portion of the High School Proficiency Test (Appendix A) was obtained along with information on its psychometric properties of that administration, the reliability (r_{xx}) and the sample standard deviation of scores s_x which is an estimate of σ_x .

Before discussing the methodology of this study, it is worthwhile to reflect on the fact that one of the calculations involves determining the probability that a student who actually meets (or does not meet) the standard will be correctly classified. One report of note has addressed a similar problem, the analysis of the accuracy of standard tests, Rogosa (1999) which was was federally funded research at Stanford's. Viadero (1999) in her *Education Week* report on the study stated:

> "How often will a student who really belongs at the 50th percentile according to national test norms actually score within 5 percentile points of that ranking on a test?
>
> The answer, a Stanford University statistician says in a new report, is only about 30 percent of the time in mathematics and 42 percent in reading (p. 3).

CHAPTER 3

METHODOLOGY

This Chapter has four sections. The first discusses the classical (non-Bayesian) statistical methodologies used in the study. Next a short thought experiment is presented to motivate an understanding of the study. Then a simplified version of the study and corresponding flowchart (Figure 11) is presented. The final section contains a more detailed description of the steps of the study (also see Appendices B through D).

Statistical Methodologies Used

In addition to the discrete form of Bayes's Theorem as presented in Equation 9 and Classical Test Theory, the empirical portion of this study relies on three statistical methodologies: (a) Monte Carlo Simulation which will be used to generate the data (Figure 11), (b) Statistical Design of Experiments (DoE), specifically Response Surface Methodology (RSM), to plan the simulations and analyze the results, and (c) Logistic Regression to do some of the intermediate calculations for a term in Bayes's Theorem.

Monte Carlo Simulation

Monte Carlo Simulation is a statistical technique to simulate stochastic phenomena. Using a computer, datapoints are drawn at random from a theoretical (such as the normal) or empirical statistical distribution. However, *physical* simulation predates computer simulation. Stigler (1999), in an article entitled "Stochastic Simulation in the Nineteenth Century," recalled a number of

physical simulations:

> There has been a tendency in recent years to date the use of simulation in statistics only from the early years of the twentieth century. For example, Teichroew (1965) and Irwin (1978) suggest that its earliest appearance may be in "Student's" classical investigation of the t-statistic (Gosset, 1908a), where "Student" (William S. Gosset) generated 750 samples of size 4. He accomplished this by shuffling 3,000 cards labeled with anthropometric measurements on 3,000 criminals, and he groups. Gosset also used the same generated samples in his investigation of the correlation coefficient (Gosset, 1908b) . . . Even more sophisticated uses of the techniques can be found in earlier literature, however, and I shall present three such examples from the last quarter of the nineteenth century. . .
>
> All concern simulation in the modern sense of the word, a modern stochastic art for the study of statistical science. All three involve generation of half-normal variates and the separate assignment of randomly generated signs to the variates, but the three involve three different randomizing devices. De Forest drew labeled cards from a box, [George H.] Darwin [son of Charles Darwin and a cousin of Francis Galton] used a spinner, and [Sir Francis] Galton used a special set of dice. . .

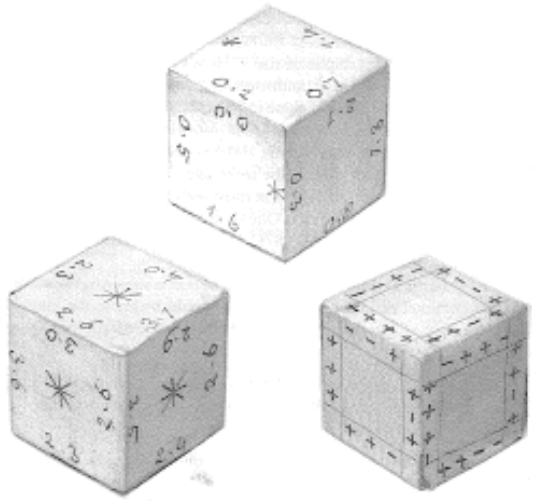


Figure 4: "Photographs of the three types of Galton's dice. . . from about 1800, perhaps the oldest surviving device for simulating normally distributed random numbers. They are presently in the Galton Collection at University College London (Stigler, 1999, Figure 7.1, p. 144, reproduced by permission of Dr. Stigler)."

Van Matre and Slovensky (2000), in an article aimed at educators involved in teaching Statistical Quality Control, stated the following:

> A pedagogic history of quality management would reveal a rich tradition of using innovative games, exercises, and experiments to convey effectively key quality principles to participants. Such demonstrations go back at least to 1931 and Walter Shewhart's bowl. Shewhart used the data of 4000 drawings (with replacement) from a bowl of 998 numbered chips to demonstrate the principles underlying his control chart. [W. Edwards] Deming. . . is famous red beads taught "by experiment a number of important principles" (1993, 158); for example, the failure of bonuses to change worker productivity in a common cause system. His funnel experiment demonstrated the detrimental effects of tampering; that is, treating common cause variation as if it were due to an assignable cause.

Moving into the computer age, Stanislaus Ulam's (1991) story of the development of Monte Carlo simulation—which had a role in solving mathematical problems connected with the development of atomic weapons for which there was no closed form solution—is quoted at length because of its importance in the development of a key methodology used in this paper:

> Two seminar talks I gave shortly after my return [to Los Alamos during the Manhattan Project] turned out to have good or lucky ideas and lead to successful further developments.
>
> The second talk was on probabilistic calculations for a class of physical problems. The idea for what was later called the Monte Carlo method occurred to me when I was playing solitaire during my illness. I noticed that it may be much more practical to get an idea of the probability of the successful outcome of a solitaire game (like Canfield or some other where the skill of the player is not important) by laying down the cards, or experimenting with the process and merely noticing what proportion that comes out successfully, rather than to try to compute all the combinatorial possibilities which are an exponentially increasing number so great that, except in very elementary cases, there is no way to estimate it. This is intellectually surprising, and if not exactly humiliating, it gives one a feeling of modesty about the limits of rational or traditional thinking. In a sufficiently complicated problem, actual sampling is better than the examination of all the chains of possibilities.
>
> It occurred to me that this could be equally true of all processes involving branching of events, as the production and further multiplication of neutrons in some kind of material containing uranium or other fissile elements. . . The elementary probabilities of each of these possibilities are individually known, to some extent, from the knowledge of the coresss

> sections. But the problem is to know what a succession and branching of perhaps hundreds of thousands or millions will do. One can write differential equations or integral differential equations for the "expected values," but to *solve* them or even to get an approximate idea of the properties of the solution, is an entirely different matter.
>
> The idea was to try out thousands of such possibilities and, at each stage, to select by chance, by means of a "random number, with suitable probability, the fate or kind of event, to follow it in a line, so to speak, instead of considering all the branches. After examining the possible histories of only a few thousand, one will have a good sample and an approximate answer to the problem. All one needed was to have a means of producing such sample histories. It so happened that computing machines were coming into existence and here was something suitable for machine calculations. . .
>
> [John] Von Neumann played a leading role in the launching of electronic computers. . . .
>
> The Monte Carlo method came into concrete form with its attendant rudiments of a theory after I proposed the possibilities of such probabilistic schemes to Johnny in 1946. . .After this conversation we developed together the mathematics of the method. It seems to me that the name Monte Carlo contributed very much to the popularization of this procedure. It was named Monte Carlo because of the element of chance, the production of random numbers with which to play the suitable games. (Ulam, 1991, pp196ff)

In 1964 Hertz, then a Director at McKinsey & Company (a major business consulting firm) sought to popularize the use of Monte Carlo simulation in business with a Harvard Business Review article, "Risk Analysis in Capital Investment." A companion article was published in the same executive-oriented journal in 1968, "Investment Policies that Pay Off." In one of these articles (1964) Hertz used an illustration of the possible outcomes of dice to illustrate a distribution (one way to get a "two," two ways to get a "three," six ways to get a "seven,", etc.). In the other, Hertz (1968) provided a conceptual illustration in which he used 'spinners' to represent how a computer might select values of possible distributions of variables that go into the calculation of Return on

Investment (market size and share, marketing and selling costs, fixed and variable manufacturing costs, and investment). Looking back on the article from the perspective of desktop computers, one can see how far the technology has advanced the calculation possibilities since the days of mainframes and punchcards, “A computer can be used to carry out the trials of the simulation method in very little time and at very little expense. Thus for one trial, 3,600 discounted cash flow calculations, each based on a selection of the nine input factors, were run in two minutes at a cost of $15 for computer time.”

Since these articles, PC software packages, such as @Risk (Palisade Corporation, 2002), are available as Excel spreadsheet add-in for a wide variety of business and technical problems. The present author was able, once provided with engineering equations, to use the @Risk program to estimate the distribution of distance required to stop a truck given inputs such as tire characteristic, brake pad friction, and driver reaction time.

Moving to the realm of educational research, Sawilowsky (1990) indicated Monte Carlo Simulation is an appropriate, and some times the only, method to evaluate statistical methodologies. It is frequently faster than developing closed form solutions, if such solutions exist. If they do not exist, it may be the only means available. Micceri (1989) has pointed out that few distributions of data in education follow a normal distribution. Thus one should consider working with empirical, rather than smooth theoretical, distributions. When an empirical distribution is used an index number is assigned to each datapoint in the distribution. Using a random number generator there are drawings (usually with

replacement) from a uniform distribution so that each of the index numbers an (approximately) equal chance to be selected. This corresponds to a datapoint is then available for further manipulation. This process is repeated, generally thousands of times, to obtain stable results. If a theoretical distribution is used, for example, the normal, random numbers are chosen from a uniform, and converted to normal deviates so that a the chances of a number within a range being selected are proportional to the probability of the normal within that range.

In this paper Minitab® Statistical Software (Minitab, 2000) is used to perform the Monte Carlo simulations.

Statistical Design of Experiments

Design of Experiments (DoE) was discussed from a philosophical point of view in the Chapter 2. This section briefly describes Statistical Design of Experiments (DoE) as a methodology which takes advantage of geometrical/mathematical properties of vector spaces (for example, orthogonally) to develop combinations of factors which can (a) produce a great deal of information in a small number of experiments, particularly compared to one-factor-at-a-time (sometimes called OFAT) experimentation, and (b) frequently provide information which could not be developed if one-factor-at-a-time experimentation is used, i.e., estimates of interactions. Standard works describing Design of Experiments include Montgomery (1997), Box, Hunter, and

Hunter (1978), Khuri and Cornell (1987), and Box and Draper (1987)[15].

Plans for DoE can be represented graphically with a dot representing one or more runs of an experimental combination. For example, an experiment with three factors, x_1, x_2, and x_3, each with two possible levels (high and low) can be represented as in Figure 5.

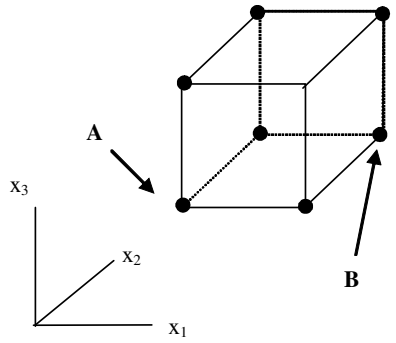


Figure 5: A three factor two level full factorial DoE. "Part A" is at the 'low' values of x_1, x_2, and x_3. "Part B" is at the 'low' value of x_3 and the 'high' values of x_1, and x_2.

As there is one experimental run for each of the possible combinations of factors, this is called "an eight run 2 level full factorial design," with $2 \times 2 \times 2 = 2^3 = 8$ experimental runs. If the phenomenon of interest is the physical performance of parts (called y[16]), eight parts would be made, each with a different combination of characteristics being studied. The experiment would

---

[15] These authors recommend an iterative approach, frequently starting with a screening experiment (simple fractional factorial with center points) to obtain a basic idea of the Response Surface, moving toward a region of interest, and exploring the that of interest with a Response Surface Design, such as a Central Composite Design.

[16] It is possible for y to be a vector. If so, DoE can be used to make trade offs between outputs.

be run, the corresponding levels of y recorded, and a graphical and statistical analysis (for example, ANOVA) performed. In this 'full factorial' case, equations for y can be developed with linear, quadratic, and interaction terms estimated free and clear of one another.

It is possible, using "generators," to develop "fractional factorial designs" (Box, Hunter, and Hunter, 1978). For example, using a "half fraction" one might study four variables in 8 experimental runs. Thus, rather than with $2x\ 2\ x\ 2\ x\ 2 = 2^4 = 16$ experimental runs, one has $2\ x\ 2\ x\ 2 = 2^{4-1} = 8$ experimental runs. For this efficiency one gives up the ability to estimate "higher order interactions" without "confounding." Fortunately for experimenters, such "higher level terms" are frequently not statistically significant (Box, Hunter, and Hunter, 1978).

Another design might be composed of "center" and "axial" points (experimental runs), as shown in Figure 6. The axial points are sometimes called "star points."

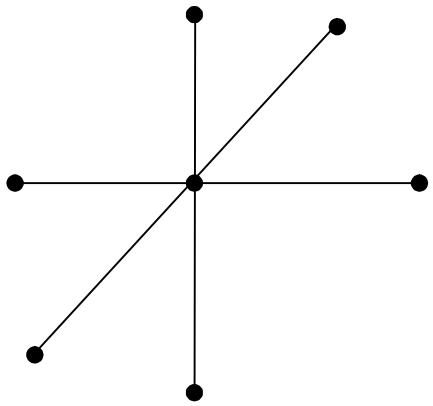


Figure 6. Center and axial (star) design DoE.

A factorial design (or fractional factorial design) and the design with center and axial points can be combined to develop what is termed a Central Composite Design, one of a class of designs termed “Response Surface Design.” This is illustrated in Figure 7. Frequently multiple runs are made at the “center point” to provide an error term to be used in ANOVA (Box and Draper, 1987). These designs have good statistical properties such as rotatability (Box and Draper, 1987).

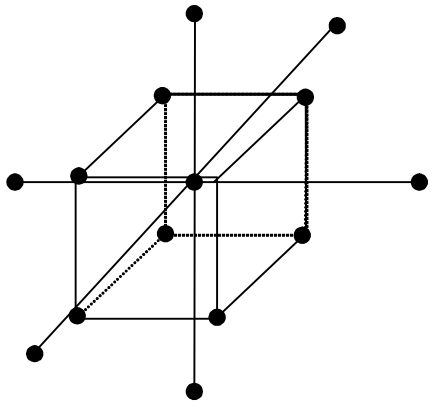


Figure 7: Central composite design (CCD) DoE.

The following graphic, Figure 8 is a slightly different representation of the Central Composite Design.

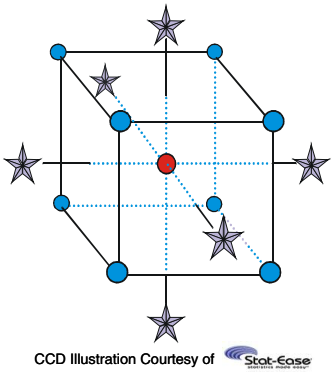


Figure 8: CCD alternate representation. Graphic by permission of Stat-Ease.

An advantage of the Central Composite Design is that it can be used to model an empirical response surface with squared and interaction terms, for example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \quad (21)$$

It is useful to visualize two of the three (predictor variable) dimensions this Response Surface as illustrated in Figure 9.

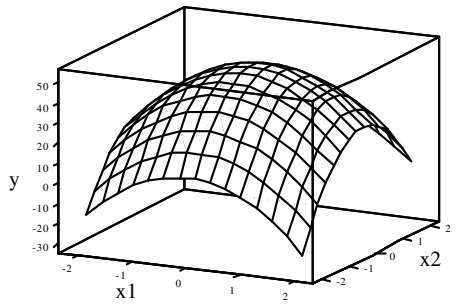


Figure 9: Response surface of y in x_1 and x_2. Compare with Figure 10.

It is more useful, however, to graph the Response Surface of y as a two dimensional contour plot as in Figure 10 which can be read much like a temperature map produced in a daily newspaper.

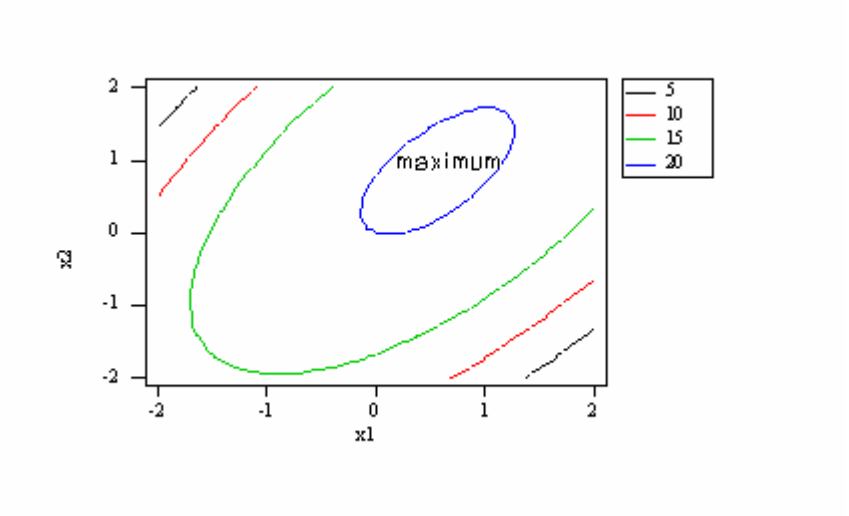


Figure 10: Contour plot of y in x_1 and x_2. Compare with Figure 9.

In this paper Design Expert software (Heleseth, et. al., 2000) is used to 'build' the Central Composite Design and generate the response surfaces which are the focus of the study.

Logistic Regression

Evert (2002) defines Logistic regression as follows:

> A form of regression used when the response variable is a binary variable. The method is based on the *logistic transformation* or *logit* of a proportion[17], namely
>
> $$\text{Logit}(p) = \ln(p/1-p)$$
>
> As p tends to 0, logit (p) tends to $-\infty$ and as p tends to 1, logit (p) tends to ∞ . The function logit (p) is a *sigmoid curve*[18] . . . . Applying this transformation, this form of regression is written as:
>
> $$\ln(p/1-p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$$
>
> where p = Pr (dependent variable = 1) and $x_1, x_2, \ldots x_q$ are the explanatory variables. Using the logistic transformation in this way overcomes the problems that might arise if p was modeled directly as a linear function of the explanatory variables, in particular, it avoids fitted probabilities outside the range (0,1). (pp. 225-226)

In this paper, a single explanatory variable is used, thus the "data probability" used in Bayes's Theorem can be calculated using the following formula and the estimates for α (a) and β (b) to obtain probability of a student who has a certain observed score (x) is indeed proficient.

$$P(x) = \frac{1}{1+e^{-(\alpha + \beta X)}} \qquad (22)$$

---

[17] Present writer's note: The proportion in question is the odds ratio, or the odds in favor, or the probability of success divided by the probability of failure.

[18] Present authors note: a sigmoid curve has an elongated s shape.

This Study

Thought Experiment

In this section a simple thought experiment is described which will provide a basic overview of the procedure used:

- Governmental body sets educational standard for high school graduation.
- A standardized test is developed and cut score for proficiency (meeting the standard) is determined.
- 1,038,044 students who are covered by the standard take the test.
- Before the test is administered a prior probability of meeting the standard is assigned to each student.
- After the test is administered and scored, a Student is classified as proficient using the Observed Score procedure if their Observed Score is greater than or equal to the Cut Score.
- After the test is administered and scored, information from the student's Observed Score is combined with that student's prior probability via Bayes's Theorem to produce a posterior probability. A Student is classified as proficient using the Bayesian procedure if the posterior is greater than or equal to 0.5.
- The state hires an "Omniscient Classifier," who can determine the exact True Scores of each of the students without error (the obvious

impossibility of this is why this is a thought experiment[19]).

- The "Omniscient Classifier" Calculates 3 numbers:
  1. The proportion of students who were classified correctly (with their individual True Score as the criterion) by the Bayesian Procedure but incorrectly classified by the Observed Procedure.
  2. The proportion of students who were classified correctly (again, with their True Score as the criterion) by the Observed procedure but incorrectly classified by the Bayesian Procedure.
  3. The Net Bayesian Advantage (NBA), which is #1 minus #2.
     - NBA can be positive (Bayesian is better) or Negative (Observed is better)
     - Cases where both procedures classified correctly or both procedures classified the student incorrectly can be ignored.

With NBA, the government can, after appropriate statistical analysis, decide whether there is evidence that the Bayesian or Observed procedure produces better results. Additional studies could be conducted with other students to determine if the results of the study on the superiority or inferiority of a Bayesian approach are generalizabile.

---

[19] Sometimes some people act as though the test were an "Omniscient

Outline of the Study (Simplified Process)

Having provided a "thought experiment" of an ideal way to judge a Bayesian approach versus the use of Observed Scores alone, this section provides a simplified version of the present study. The goal of the thought experiment was to calculate a simple metric to judge a Bayesian approach—NBA. In the "thought experiment" the calculation of NBA required collection of three pieces of information corresponding[20] to each of 1,038,044 students:

- Whether the student's True Score was above or below the Cut Score
- Whether the student's Observed Score was above or below the Cut Score
- Whether the student's posterior probability was above or below 0.5.

In other words, each student was a datapoint for which there were three pieces of information.

Without an "Omniscient Classifier" like that in the thought experiment, it is impossible to *directly* evaluate whether a *specific* Bayesian procedure will classify students more accurately than relying on Observed Scores alone (that is, have a positive NBA). However, because simulation permits the researcher to posit the existence of a set of True Scores, it is possible to evaluate the *circumstances* under which a Bayesian procedure would be superior. This section outlines that simulation by reference to Figure 11. In this there are two important items for the reader to keep in mind:

Classifier"

- The steps in Figure 11 between the rounded boxes at the top and bottom represent calculations performed on ONE True Score. They are repeated 1,038,044 times in the simulation for *each* of the 30 combinations of Bias, SDBias, Cut Score, and Reliability in the Central Composite Design (a total of 31,141,320 simulated testers).
- The details of this outline are covered in the next section, *The Study.* In that section the reader will find *much* fuller descriptions of L, description of the logistic regression, SEM (standard error of Measurement), "True Probability," the method the simulate deviations, source of the distribution of Estimated True Scores, etc.

As frequently is the case with processes with multiple parallel steps, it is easiest to start with the output. Thus we will start at the bottom (above the rounded box) and work up to the other rounded box. The final last output is found in the unshaded hexagon near the bottom right hand side of the flowchart. Note there are two other unshaded hexagons. These correspond to the classifications in the thought experiment:

- Posterior >= 0.5,
- Observed Score >= Cut Score, and
- True Score >= Cut Score.

The posterior is calculated by Bayes's Theorem in the pink box. This is the Likelihood form of Bayes's Theorem, Equation 9 [with L = P (e | not h) / P (e | h)].

In addition to L, the posterior has one other input, the prior.

The steps to calculate the prior are shaded yellow. It is the sum of the "True Probability" plus a deviation from that probability. The deviation is, in turn, the result of a Monte Carlo simulation. This simulation produces deviations which are normally distributed with a mean equal to the "Bias" and a standard deviation equal to the "SDBias." Bias and SDBias are selected according to a Central Composite Design (see Figures 20 and 21). The "True Probability" is a simple calculation of the probability of tester with a given True Score obtaining an Observed Score equal to or above the Cut Score given an SEM, which is turn is a function of the Reliability of the Test (see all green boxes, white SEM box, and blue "One Reliability" box).

Returning to Bayes's Theorem (Equation 9) in the pink box, one other piece of data is needed to calculate the posterior, L. L is the ratio of a) the probability that a student is NOT being proficient given their Observed Score on a test to ii) the probability that they are proficient given their Observed Score.[21] This is a result of applying using one of several logistic equations (the factors in the run of the CCD determine which equation) and the Observed Score (see blue box with rounded corners on the right). The Observed Score is also simulated. This is done by a deviation to the Estimated True Score that is under examination (central green box). The deviation is based in turn on a simulation with mean zero and the SEM (Standard Error of Measurement).

---

[21] As discussed in other sections of this dissertation, L relates to Popper's bold conjectures.

There are 1 million-plus iterations (trips through this flowchart) for each run of the Central Composite Design. Once the iterations of a run are completed, the classifications (True, Observed, and Posterior indicating proficient or not proficient) are summarized and the Net Bayesian Advantage is calculated (see large lower curved box). This is repeated for each of the 30 runs of the Central Composite Design, producing 30 observations of Net Bayesian Advantage which are modeled using Response Surface Methodology.

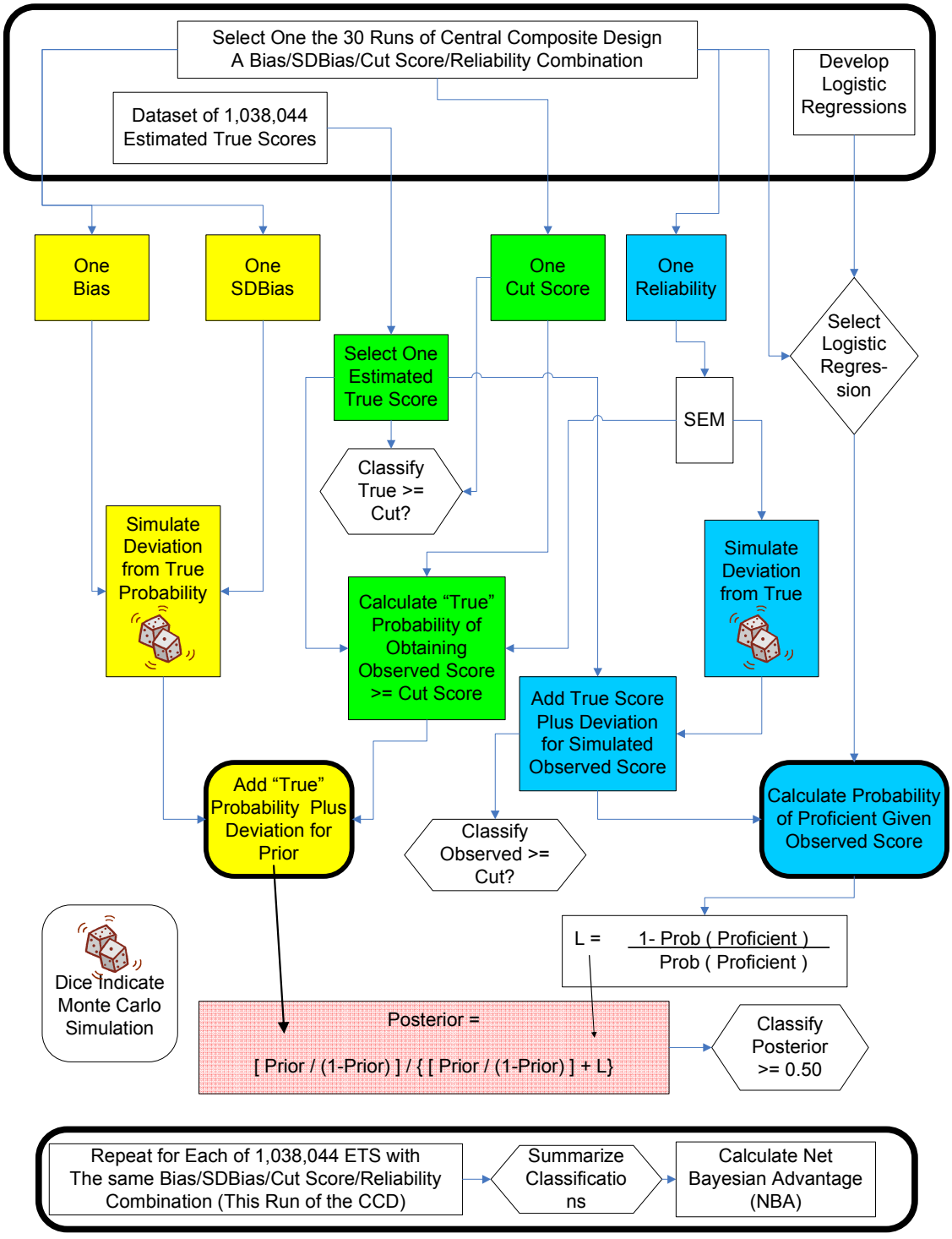


Figure 11: Flowchart of a simplified version of steps in this study, with emphasis on the steps in the simulation. A more detailed flowchart is in Appendix D.

The Study

Having provided a "bottom up" overview of the study in the prior section, this section provides the "top down" detail. It is flowcharted in Appendix D.

Preliminary Step

This study makes extensive use of simulations based on Estimated True Scores (ETS). This section describes the development of those Estimated True Scores from the results of the High School Mathematics portion of the Michigan Educational Assessment Program for Grade 11 First Time Testers Spring 1999 (hereafter referred to as 99HS-MEAP-M). This dataset contained 74,146 scores (Appendix A). To obtain ETS, the following procedure was used:

Step a. Subtract the mean of the scores (28.887) from each score to obtain the deviation score, x.

Step b. Obtain an estimate of the true deviation score by multiplying the deviation score by the reliability using equation 18:

$$t' = r_{xx}\, x \qquad (18)$$

where r_{xx} is the reliability of the test, which was 89.2 for the 99HS-MEAP-M.

Step c. Add back the mean (28.887) to obtain the 74,146 ETS.

Step d. "Stack" the 74,146 ETS fourteen times to obtain 1,038,044 ETS which will be used throughout the rest of the study. (In the simulation, random disturbances are added to the 1,038,044 ETS).

Likelihood Form of Bayes's Theorem

In this study, the simple, likelihood form of Bayes's Theorem, equation 9, is use. It is presented in equation 23 in an alternative notation:

$$[\text{Prior} / (1-\text{Prior})] / \{ [\text{Prior} / (1-\text{Prior})] + L\} \quad (23)$$

This equation has two inputs, the Prior and L. The Prior is the Bayesian Prior probability that the student meets the standard. It is the object of this paper to develop specification for two characteristics of this Prior, Bias and Consistency, which can be used by researchers as guidance in developing and evaluating specific procedures for generating these priors if, as suggested in Chapter 5, a decision is made to evaluate the costs and benefits of using a simple Bayesian approach to classify students. In this case, L is the information contained in the Observed score. It is defined as:

$$L = P (T = \text{not } M \mid O) / P (T = M \mid O) \quad (24)$$

Where T = True Score
M = Meets the Standard
O = Observed Score.

In other words, L is an estimate of the ratio of the conditional probability (conditioned on the Observed Score) that the test taker does <u>not</u> meet the standard divided by the conditional probability (again, conditioned on the Observed Score) that the test take <u>does</u> meet the standard. As this is a binary case where either the student meets the standard or does not (that is, the True Score is either is at or above the Cut Score or it is not), L is to be estimated by calculating P (T = M | O). A reasonable method of doing this is as is described in section below. A reasonable method for generating priors for use in this study

(which focuses on their Bias and Consistency) is described in the second following section.

It is worth stressing at this point in the paper that although the process for generating the probabilities that are entered into Bayes's Theorem (equation 9) are *considered* reasonable, the reasonableness of these inputs is not *required*. What is important for a study which looks at characteristics of a Bayesian Prior is that the process for producing the Priors and the Data Probabilities is well defined (as it is in this study by a computer program, Appendix B) and that there is a classification of the students as meeting or not meeting standards that results from the probabilities (see Figure 15). It is these classification that determine the calculation of Net Bayesian Advantage (Equation 30), which is the measure of the improvement in classification as proficient or not proficient (with True Score as the criterion) from a Bayesian approach compared with relying on Observed Scores alone.

Figures 12 through 14 provide a graphical comparison of Observed Scores from the 99HS-MEAP-M with ETS.

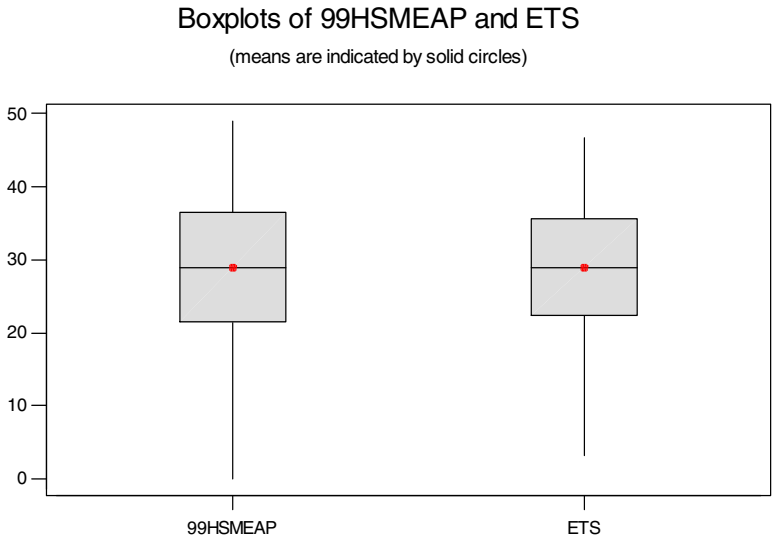


Figure 12: Boxplot of 99HS-MEAP-M and ETS (Estimated True Scores) used in the simulation. Note the attenuation of ETS relative to 99HS-MEAP-M.

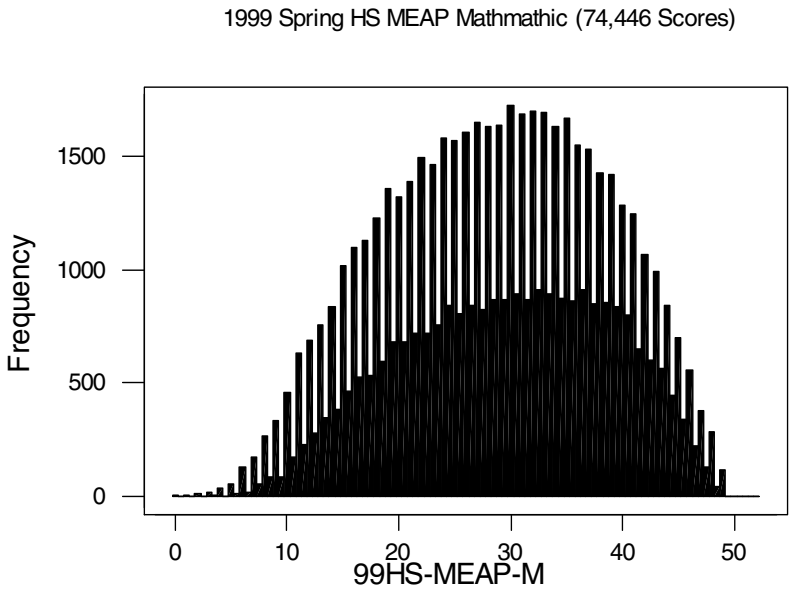


Figure 13: Histogram of 99HS-MEAP-M showing digit preference.

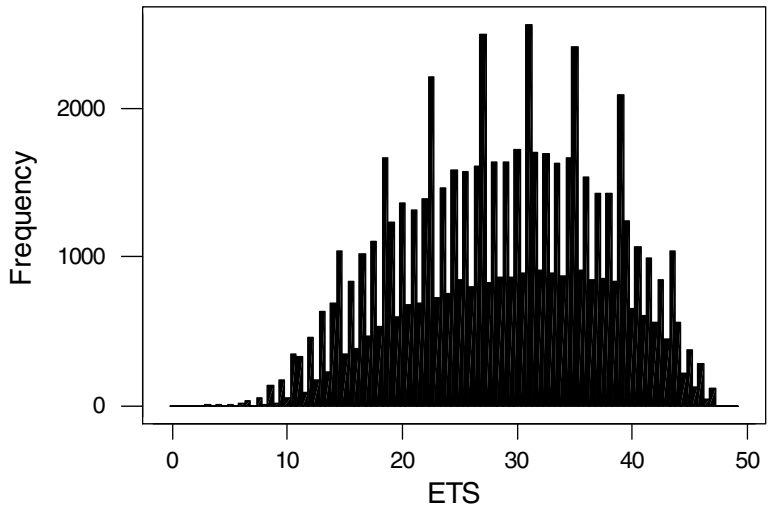


Figure 14: Histogram of ETS showing digit preference.

Generating L

Step a. Take 1,038,044 ETS the development of which is described in above.

Step b. Generate an 'error' term for each ETS as follows:

Step b.i. Determine the Reliability of the test of interest.

Step b.ii. For that reliability, calculate a Standard Error of Measurement (SEM) using equation 20 (repeated below for convenience):

$$\sigma_{meas} = \sigma_x \sqrt{(1-r_{xx})} \qquad (20)$$

where σ_{meas} = SEM
σ_x = standard deviation of the original data (74,146 actual observed scores in the HS-MEAP-M, and
r_{xx} = the Reliability of the test of interest.

Step b. iii. Generate 1,038,044 random 'errors' from a normal distribution with mean 0 and standard deviation equal to the Standard Error of Measurement (SEM) of interest.

Step b. iv. Add the random 'errors' to the ETS.

Step b. v. Truncate the new simulated observed scores at 0 and 49 (the range of scores on the 99HS-MEAP-M).

Step b. vi. A. 1. Code the ETS as follows:

1 if ETS >= Cut Score of Interest
0 if ETS < Cut Score.

Step b. vii: Using the Minitab Logistic Regression subroutine (Appendix C), develop equations for the Logit with the Coded ETS as the dependent variable and the simulated Observed Score (O) as the independent variable. The coefficients for these equations will be used in the Main Simulations to calculate P (T = M | O), and thus L = P (T = not M | O) / P (T = M | O) for each O (Observed Score in the Main Simulations) using equation 22 (repeated below for convenience):

$$P(x) = \frac{1}{1+e^{-(\alpha + \beta X)}} \qquad (22)$$

Generating Priors

The process used for generating prior is a follows:

Step a. Take 1,038,044 ETS the development of which is described above.

Step b. For each ETS, calculate the probability of obtaining an Observed

Score greater than or equal to the Cut Score given the Observed Scores and assuming the Observed Scores are normally distributed with mean equal to the ETS and standard deviation equal to the SEM (the rational for step is discussed in the next section):

$$P (O >= C \mid \mu = ETS , \sigma = SEM) \quad (25)$$

This is mathematically equivalent to calculating the following:

$$P (O < ETS \mid \mu = Cut , \sigma = SEM) \quad (26)$$

That is, for each EST, calculate the probability cumulative probability to the ETS of a normal distribution with mean equal to the Cut Score and standard deviation equal to the SEM. This identity is used because equation 26 can be easily calculated for the vector of 1,038,044 observations of ETS within Minitab.

Step c. Generate, 'error' terms from a normal distribution with mean equal to the Bias and standard deviation equal to the SDBias (a measure of (lack of) Consistency).

Step d. Add the results of Steps b and c.

Step e. Windsorize the result of Step d at 0.5 and 99.5. This insures that all students will have be estimated to have some (be it only 0.5%) chance of meeting the standard and not student will be estimated as having 100% chance of meeting the standard.

A Note on the Reasonableness of Priors

The focus of this study is to develop specifications for characteristics of a Bayesian estimate of the probability that a student meet the standards of a high stakes test that, when combined with information from the Observed Scores using Bayes's Theorem (equation 9) will improve classification compared with the present procedure using Observed Scores alone. Two reasonable choices of characteristics of these Priors are the following:

- Accuracy, which is measured by its absence, Bias (average distance from "true" probability), and
- Consistency, which is also measured by the its absence, the standard deviation of the Bias, called SDBias in this study.

To discuss the meaning of these characteristics, it is helpful to first consider an "ideal" procedure, which would produce Unbiased and Consistent Priors, that is produce, on average, the correct Prior estimate for all potential test takers who have the same set of characteristics (thus Accurate or Unbiased) and have a small spread around this average (thus Consistent). For example, estimates of Priors based on a procedure involving a transformation of an expected observed score which, in turn, could be predicted by a regression equation would produce the same predicted Prior for each and every student who had the same value of the dependent variables in the equation. Moreover, if the confidence interval of the regression is small, the Prior would be Consistent (in the meaning used in this paper).

A measure of the idea of Consistency is fairly straightforward. The absence of Consistency can be measured by a standard deviation. For example, if an estimation method has a low standard deviation of say 2.167 points, students with the same profiles will all have Priors within +/- 6.5 (3 σ) points 99.73% of the time and +/- 4.334 (2 σ) points 95% of the time.

Determining the Bias is somewhat more difficult because the True Scores (and thus the True Probabilities) are unobserved and unobservable objects. For purposes of this study, Bias is defined as the deviation from an 'Estimated True Probability' that the student meets the standard. While this is unknown and unknowable, one can use a Frequentist definitional thought experiment to describe the best estimate. One can then go on to develop an alternative method to estimate that probability (a task left to other researchers). We are interested in the "True Probability" of obtaining an Observed Score about the Cut Score (on a given administration of a test):

- A Frequentist definition of of the 'true' probability that a student meets the standard would be the proportion of times that the student passes the test in a large number of administrations o the test[22]:

$$TP = \lim_{n \to \infty} \Sigma Q_i / n \qquad (27)$$

[22] This could be exactly the same test or different forms of a given test.

where

TP = A Student's True Probability of Meeting Standards (Frequentist definition)

O_i is the Observed Score of student takes the test for the i^{th} time

Qi = 1 if $O_i >= C$

0 if $O_i >= C$

ΣQ_i = the number of times the Observed Score is greater than the Cut Score, and

n = Number of Times the Test Is Administered to the student.

- Unfortunately, this calculation is impossible because it is impossible to have even a small number of test administrations (n) without the student learning more or forgetting something that the Test is measuring[23]. However, the practical difficulties in estimating this TP is a useful unobserved object. Moreover, it is as reasonable to use it as it is to use the concept of a True Score.
- It has been assumed that a True Score exists, or at least it is reasonable to use the concept of a True Score. If a TP exists and the True Score exists, there must be in a one-to-one correspondence between the two.
- If one knows the distribution which Observed Scores follow given a True Score, one can calculate the following:

[23] For example, allowing 3 weeks between tests to prevent an impact of a quick retest on performance, it would take over half a year to administer ten tests. One would hope a student

$$P(O >= C \mid \text{True Score, other parameters}). \quad (28)$$

where O = Observed Score and
C = Cut Score

- This must be equal to the True Probability of equation 27.
- Classical Test Theory postulates that a confidence interval can be built around the true score (Nunnally and Berstein, p. 240, and 260) and that this confidence interval involves the SEM.
- If it is further assumed that the distribution is normal, we have the result:

$$TP = P(O >= C \mid \mu = TS, \sigma = SEM) \quad (29)$$

where $O \sim N(TS, SEM)$.

It is not unreasonable, as a first approximation, to think that the Observed Score has a normal distribution because many phenomena which are impacted by a multitude of factors are approximately normally distributed (further research might look at alternative distributions) and a student's score on an individual test is certainly influenced by a number of factors, including the following:

- The room in which the test is taken (from environment (light, temperature, air flow, etc) to information posted on the walls),
- The student's temporary state of health,
- Impact of 'cramming,'
- Personal factors (for example, a recent fight with a friend or parent), and
- Whether or not breakfast was eaten the day of the test.

---

would learn during a half-year in school. In addition, a student could be expected to forget some

Measures of Effect on Classification Success of Bayesian Estimate

In order to develop specifications for a Bayesian approach, it is necessary to a) decide on a distribution of data to use as the basis for simulations, and b) develop a metric of the degree to which a Bayesian classification is better or worse than Observed Scores alone.

In determining the distribution of data to use as a specification, it was decided to use an actual distribution of high stakes test scores rather than a theoretical distribution (for example, the normal). There are three reasons for this. First, working with a real distribution provides insights that might not be available if one used a theoretical distribution[24]. Second, use of a actual distribution of data provides a more realistic estimate of order of magnitude of potential improvement possible if one applies a Bayesian approach in a single 'real world' case. Finally, the first two reasons combined will not only help motivate other researchers to develop such approaches, but also positively dispose policy makers to use them.

In addressing the other question, the development of a metric of the degree to which a Bayesian classification is better or worse than Observed Scores alone, it is helpful to look at Figure 15, which lists the combinations of True, Observed, and Bayesian classifications which are possible for a single student. Given that there are two possibilities (either Meeting or Failing to Meet the Standard) for each of the three classifications (True, Observed, and

---

of the test content if not in school for 30 weeks.

[24] The distribution chosen exhibits 'digit preference.'

Bayesian), the following table lists the eight (2^3) mutually exclusive and exhaustive possible classifications. The computer simulation provides data with which to calculate the proportion of these classifications and they are used to calculate the metrics.

No.	Code	TRUE SCORE	Observed Score Method	Bayesian Method	Compared With Observed Score Method, Bayesian Classification Is
1	tfofbf	fail	fail	fail	Same
2	**TfofBM**	**fail**	**fail**	**Meet**	**Worse**
3	**tfOMbf**	**fail**	**Meet**	**fail**	**Better**
4	tfOMBM	fail	Meet	Meet	Same
5	TMofbf	Meet	fail	fail	Same
6	**TMofBM**	**Meet**	**fail**	**Meet**	**Better**
7	**TMPMbf**	**Meet**	**Meet**	**fail**	**Worse**
8	TMOMBM	Meet	Meet	Meet	Same

Figure 15: Possible combinations of classifications used in the main simulation (Appendix B)

A reasonable measure of the degree to which the Bayesian approach results in an improved classification will be called the Net Bayesian Advantage (NBA). This is defined as a) the proportion of test takers classified correctly by Bayesian Method (groups 3 and 6 in Figure 15) but incorrectly by the Observed method minus b) the proportion of test takers classified incorrectly by the Bayesian method but correctly by the Observed method (groups 2 and 7 in Figure 15). Using the codes in the above table, this can be summarized with the

following equation (where each code refers to the proportion of results with that code):

$$NBA = (tfOMbf + TMofBM) - (tfofBM + TMOMbf) \quad (30)$$

Note: NBA can be positive or negative. Positive NBA indicates regions where a Bayesian classification procedure with a given combination of the four factors, Bias, Consistency, Reliability, and Cut Score is better than a procedure using Observed Scores alone. Negative NBA indicates regions where a Bayesian classification procedure with a given combination of the four factors, Bias, Consistency, Reliability, and Cut Score is worse than a procedure using Observed Scores alone.

While NBA is the main measure of a Bayesian method, a secondary measure might be helpful in cases like the following. Two methods might have the same NBA, but one of the "NBA's" results could include in a larger number of students who *would* have been correctly classified with the Observed method being 'reclassified' incorrectly with the Bayesian method. In general, one would prefer the method which does results in fewer new misclassifications of students who would have been correctly classified under the Observed method. Thus the following measure, Bayesian Worse than Observed (BWO) will also be used:

$$BWO = tfofBM + TMOMbf \quad (31)$$

Note: BWO is the second parenthetical term of NBA.

Ranges of Interest and the Central Composite Response Surface Design

Next, the range of values for variables under study is discussed. Two groups of variables are studied. The first group relates to characteristics of the Bayesian estimate (Bias and Consistency). The second group relates to characteristics of the tests for which a Bayesian approach might be used (Reliability and Cut Score). It is important to study the responses over ranges of this second group of factors because one wants to insure that the procedures can be use over a range of tests characteristics that are outside of the control of those who are applying the procedure. The ranges selected for study are summarized in the Figure 16 and discussed below:

	Measure of	**Variable Name**	**RANGE OF INTEREST**	
			Low Level	**High Level**
Aspects Bayesian Estimated. Prior Probability that the Student Meets the Standard	Bias	Bias	-15	15
	Consistency: Range (3 Standard Deviations)	N/A (see below)	6.50	18.00
	1/3 of Range (One Standard Deviation)	SDBias	2.1667	6.0000
Aspects of the Tests	Reliability	Reliability	84.0	94.4
	Cut Score	Cut Score	23	29

Figure 16: Range of variables of interest

The primary[25] rational for the ranges of Bias and SDBias that were selected is to provide a wide coverage of deviations from a "True Probability" so that the potential for improvement in classification for a potential Bayesian approach could be studied under this wide range. This was done by looking at the ranges that would be produced by combinations of Bias and SDBias. It is reasonable to expect that any test will be most likely to misclassify a student whose "True Probability" of obtaining a Observed Score about the Cut Score is approximately 0.50, which is equivalent to saying that the True Score is approximately equal to the Cut Score.

Figure 17 and 18 provide examples of calculation using addition and 'build up' reasonable levels of the priors. The reader will notice that the Bias and SDBias when added to the 50% can result in a wide range of prior probability of meeting the standard. Calculations show resulting levels of prior for different levels of Bias and Consistency (SDBias).

[25] Other considerations included producing possible combination in a Central Composite Design.

	High Bias/ Low Consistency		Low Bias/ High Consistency	
	Low	High	Low	High
"True Probability"	50.0	50.0	50.0	50.0
Bias	-15.0	15.0	0.0	0.0
+/- 2 SDBias Impact	-18.0	18.0	-6.5	6.5
Range	17.0	83.0	43.5	56.5

Figure 17: Build up of range of interest of Prior Probability, +/- 3 SDBias example.

	High Bias/ Low Consistency		Low Bias/ High Consistency	
	Low	High	Low	High
"True Probability"	50.0	50.0	50.0	50.0
Bias	-15.0	15.0	0.0	0.0
+/- 2 SDBias Impact	-12.0	12.0	-4.3	4.3
Range	23.0	77.0	45.7	54.3

Figure 18: Build up of range of interest of Prior Probability +/- 2 SDBias example.

Specifically, for a True Score equal to the Cut Score, this study explores ranges of Bias and SDBias that result in Bayesian Priors that have a 95% confidence interval between 0.23 and 0.77 and a 99.73 % confidence interval between 17 and 83. These High Bias/Low Consistency ranges, representing the “worst case combination” examined in this study, are detailed in the two

following tables. The “best case” ranges for the Low Bias/High Consistency combinations are also provided. However, as the risk of any new Bayesian Method would be that the ranges studied were not wide enough, it is the High Bias/Low Consistency which are of most interest.

Next, the range of interest of characteristics of the test, Reliability and Cut Score, will be discussed (See Figure 16). The range of Reliability that was selected for study is 84.0 to 94.4. It is believed that this range of Reliability covers a large number of modern high stakes tests. As the Reliability approaches 100, the potential usefulness of any Bayesian method will diminish. Indeed, a perfectly Reliable test would correctly classify almost[26] every student 100% of the time. Few tests are anticipated to have Reliability above 94, and thus 94.4 is a reasonable higher bound. The reason the high level is not “round” number of 94.0 is due to the formula to selection of axial points in the Central Composite Design. This will be discussed in the next section.

The Reliability and Cut Score centerpoints (defined below) of 89.2 and 26.0 respectively were deliberately selected to be the same as the actual Reliability and Cut Score of the 99HS-MEAP-M. The lower bound of the Cut Score range, 23.0, was selected to be half-way between the actual Cut Score and the level at which the student is considered to meet the Michigan Standards for the 99HS-MEAP-M at a “basic” level. This lower level is three away from the centerpoint (26.0 – 23.0 = 3.0). The high bound of the range of interest, 29.0,

was obtained by adding 3.0 to 26.0 (providing a symmetric range of 3.0 around the centerpoint).

Central Composite Design (CCD)

A Central Composite Response Surface Design allows the estimation of a quadratic equation that has good statistical properties in the range of interest. To provide data to estimate this equation, in addition to points at the Low and High end of the range of interest (which can be coded as "-1" and "+1"), a Central Composite Design also has experimental runs at the center points (halfway between the –1 and + 1 levels) which are coded "0" and at eight axial points. The axial points are placed at a distance from the centerpoints equal to +/- √ (number of factors). The +/- 1 levels are called the factorial points because they constitute a two-level factorial design (in this case a 2^4 factorial). In this study the number of factors is 4, so the axial points are at +/- 2 coded distances from the centerpoints. For example, the Low Level of reliability is 84. This is 5.2 points from the centerpoint of 89.2. Thus, 5.2 if equal to a distance of 1, the axial points are at 2 x 5.2 or +/- 10.4, or 78.8 (89.2 - 10.4) and 94.4 (89.2 + 10.4). These axial points are only used for estimation of the quadratic equations and thus one does not view the Central Composite Design as providing accurate information about the quadratic response surface at the axial points (or beyond). Thus, the fact that in practice one would never encounter

---

[26] The exception is that student whose True Score is <u>exactly</u> (at an infinite number of decimal

some of the levels of the axial points (Reliability of 99.6, SDBias of 0.25 percentage points) is not a concern. Thus the levels of the variables in the boxed area are of particular interest. The Axial points are used for estimation.

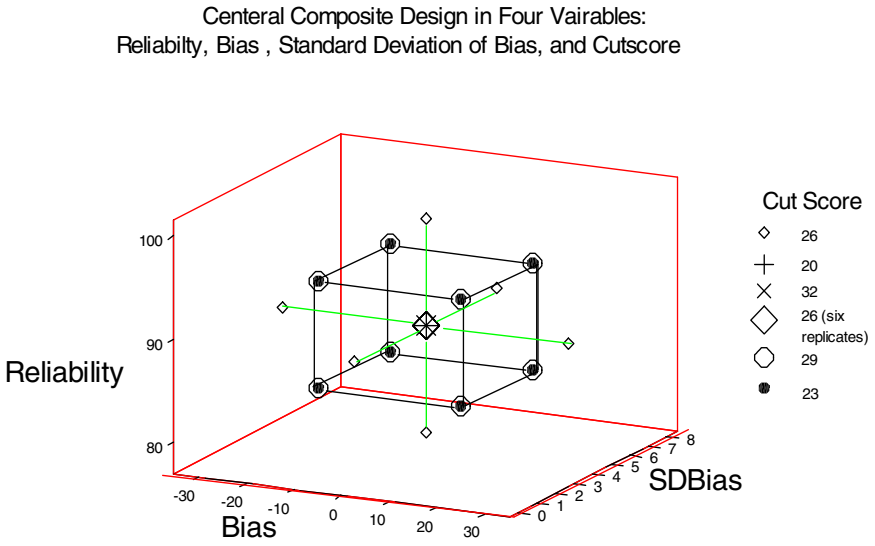
Variable	Axial	Low	Center	High	Axial
RELIABILITY	78.8	**84**	**89.2**	**94.4**	99.6
CUTSCORE	20	**23**	**26**	**29**	32
BIAS	-30	**-15**	**0**	**15**	30
SDBias	0.25	**2.16667**	**4.08333**	**6**	7.91667
Memo: 3 SDBias	0.75	**6.5**	**12.25**	**18**	23.75
Memo: Central Composite Design Code/Distance From Centerpoint	-2	-1	0	1	2

Figure 19: Points of the Central Composite Design (CCD) used in the study. Points of interest in box. Axial points are used for estimation of quadratic response. Compare the 3 SDBias memo with Figure 17.

The following geometric representation of the 30 run design may help to visualize the design. It includes 16 factorial points (at the corners of the square, equal to the High [+1] and Low [-1] levels), six center points (coded 0, the center points of the design is repeated six times to get a measure of pure error), and eight axial points (two of which are projected in 3-space at the same point as the center points on this figure).

---

places) equal to the Cut Score. That student would be correctly classified 50% of the time.



Note: The six small diamond axial points are equidistant from the centerpoint

Figure 20. Graphical representation of 4 dimensions of Central Composite Design used in this study.

A reason the Central Composite Response Surface Design is so useful can be seen in the following correlation matrix (Table 1). Specifically, the four independent variables are completely uncorrelated. Thus, the resulting equations for any response(s) have excellent statistical properties.

Table 1: Correlation of Reliability, Cut Score, Bias, SDBias

	Reliabily	Cut Score	Bias
Cut Score	0.000		
Bias	0.000	0.000	
SDBias	0.000	0.000	-0.000

Cell Contents: Pearson correlation

Specifically, the interaction of these terms have zero correlation, as do their squares. The fact that the given the 30 points selected, the four factors are orthogonal and thus any difference estimated by the model is not due to the selection of points and is “free and clear” of influence of other factors.

Figure 21 contains a list of the levels of the 30 runs of the simulation. The “Standard Order” helps one understand the geometry of Central Composite Designs. The “Run Order” is a random order in which the simulations will be run in one computer session. It is also contains an indication of whether a particular run is a Center Point (Center), Factorial Point (Fact), or Axial point (Axial). This design was generated with Design Expert ® (Helseth, et. al, 2000).

Standard Order	Run Order	Point Type	Reliability	Cut Score	Bias	SDBias	Memo: 3 x SDBias
1	3	Fact	84.00	23	-15	2.16667	6.50
2	15	Fact	94.40	23	-15	2.16667	6.50
3	10	Fact	84.00	29	-15	2.16667	6.50
4	29	Fact	94.40	29	-15	2.16667	6.50
5	9	Fact	84.00	23	15	2.16667	6.50
6	21	Fact	94.40	23	15	2.16667	6.50
7	30	Fact	84.00	29	15	2.16667	6.50
8	25	Fact	94.40	29	15	2.16667	6.50
9	26	Fact	84.00	23	-15	6.00000	18.00
10	4	Fact	94.40	23	-15	6.00000	18.00
11	8	Fact	84.00	29	-15	6.00000	18.00
12	24	Fact	94.40	29	-15	6.00000	18.00
13	16	Fact	84.00	23	15	6.00000	18.00
14	20	Fact	94.40	23	15	6.00000	18.00
15	19	Fact	84.00	29	15	6.00000	18.00
16	14	Fact	94.40	29	15	6.00000	18.00
17	22	Axial	78.80	26	0	4.083335	12.25
18	1	Axial	99.60	26	0	4.083335	12.25
19	7	Axial	89.20	20	0	4.083335	12.25
20	11	Axial	89.20	32	0	4.083335	12.25
21	13	Axial	89.20	26	-30	4.083335	12.25
22	2	Axial	89.20	26	30	4.083335	12.25
23	23	Axial	89.20	26	0	0.250005	0.75
24	5	Axial	89.20	26	0	7.916665	23.75
25	27	Center	89.20	26	0	4.083335	12.25
26	18	Center	89.20	26	0	4.083335	12.25
27	17	Center	89.20	26	0	4.083335	12.25
28	12	Center	89.20	26	0	4.083335	12.25
29	6	Center	89.20	26	0	4.083335	12.25
30	28	Center	89.20	26	0	4.083335	12.25

Figure 21: 30 runs of the Central Composite Design (CCD) used in this study. Note that the first 16 points (in the Standard Order 1-16) are factorial points, corresponding to points on the corners of the square in Figure 20. Points 17 through 24 are Axial points, the “star” points in Figure 20, and Points 25 through 30 are centerpoints.

As stated at the beginning of this chapter, SEM is a mathematical function of Reliability. The SEMs which are the actual input into the computer programs are in Figure 22. These are calculated from equation 20.

Reliability	Cut Score	SEM
78.8	26	4.455165
84.0	23	3.870400
84.0	29	3.870400
89.2	20	3.179858
89.2	26	3.179858
89.2	32	3.179858
94.4	23	2.289760
94.4	29	2.289760
99.6	26	0.611964

Figure 22: Standard Error of Measurement used in simulation as function of Reliability and Cut Score (factors in the Central Composite Design)

CHAPTER 4

RESULTS AND RESPONSE SURFACE INVESTIGATIONS

Results of The Main Simulation

A simulation was run for each combination of the 30 combinations of Reliability, Cut Score, Bias, and SDBias listed for the Central Composite Design in Figure 21. Minitab® Statistical Software (Minitab, 2000) was used for the simulations (Appendix B). The basis of these simulations was the same 1,038,044 Estimated True Scores (ETS) with each ETS being a row in the Minitab® Worksheet. These were, in turn, 14 replicates based on each of the 74,146 actual scores of the 99HS-MEAP-M (Appendix A). A local Minitab® macro was used to calculate selected constants corresponding for input the 30 runs of the CCD for the simulation (Appendix C). Figure 22 constants definitions of these constants . Figures 23 and 24 contain definitions the names of variables used in the main simulation (Appendix B).

Inputs To Macro Which Changed For Experimental Run			
Constant	**Name Used in Macro (In Appendix B)**	**Represents**	**Factor(s) In CCD Determining Constant**
k1	Reliability	Reliability	Reliability
k2	SEM	Standard Error of Measurement	Reliability
k4	cut	Cut Score	Cut Score
k5	Bias	Bias of the Bayesian Estimate	Bias
k6	SDBias	Standard Deviation of the Bias of the Bayesian Estimate	SDBias
k7	a	Constant from Logistic Regression	Reliability and Cut Score
k8	b	Coefficient from Logistic Regression	Reliability and Cut Score
Note:	k7 is also called Const-LgstRegress		
Note:	K8 is also called Coeff-LgstRegress		

Figure 23: Constants used in the Macro (Appendix C)

Col.	Name	Represents	How Calculated	Comment
c1	ETS	Estimated True Scores	An input. Each simulation begins with these same numbers	
c2	TM-Code	TM if the True >= CutScore, tf if True < CutScore	Minitab "Code" Subcommand	
C3	e-O	Random number to be added to the true to produce OO	Minitab Generate Random Number for a normal random deviate with mean zero and Std Dev = SEM	
c4	OO	Origional Observed	c1 Plus c4	
c5	Obs1	Trims c4 to be between 0.0 and 49.0	Minitab "Code" Command	
c18	DcmlPt	The decimal part of c5	Minitab "Floor" Command	
c19	DcmlPt Rndd	Round c18 to 0, 0.5, or 1.0	Minitab "Code" Command (If less than 0.33333), 0.5 (if between 0.33334 and 0.66667), and 1.0 if >= 0.66667.	
c20	Obs	Non-Decimal part of C5 plus c19	Minitab "Calc" Command: C20 = Floor C5 + C19	
c6	OM-Code	OM if the Obs >= CutScore, of if Obs < CutScore	Minitab "Code" Command	
c7	T-Prob	P (O >= C | μ = ETS , σ = SEM)	Minitab CDF Command for P (O < ETS | μ = Cut , σ = SEM)	Equations 25 & 26
c8	e-bv-t-P	Error term to add to the true probability to get the bayesian estimate	Minitab Generate Random Number Command for a normal random deviate with Mean = Bias and Std Dev = STBias	
c9	ibe	Initial Bayesian Estimate (Untrimmed)	c7 + c8	

Figure 24: Description of Columns c1-c9 in the main simulation (Appendix B)

Col.	Name	Represents	How Calculated	Comment
c10	Prior	Prior Probability Trims c9 to be between 0.005 and 0.995	Minitab "Code" Command	
c12	Data Prob	"Data Probability," The probability of The True Score Beign above the Cut Score given the logistic regression	Minitab "Calc" Command: C12 = 1/[1+E()**(- 1*['a'+'b'*'obs'])]	Equation 22
c13	Posterior	Result of Bayes'sTheorem	Minitab "Calc" Command: 'Posterior' = ['Prior'/(1-'Prior')]/(['Prior'/(1-'Prior')] + [(1-'Dat Prob')/'Dat Prob'])	Equations 9 and 23
c14	BM-Code	BM if the True >= 0.50, bf if Bayesian < 0.50	Minitab "Code" Command	
c15	Result	The classification of True, Observed Score, and Bayesian	Conconnate c2, c6, and 14, with The Result being One of the Following: tfofbf TfofBM tfOMbf tfOMBM TMofbf TMofBM TMPMbf TMOMBM	See Figure 15 Note: Tallies are used to compute NBA and BYO in Equation 29 Equation 30

Figure 25: Description of Columns c10-c15 in the main simulation (Appendix B)

Results Of Logistic Regression

Before discussing the results of the 30 experimental runs, the results of the Logistic Regressions that are then inputs as the constants a (k7) and b (k8) to the Macro which controls the simulations. There are unique regressions for the unique Reliability/Cut Score combinations.

Inputs Into Macro (Appendix C)			Logistic Regression	
Reliability	**Cut Score**	**SEM**	**Constant (a)**	**Coefficient (b)**
78.8	26	4.4552	-9.629	0.37735
84.0	23	3.8704	-9.833	0.43923
84.0	29	3.8704	-12.863	0.44116
89.2	20	3.1799	-10.127	0.53330
89.2	26	3.1799	-14.216	0.55140
89.2	32	3.1799	-17.456	0.54506
94.4	23	2.2898	-17.772	0.77809
94.4	29	2.2898	-22.868	0.78238

Figure 26: Inputs and outputs of logistic regression. Note: Properly speaking, Reliability is not an input into the Macro. It does, however, determine the SEM used in the Macro to generate the Observed Scores.

.

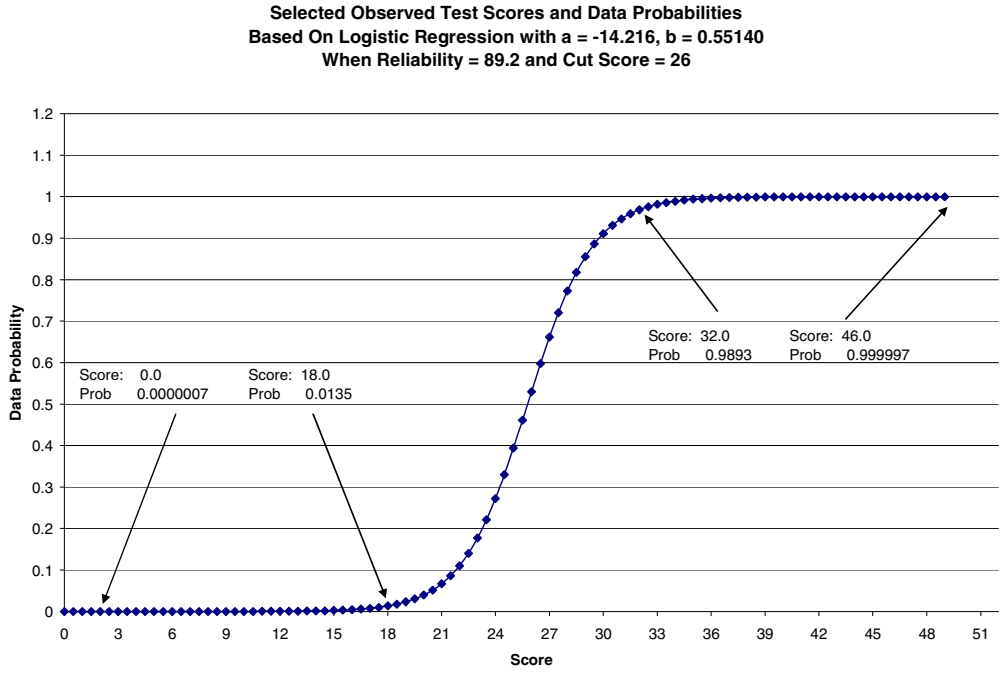


Figure 27: Example of one of the logistic regressions

Results Of Experimental Runs

Results Output

Figure 28, below, contains the results of the 30 experimental runs of the Central Composite Design. In addition to NBA (Equation 30) and BWO (Equation 31), their sum which is termed Gross Bayesian Advantage (GBA) and the ratio of the two NBA to GBA is presented.

$$GBA = NBA + BWO \tag{32}$$

When the ratio of NBA to GBA is close 1, the improvement in classification is without cost of *new* misclassification. While the main analysis of this study is accomplished by an examination of response surfaces generated by regression equations for NBA (Net Bayesian Advantage) and BWO (Bayesian Worse than Observed)—which will not be reviewed until after a discussion of model diagnostics—looking at the results in two other ways (inspection of the table and a simple graphic) will motivate interest in the response surfaces[27].

[27] Had the regression diagnostics not been adequate it would have been inappropriate to conduct this initial review.

Information on CCD			Aspects of the High Stakes Test		Characteristics of the Bayesian Estimate		Responses (All In Percent)			NBA/GBA Ratio
Std Odr	Run Odr	Point Type	Reliability	CutScore	Bias	SDBias	**NBA**	BWO	GBA	
					Bias Vs. Estimated True Prob	Standard Deviation of Bais	Net Bayesian Advantage	Bayesian Worse Than Observed	Gross Bayes Advtg	
1	3	Fact	84.0	23	-15	2.167	4.69	0.22	4.91	0.955
2	15	Fact	94.4	23	-15	2.167	2.75	0.28	3.03	0.908
3	10	Fact	84.0	29	-15	2.167	**5.41**	0.54	5.95	0.909
4	29	Fact	94.4	29	-15	2.167	3.37	0.46	3.83	0.880
5	9	Fact	84.0	23	15	2.167	4.21	0.5	4.71	0.894
6	21	Fact	94.4	23	15	2.167	2.68	0.12	2.8	0.957
7	30	Fact	84.0	29	15	2.167	5.19	0.29	5.48	0.947
8	25	Fact	94.4	29	15	2.167	3.22	0.08	3.3	0.976
9	26	Fact	84.0	23	-15	6.000	4.64	0.25	4.89	0.949
10	4	Fact	94.4	23	-15	6.000	2.79	0.28	3.07	0.909
11	8	Fact	84.0	29	-15	6.000	5.35	0.61	5.96	0.898
12	24	Fact	94.4	29	-15	6.000	3.38	0.46	3.84	0.880
13	16	Fact	84.0	23	15	6.000	4.15	0.55	4.7	0.883
14	20	Fact	94.4	23	15	6.000	**2.62**	0.15	2.77	0.946
15	19	Fact	84.0	29	15	6.000	5.1	0.34	5.44	0.938
16	14	Fact	94.4	29	15	6.000	3.22	0.11	3.33	0.967
17	22	Axial	78.8	26	0	4.083	6	0.04	6.04	0.993
18	1	Axial	99.6	26	0	4.083	1.33	0	1.33	1.000
19	7	Axial	89.2	20	0	4.083	3.38	0.05	3.43	0.985
20	11	Axial	89.2	32	0	4.083	4.69	0.05	4.74	0.989
21	13	Axial	89.2	26	-30	4.083	3.33	1.43	4.76	0.700
22	2	Axial	89.2	26	30	4.083	3.07	1.16	4.23	0.726
23	23	Axial	89.2	26	0	0.250	4.36	0	4.36	1.000
24	5	Axial	89.2	26	0	7.917	4.29	0.08	4.37	0.982
25	27	Center	89.2	26	0	4.083	4.35	0.01	4.36	0.998
26	18	Center	89.2	26	0	4.083	4.34	0.01	4.35	0.998
27	17	Center	89.2	26	0	4.083	4.37	0.01	4.38	0.998
28	12	Center	89.2	26	0	4.083	**4.35**	0.01	4.36	0.998
29	6	Center	89.2	26	0	4.083	4.33	0.01	4.34	0.998
30	28	Center	89.2	26	0	4.083	4.33	0.01	4.34	0.998

Figure 28: Result of Central Composite Design. Runs with NBA responses of particular interest have added emphasis: Standard order #3 (largest response 5.41), #14 (smallest response), and #28 (a centerpoint).

Two things stand out from an examination of this table. First, the Net Bayesian Advantage is not only always non-negative, it is at least 2.6 percentage points. It reaches a maximum of 5.41. Second, the ratio of NBA to GBA (which might be seen as a measure of “pure” advantage to a Bayesian approach) is at least 0.88 at the factorial points, which span the values of interest. Another item worth noting is the closeness of the six experimental runs at the centerpoint. NBA ranges from 4.33 to 4.37, a range of only 0.04 percentage points. This indicates that the simulation is indeed stable.

Before a discussion of modeling diagnostics, one more initial impression of the results of the simulations can be obtained by investigating the cube plot (or cube graph), Figure 29.

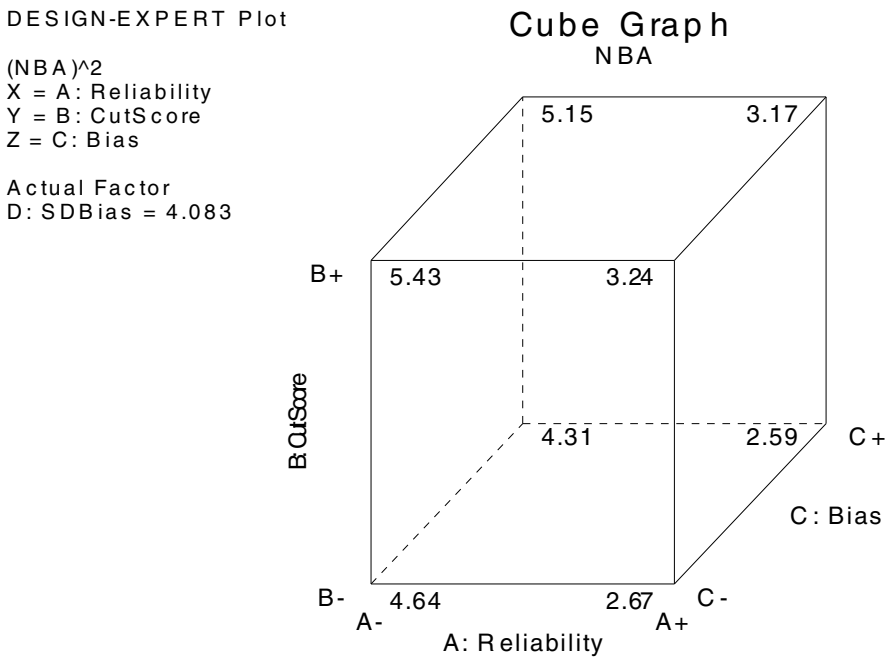


Figure 29: Cube Plot of NBA illustrating the average result of NBA at –1 and + 1 levels or Reliability, Cut Score, and Bias (with SDBias, a forth dimension of the independent variables, held at the centerpoint of 4.083). Not that moving along from the low level of reliability to the high level, whatever the level of Bias or SD Bias, results in a change in NBA of about 2, more than the change along any other axis.

Model of NBA (Net Bayesian Advantage)

In this section the Response Surface Model (RSM) for NBA (Net Bayesian Advantage) is discussed. First, the summary equation, ANOVA, summary statistics and model diagnostics (plots of residuals, leverage, etc.) are presented. Next, the impact of the four factors (Reliability, Cut Score, Bias, and SDBias) on NBA are explored through the Response Surface graphics, primarily contour plots.

NBA Equations, Summary Statistics, ANOVA, and Diagnostics of Functional Form

Before presenting the regression equation, a note should be made about the power transformation used. NBA^2 rather than NBA was modeled. The reason is that the diagnostics are much better for NBA^2 than NBA indicating better meeting of assumptions of a linear model. The decision to square the response variable was guided by a Box-Cox Plot which indicated squaring the response was a transformation which would stabilize the variance, that is, satisfy the constant variance assumption. In the Box-Cox plot below, a transformation of 1 (no transformation) is out of the 95% confidence interval for λ (the Greek leter Lambda), the power transformation which minimizes the natural logorithm of the sum of squares. Squaring is a 'standard transformation' which is esentially the same as the best transformation, 1.98 which is recommended by the Design-Expert® software (Helseth, 2000).

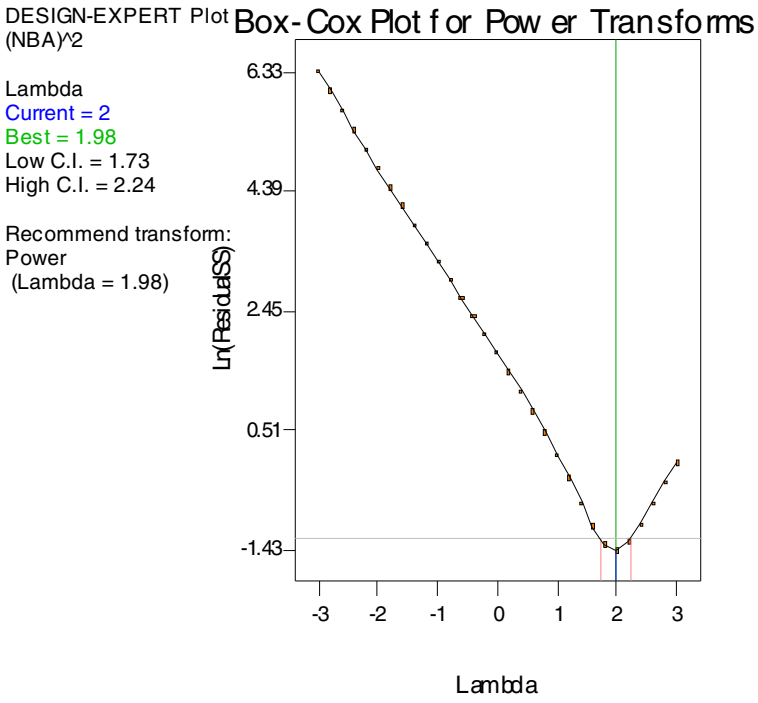


Figure 30: Lambda plot indicating appropriateness of square transformation for NBA.

The response surface equation for NBA^2 (Equation 33) is both parsimonious and useful. It is parsimonious because it does not include all of the possible squared and interaction terms that were estimated in the full Central Composite Design Response Surface Model (most terms which were not significant were dropped in the reduced model presented here). In keeping with standard Response Surface practice, all hierarchies are maintained. Thus, as SDBias is incuded as an interaction term (the final term), the main effect of SDBias is also included. (Note: the final term is included because the focus of the study is Bias and SDBias.) It is useful in that it does include an interaction term for the variables that are the focus of the study, Bias and SDBias, even

though this interaction is not statistically significant (see ANOVA, Table 3 below). (Design-Expert software only produces contour and 3D plots for interactions included in the model). In addition, in keeping with standard Response Surface Methodology practice, as SDBias is included in interaction terms, its main effect is also included.

$$NBA^2 = -80.513 + 0.419 * Reliability + 10.381 * CutScore - 0.775 * Bias - 0.082 * SDBias - 0.056 * CutScore^2 - 0.009 * Bias^2 - 0.073 * (Reliability * CutScore) + 0.008 * (Reliability * Bias) -.002* (Bias * SDBias) \quad (33)$$

Summary statistics for the Equation 33 are presented in Table 2 and the Analysis of Variance in Table 3 below. This model has a very high Adjusted R^2, 0.9887. Normally an R^2 this high is suspect in Social Science research. However, one must consider that the model is actually a summary of 30 simulations (each with 1,038,044 iterations, resulting in a total of 31,141,320 simulated test results) and that the simulations are use equations based on the same four factors as the model. The closeness of fit can also be seen in Figure 31, a plot of actual and predicted values of NBA^2.

Table 2:
Summary statistics for Net Bayesian Advantage, Equation 33, Power Transformation Lambda = 2

Statistic	Value	Statistic	Value
Std. Dev.	0.84	R-Squared	0.9922
Mean	16.82	Adj R-Squared	0.9887
C.V.	4.97	Pred R-Squared	0.9796
PRESS	36.74	Adeq Precision	63.7980

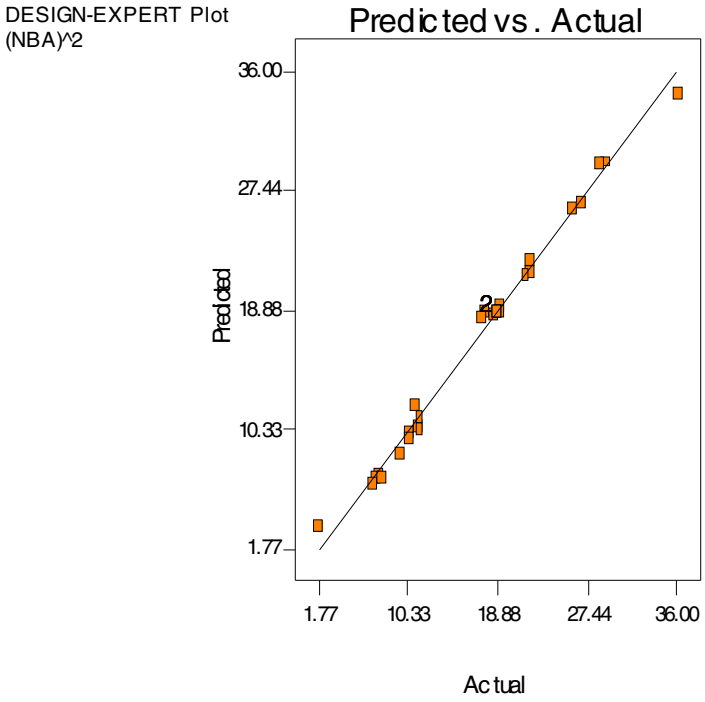


Figure 31: Actual vs. Predicted values of NBA^2 indicating close agreement.

The ANOVA table for NBA^2 is presented in Table 3. With an F statistic of 284, the model is definitely significant. The model has significant "Lack of Fit." On reflection, this is not as serious of a problem as one might think. First, with an R^2 of over 0.99, the model does fit the data well. Second, when the residual plots are reviewed, it will be see that the most extreme residuals are generally axial points, which although they are used to estimate the model, they are not within the range of interest. Third, the Sum of Squares for Lack of Fit is calculated by subtracting the Sum of Squares for Pure Error from the Residual Sum of Squares. Not only is the total Sum of Squares relatively small (producing the high R^2 of the first point), the Sum of Squares for Pure Error is based on the six replicates at the centerpoints, which, as discussed above, produce very similar results because the simulations are stable. Finally, the response has already been transformed, and there is little that can be done to improve the fit. While future research might model the results with different functional forms (which would have to be more complex than the simple Central Composite Design to prevent confounding—for example, a cubic term for Reliability suggested by the Residuals vs. Reliability plot in Figure 39—for purposes of this study it is concluded that the statistically significant Lack of Fit does not have any practical significance for the purposes of this study.

Next, the statistical significance of variables is described. Of the variables that are the focus of the study, Bias appears as statistically significant in three terms: It main effect (Bias, variable C), its squared term ($Bias^2$ or C^2) and in an interaction with Reliability (Reliability * Bias or AC). SDBias (variable D) is not

statistically significant in any term. However, it is included in the model so the response surface in Bias/SDBias (C/D) space can be examined in the response surface and contour plots below[28].

The other two variables, Reliability (A) and Cut Score (B), which are aspects of tests to which a Bayesian approach might be applied, are statistically significant. Reliability's main effect (A), and its interactions with both Cut Score (Reliability * Cut Score or AB) and Bias (Reliability * Bias or AC) are significant. Cut score (B) and its square (B^2) are included in the model. All significant terms involving Reliability and Cut Score are significant at the 0.01 level.

[28] One can not construct contour plots of variables excluded from the model in Design Expert ® software (Helseth, et. al. 2000)

Table 3
Analysis of Variance for Net Bayesian Advantage (NBA) Equation 33, with Power Transforamtion Lambda = 2

Source		Sum of Squares	DF	Mean Square	F Value	Prob > F
Model		1789.23	9	198.80	284.14	< 0.0001
Reliability	A	1423.88	1	1423.88	2035.06	< 0.0001
CutScore	B	190.59	1	190.59	272.40	< 0.0001
Bias	C	16.99	1	16.99	24.29	< 0.0001
SDBias	D	0.60	1	0.60	0.85	0.3673
	B^2	7.28	1	7.28	10.40	0.0042
	C^2	127.75	1	127.75	182.58	< 0.0001
	AB	20.79	1	20.79	29.72	< 0.0001
	AC	6.48	1	6.48	9.26	0.0064
	CD	0.053	1	0.053	0.076	0.7856
Residual		13.99	20	0.70		
Lack of Fit		13.91	15	0.93	53.28	0.0002
Pure Error		0.087	5	0.017		
Cor Total		1803.23	29			

Model Diagnostics Plots

As one of the major assumptions of linear models is that residuals are normally distributed, examination of a Normal Probability Plot of residuals is an important step. Figure 32, the normal probability plot of Studentized Residuals does not serious provide evidence for severe non-normality, although there is a slight departure near the center of the range, which may be due to centerpoints.

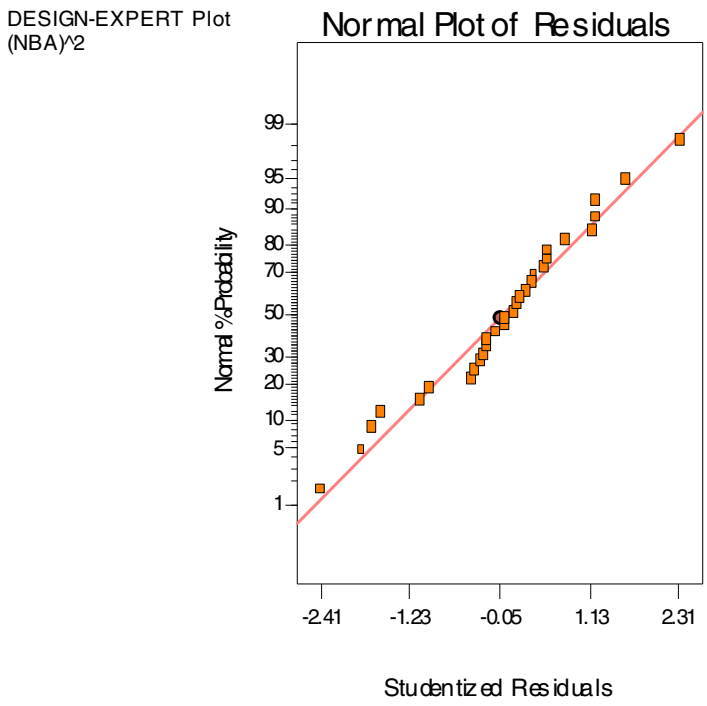


Figure 32: Normal probability plot of Studentized residuals for equation 33.

Another basic diagnostic is to look for individual observations with residuals that have extreme values, measured by the T-Statistic, plotted in Figure 33. Such outliers may indicate simple problems (such as transcription errors) or lead the researcher to investigate the observation to see if there is something identifiable about the observation. Frequently used rules of thumb for outliers are 3.0 and 3.5. None of the outliers here are beyond this level. An investigation of one of the larger residuals, Standard Order #18/Run Order #1[29], indicates that this is an axial point, which is beyond the range where we explore the response surface. There is also a slight hint that residuals increase with as

[29] Readers wishing more information on a residual may want to look at the results in Figure 28.

the runs progress (which, since these are simulations, could only be due to autocorrelation of the random number generators). However, this is not dramatic.

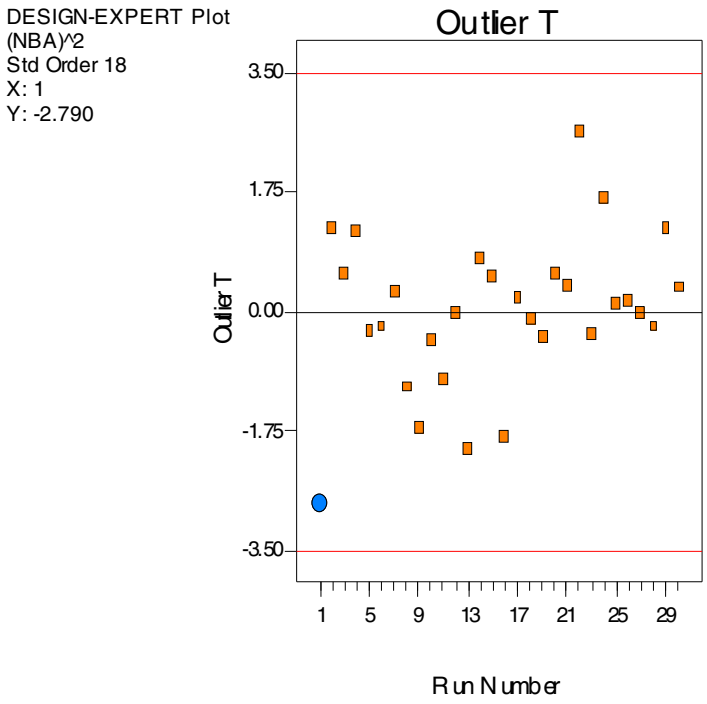


Figure 33: Outlier T statistics for residuals for Equation 33.

Another basic diagnostic is to look Residuals vs. Predicted values, presented in Figure 34, below. In this case studentized values are examined. Although all are within the rule of thumb of 3.0, one has been highlighted. Again, this is an axial point (Standard Order #17). This plot provides mild evidence of heterostochasticity, indicated by a hint of a v-shape, with a slightly greater spread at the lower values of predicted. However, the response (NBA) has already been squared to address heterostochasticity. Thus, few remedial measures are available. Moreover, Figure 35, the graph of the Standard Error of

the Estimate indicates small variation in the areas of interest. This area is Cut Score from 23 to 29 and Reliability from 84.0 to 94.4. In this area SEE ranges from about 0.26 to 0.38.

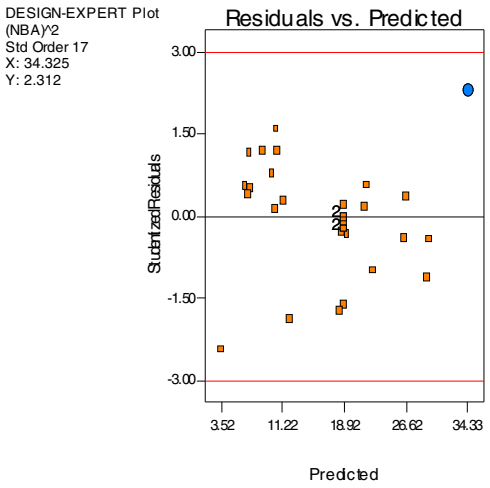


Figure 34: Residuals versus fitted values for Equation 33.

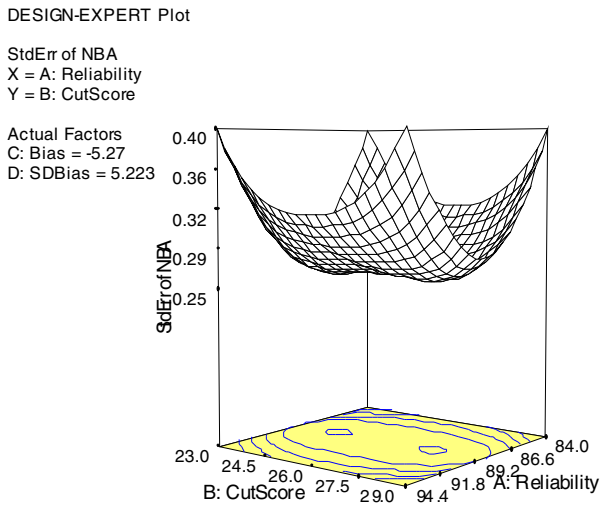


Figure 35: Standard Error of the Estimate (SEE) for Equation 33.

The next four plots, Figure 36 through 39, display residuals versus each of the four independent variables. Among the uses of these plots is the determination if transformed factors should be added to the model. For example, a quadratic shape for residuals versus a factor would indicate its squared term should be added to the equation. None of the four plots is dramatic, indicating no serious need for additional terms, such as raising one of the variables raised to a higher power. There is, however, a vague suggestion of a sinusoidal shape in the plot of Residuals versus Reliability, Figure 36. Again, while this is not dramatic, future research might explore the potential of improving the fit by adding cubic term. Such additions would require a much larger number of simulations in and a much larger Central Composite Design than is used in this study.

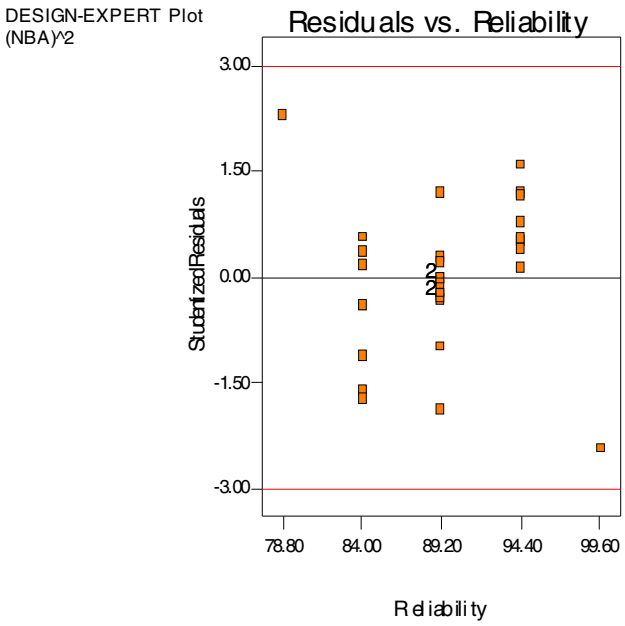


Figure 36: Residuals of Equation 33 versus Reliability. Indicating a slight, but not dramatic, sinusoidal shape.

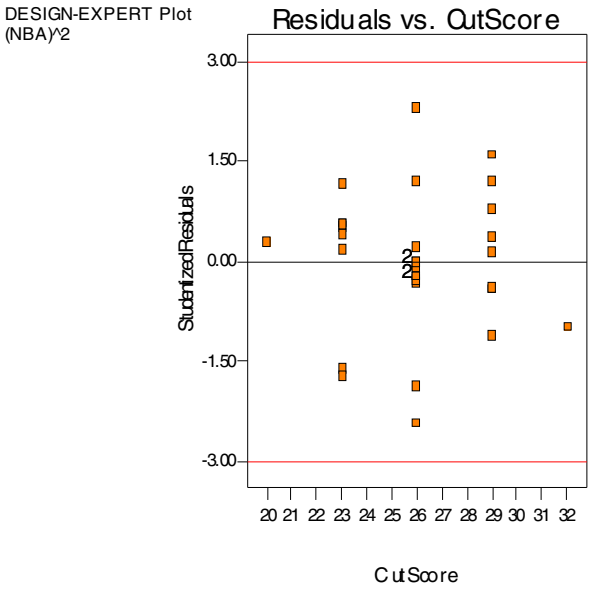


Figure 37: Residuals of Equation 33 versus Cut Score.

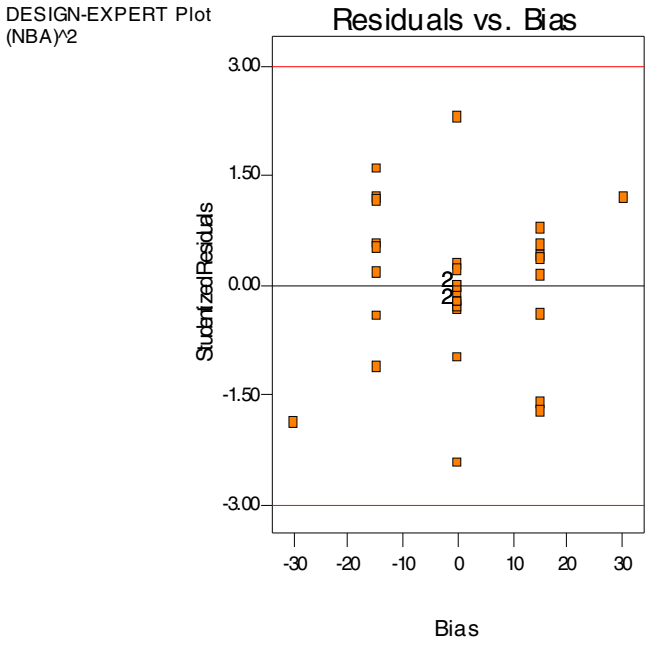


Figure 38: Residuals of Equation 33 versus Bias.

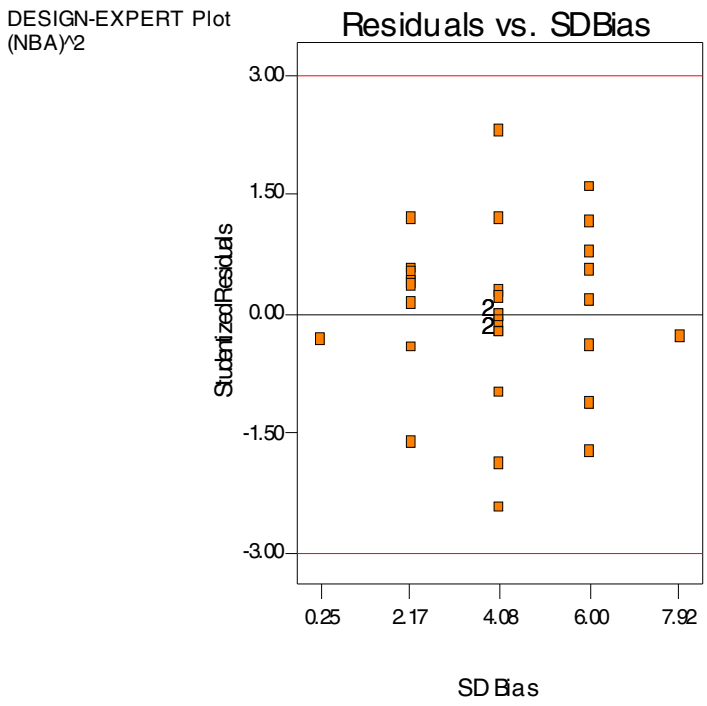


Figure 39: Residuals of Equation 33 versus SDBias.

The final two diagnostic plots for NBA (equation 33) are Leverage and Cooks Distance. The leverage plot, Figure 40 indicates the weight of each observation on the model. The weights are spread about their average, indicating no unusually influential point. Not surprisingly, the four points with the heaviest weight (all at 0.56) are axial points (Standard Order 19 through 22).

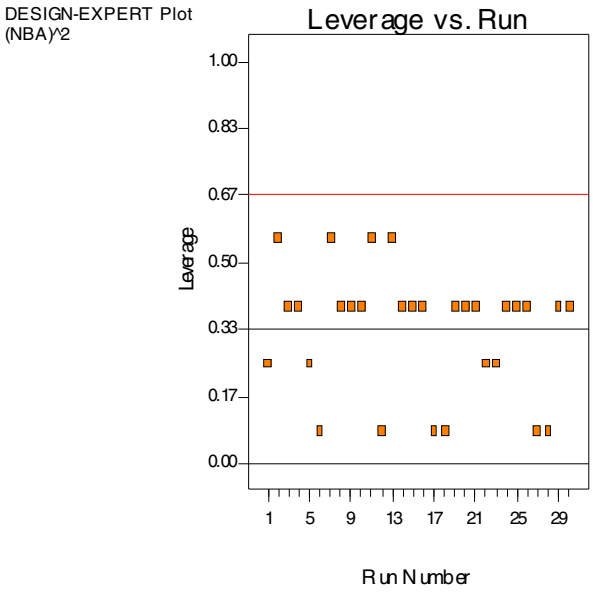


Figure 40: Leverage plot for Equation 33.

The Cook's Distances plotted in Figure 41 are generally well behaved. One which stands out slightly is an axial point (Standard Order 21).

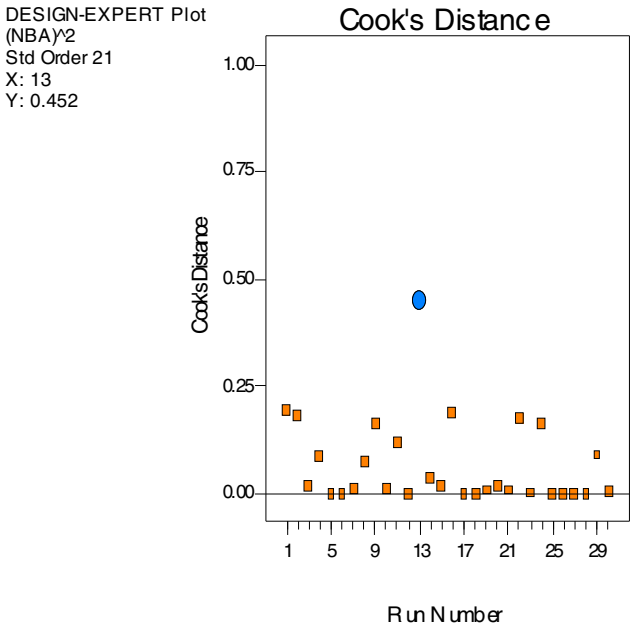


Figure 41: Cooks Distances for Equation 33.

Based on the model summary statistics and model diagnostics one can conclude that it is appropriate to use the Response Surface Model for NBA in Equation 33 which generated by a Central Composite Design. One can confidently explore the impact of Bias and SDBias (the focus of the study) as well as Reliability and Cut Score on Net Bayesian Advantage with response surface contour plots and other plots.

Graphical Analysis of response surfaces: THE CORE OF THE ANALYSIS

The 30 simulations of this study, Figure 28, are summarized by the Response Surface Model Equation 33. This equation is a five-dimensional manifold, with one independent variable, Net Bayesian Advantage (NBA) and four independent variables; the right hand side of the equation has a total of nine terms (including regression constant, squared, and interaction terms). As a nine term equation is beyond the ability of most observers to grasp easily, a number of Response Surface graphs have been developed to aid in understanding this equation. In this section, single factor, interaction, 3D (three dimension), and contour plots for NBA are presented. A contour plot is a two dimensional representation of a 3 dimensional surface (see Figure 43 for an example). Just as a topographical map presents elevation for a given longitude or latitude, or a weather map presents temperature, so too a contour plot for NBA represents equal levels (slices) through the manifold at various X/Y combinations. In the study, the X/Y combinations of most interest are Bias/SDBias (two aspects of a Bayesian Prior), Reliability/Cut Score (two aspects of tests to which a Bayesian approach might be applied), and Reliability/Bias.

Three graphs will be of particular interest, Figure 44, Figure 49, and Figure 55. Figure 44 is the contour plot for NBA as a function of Bias vs. SDBias. These two characteristics were the initial focus of the study because they are quantifications of the two important aspects of a Bayesian Prior, Accuracy and Consistency (measured by their absence, Bias of the Prior and Standard Deviation of the Prior, respectively). These are important because *if* there were large ranges where combinations of Bias or SDBias result in low or even negative values of NBA, it would be unwise to use a Bayesian approach. Figure 49 is a plot of two aspects of the test, Reliability vs. Cut Score. These are important as they can guide decision makers as to which tests would be most susceptible to improved classification of students by use of a Bayesian approach. The third graph of particular importance is the Perturbation Graph, Figure 55. This graph is essentially a graphical summary of the first and second order terms (excluding interactions) of Equation 33.

Several aspects of the following graphs are worth noting. Unless otherwise indicated the actual response (that is untransformed NBA rather than its square which was modeled) is graphed. This is an option of the Design-Expert ®Software (Helseth, et. al. 2000). Another feature of the software is that it adds to the contour plots a dot at design points, that is, combinations of the factors at which experiments/simulations were run. If more than one point exists at that combination, the number of runs is indicated. Finally, as a 3D or contour plot presents 3 dimensions of a 5 dimension manifold, levels of the other two independent variables must be chosen and held constant when a plot is printed.

These are listed to the left of the graph. In general these are the centerpoints of the CCD. It should be noted that Design-Expert software allows the careful data analyst to dynamically vary independent variables which are not on the plot to see the impact.

Impact of Bias and SDBias on Net Bayesian Advantage (NBA)

The initial purpose of this study is to determine what levels of two characteristics of a Bayesian Prior (the probability that a student meets State Standards), when combined with information from Observed Scores on a high stakes test, will result in improved classification of the test taker as either meeting or not meeting the standard when compared with the use of Observed Scores alone. The two characteristics that are the focus of the study are Bias and Consistency (which is measured by SDBias). An examination of their 3D Figure 43 and Contour Plot, Figure 44 conveys a great deal of information despite the fact that the interaction of Bias and SDBias is not statistically significant in the ANOVA, $p = 0.7856$ (Table 3). This lack of statistical significance is consistent with the Interaction Graph of Bias and SD Bias, Figure 42.

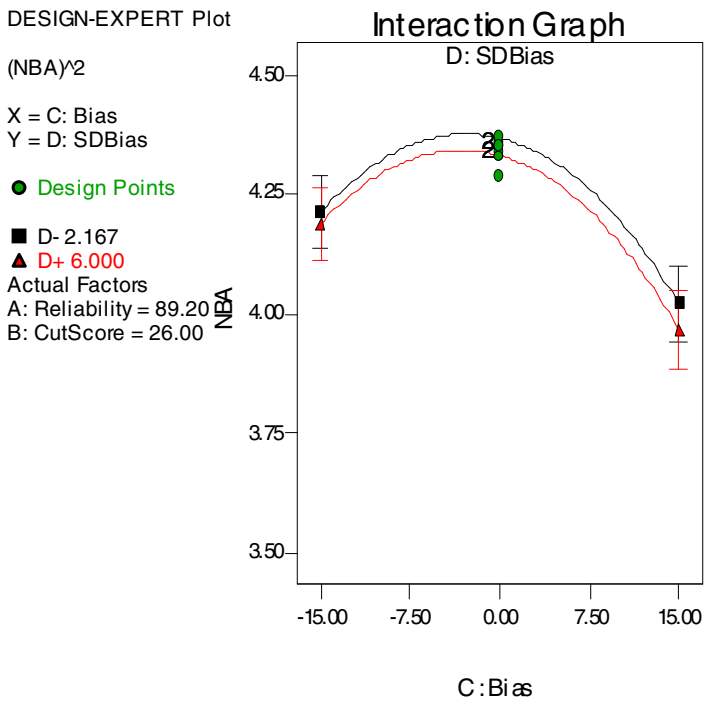


Figure 42. Interaction graph of Bias and SDBias for Equation 33. As the line do not cross and the standard error bars overlap, there is no evidence of interaction between Bias and SD Bias.

From the 3D plot, Figure 43, below, one can gain a perspective regarding the response of NBA to simultaneous changes in Bias and SDBias within the area of interest (0% +/- 15 percentage points for Bias and SDBias between 2.167 and 6.000). As NBA, Bias, and SDBias represent only three of the five dimensions of the Response Surface, levels must be chosen for the other two factors, Reliability and Cut Score, which are aspects of any test to which a Bayesian method for improving classification might be applied. In Figure 43 and 44 these levels are fixed at the centerpoints of 89.2 for Reliability and 26.0 for

Cut Score, the actual Reliability and Cut Score (for Meets Standards) for 99HS-MEAP-M, the results of which are the basis of the simulation.

A fruitful examination of the 3D plot, Figure 43, can begin by looking at the Z-axis (the vertical axis) along which NBA is graphed. Three aspects are interesting. First, NBA is always positive. Thus, at the centerpoints for Reliability and Cut Score, a Bayesian method is always better than relying on Observed Scores alone. Second, the range of NBA is between 4.0 and 4.4. This is a fairly flat response, indicating anywhere in the range of interest of Bias and SDBias, one would get approximately the same improvement. Again, this is consistent with SDBias not being significant (see Equation 33 and Table 3).Thus, if the test were again administered to a 75,000 group of students with the same distribution of True Scores, a Bayesian approach with as much Bias as +/- 15 (on average) and a SDBias of much as 6.0, would correctly classify at least approximately 3,000 more students than reliance on Observed Scores alone (4% of 75,000 = 3000). Third, there is definite curvature to the Response Surface relative to changes in Bias. Thus, as one moves from –15 to + 15 NBA first goes up and then down. This is due to the inclusion of a square term for Bias (C^2) in the model (see Table 3). Thus, as one would expect, Bias near zero results in better classification.

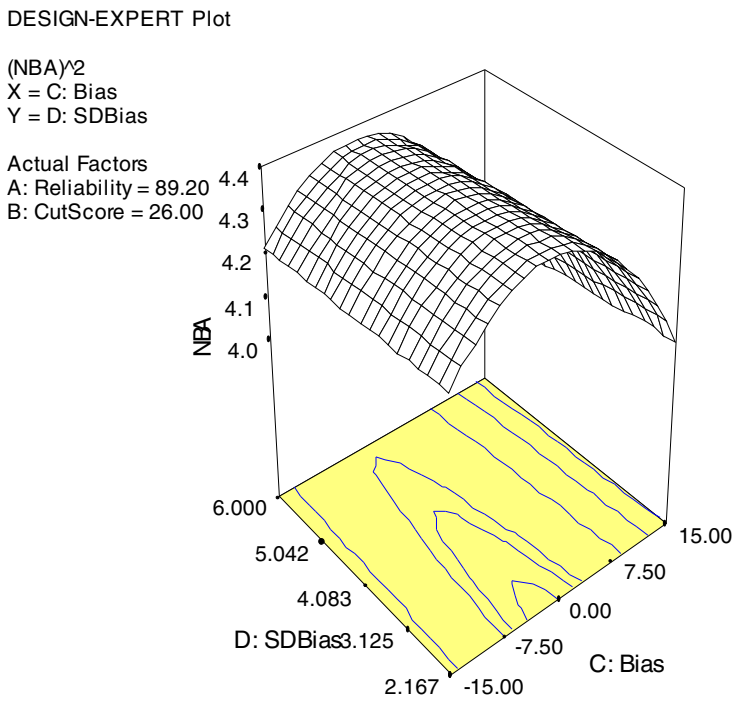


Figure 43: Response Surface of NBA (z-axis) in Bias and SDBias

The 'floor' of the 3D plot, a projection of the response surface, is the Contour Plot. It is presented in Figure 44 to provide easier visual analysis. The quadratic nature of the response of NBA as one mores from Bias of –15 to + 15 is clearly seen (the contours going from 4.2 to 4.4 and down to 4.0). Again, it is primarily the result of the $Bias^2$ term in Equation 33. Moreover, as one might expect, the optimum level of NBA is reached near zero Bias. (The fact that it is not at zero is addressed in the next paragraph.) On the other hand when one goes from 2.167 to 6.000 in the SDBias dimension, the largest change in NBA is 0.1, 1/10 of a percentage point. This is consistent with SDBias (termed variable D in the ANOVA, Table 3) not being significant in any term of the ANOVA. The

lack of statistical significance of the Bias by SDBias interaction means that the Response Surface is not twisted in this dimension. Thus, the 3D surface of Figure 43 looks like a smooth vault of Roman arch.

In the Contour Plot, Figure 44, the contour line with a maximum of 4.4 has been drawn emphasized to help the reader notice that the maximum is near the design point where 6 experiments were, as indicated by the 6 beside the dot. The levels of the experimental variables at this point were Bias equal to 0, SDBias equal to 4.083, (Bias and SDBias being read from the graph's axes), Reliability equal to 89.2, and Cut Score equal to 26.0 (Reliability and Cut Score levels indicated in the "Actual Factor" note).

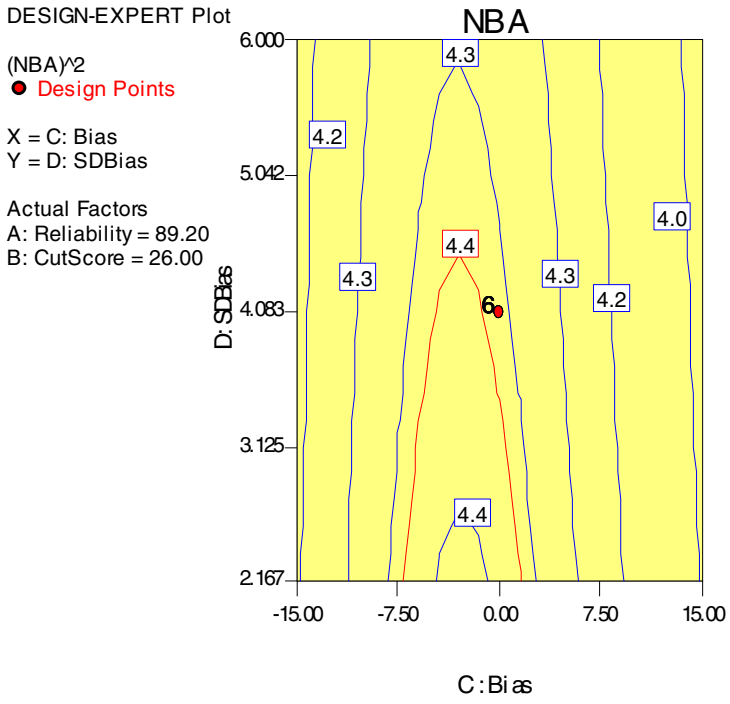


Figure 44: Contour plot for NBA versus Bias and SDBias, detail.

This quadratic curvature of NBA with respect to Bias is very clear in the One Factor Plot, Figure 45, below. This plot is essentially looking at the 'vault' of the 3D plot of Figure 43 edge-on when Reliability, Cut Score, and SDBias are set at their centerpoints (89.2, 26.0, and 4.083 respectively). This plot reveals an interesting aspect of the response surface. While the maximum for NBA (given the other factors) is *near* zero Bias, it is not *at* zero Bias.

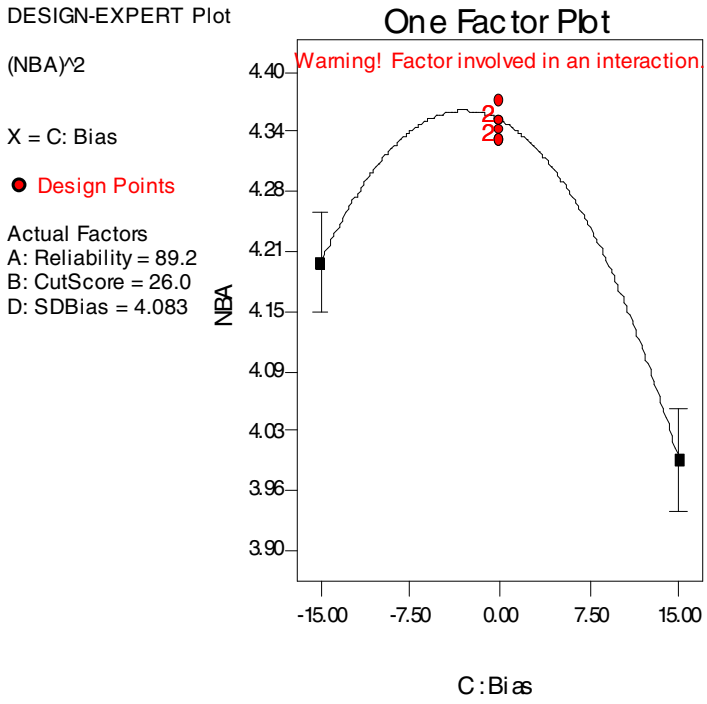


Figure 45: One factor plot for NBA versus Bias.

The One Factor Plot for SDBias is presented in Figure 46. Given that SDBias is not statistically significant in any term (main, squared, or interaction) it is not surprising that its graph is essentially a flat line as one goes from 2.217 to 6.0. Moreover, the error bars at the end points overlap.

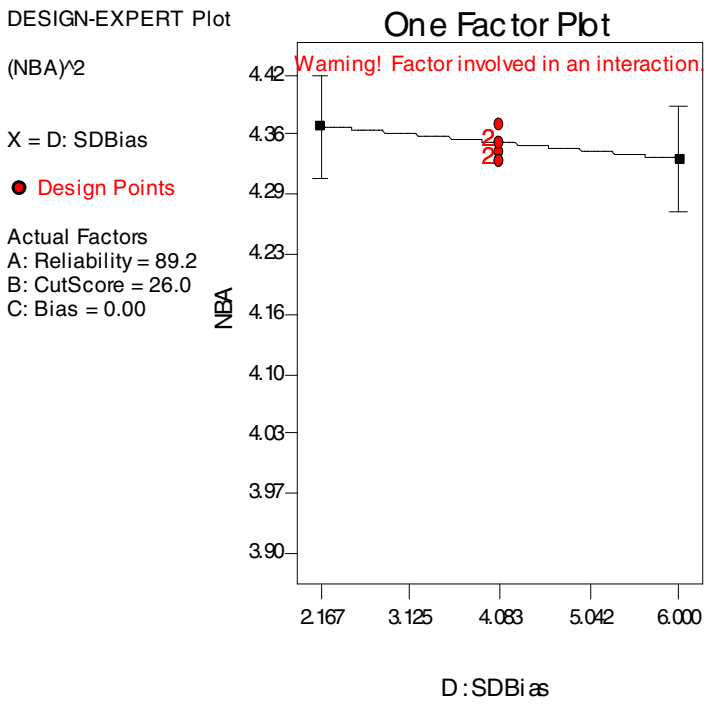


Figure 46: One factor plot for NBA versus Bias.

The initial purpose of this study was to determine what levels of two characteristics of a Bayesian Prior (the probability that a student meets State Standards), when combined with information from Observed Scores on a high stakes test, will result in improved classification of the test taker as either meeting or not meeting the standard when compared with the use of Observed Scores alone. In the above One Factor Plot for Bias, Figure 46, as well as the Contour Plot, Figure 44, one might expect the maximum NBA to occur where Bias is 0 and that NBA would decrease symmetrically as the magnitude of Bias increases whether in a positive or negative direction. This would be consistent with a statistically significant term for expect Bias^2 and no statistical significance

for the main effect of Bias. While the ANOVA Table 3 does indicate the expected statistically significance for $Bias^2$ (F-Statistic of 182.58 with P-value less than 0.0001), the main effect of Bias is also statistically significant (F-Statistic for of 24.29 with P –value of less than 0.0001). The impact of this combination of significance for main and square effects of Bias can be further explored by limiting the range of Bias in a contour plot to +/- 8.0, as in Figure 45, below. NBA reaches a maximum at approximately a Bias of –2.8 (the 'peak' indicated by the Predi(ction) flag). The fact that the offset in the negative direction is due to the negative coefficient on Bias in equation 33. The strength of the statistical signal ($p < 0.0001$) suggests that this is not just due to random chance, i.e., a result of the samples resulting from the simulation. While the opportunities for further research on this topic are discussed in Chapter 5, the following discussion is sufficient for the present.

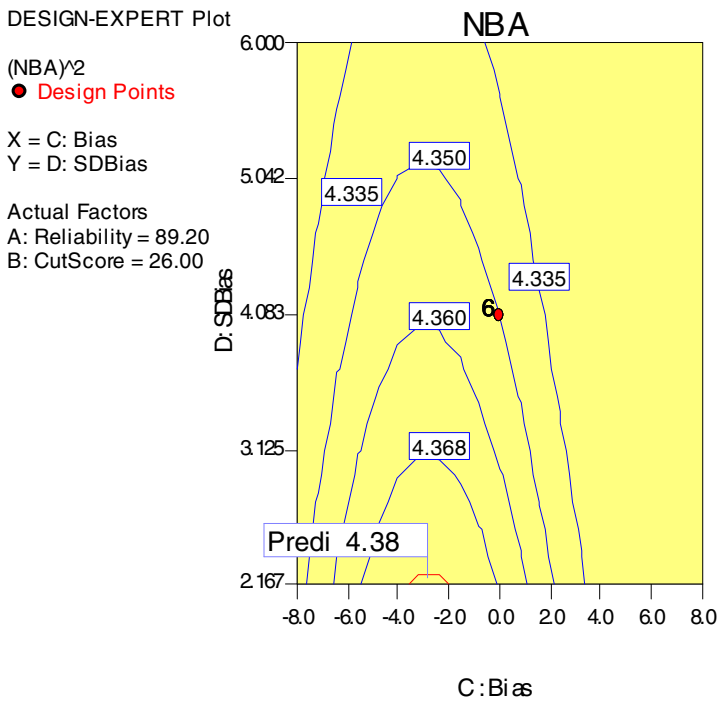


Figure 47: "Close up" contour plot of NBA as a function of Bias and SDBias.

When the planning for this dissertation was in the formative stages, the major advisor, Professor Shlomo S. Sawilowsky, suggested that using an *actual* distribution of high stakes test scores be used rather than relying on simulating scores (for example, using a normal distribution). Among the reasons for this was the work of Micceri (1989) who found that the Observed Scores of many educational and other psychometric tests were severely non-normal and might exhibit features like digit preference. Digit preference is means that there might is not a smooth transition between adjacent scores. Looking at the underlying data used in this study, actual scores of the spring 1999 administration of the Michigan Educational Assessment Program High School Mathematics test

(99HS-MEAP-M), Figure 12, one sees that the distribution is far from normal and a substantial digit preference exists.

Although it is a subject for further study in this research program, it may be that the distribution of true scores makes a difference to where the optimum Bias falls. However, while this is an interesting aspect of Bias, it is unlikely to be of great practically significant.

Impact of Reliability and Cut Score on Net Bayesian Advantage (NBA)

There are a number types of factors which would influence the usefulness of a Bayesian approach in classifying more of test takers correctly as either meeting or not meeting a standard of a high stakes test. Among these are aspects of the test takers, the Bayesian Priors, and the tests. The abilities of the students, reflected in the distribution of skills/knowledge of the students, has been held constant by the use of a single distribution of Estimated True Scores which is based on the actual scores of the 99HS-MEAP-M. Discussions above indicate that exploring this might be of interest, though perhaps of limited practical significance. This same discussion indicated that, within range under investigation, Bayesian Priors within a broad range of Bias and SDBias consistently improve NBA (Net Bayesian Advantage) which is the measure of improved classification though a Bayesian approach. Next we will explore aspects of the tests themselves. It is important to look at aspects of the tests to insure that a Bayesian approach might have wide applicability. The two aspects of the test selected here are Reliability and Cut Score.

In this section the impact of two aspects of the test—as opposed to the estimators of the Prior Probabilities—are explored. Specifically the impact of Reliability and Cut Score on the improvement in classification of test takers obtained by using a Bayesian approach versus relying on Observed Scores alone is examined using Response Surfaces and Contour Plots. This improvement is measured by NBA and modeled by Equation 33. As with the other two factors (Bias and SDBias) the analysis begins by reviewing a 3D plot. Figure 28 has NBA on the Z-axis and Reliability and Cut Score on the X/Y axes. The levels of the other two factors must be chosen for a 3D representation of a five dimensional manifold. The centerpoints of Bias (0.0) and SDBias (4.083) are used. When compared with the 3D surface of NBA versus Bias/SDBias presented in Figure 43, the Relaibility/CutScore surface (Figure 48) has a very different shape (a plane with a slight twist rather than a smooth vault) and a much wider range of NBA (about a 2.5 percentage point range from about 3 to 5-½). This relationship of Reliability, Cut Score, and NBA is more clear in Figure 49, the Contour Plot. This Contour Plot has the same center point levels of Bias and SDBias, 0.0 and 4.083, respectively.

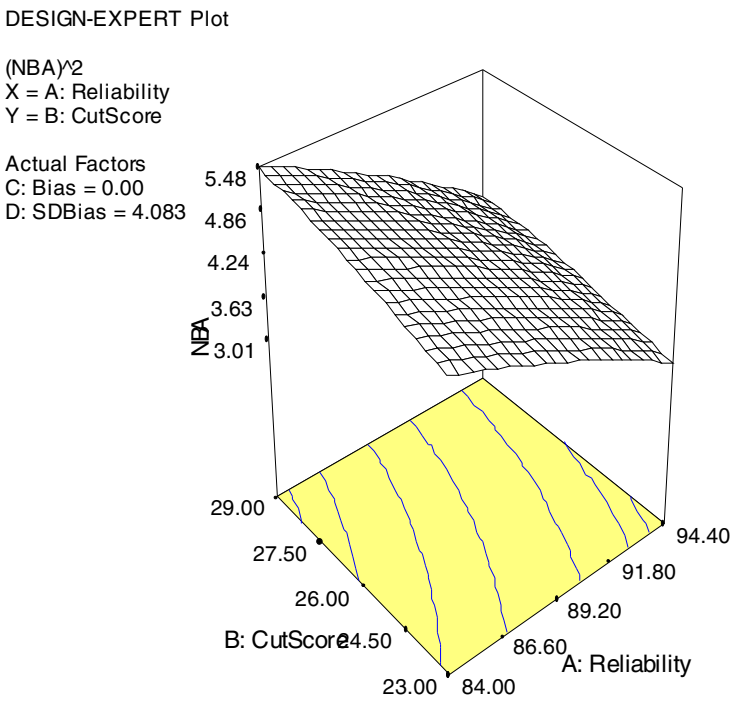


Figure 48: Response surface of NBA as a function of Reliability and Cut Score (equation 33).

In the contour plot below, Figure 49, which covers the range of interest of Reliability and Cut Score, NBA is at a minimum (3.01) in the lower right hand corner (Reliability equal to 94.4 and Cut Score equal to 23) and at a maximum (more than 5.47) in the upper right hand corner (Reliability equal to 84.0 and Cut Score equal to 23). The steepest downward gradient is in the direction of increased Reliability. For example, moving from Reliability of 84.0 to 94.4 along a Cut Score of 23, NBA drops 1.69 percentage points, from 4.70 (as seen in the contour that intersects the origin) to 3.01 (as indicated by the Predi[cted] flag). In contrast, going up the Cut Score axis from 23 to 29 results in an change which is nearly a full percentage point less in absolute magnitude (0.77 vs. the previously discussed 1.69), from 4.70 to 5.47. In addition, by inspecting the contours of

NBA at the various levels of Cut Score, we can see that moving from Reliability of 84.0 to 94.4 results in decreases of 1.69 to 1.94.

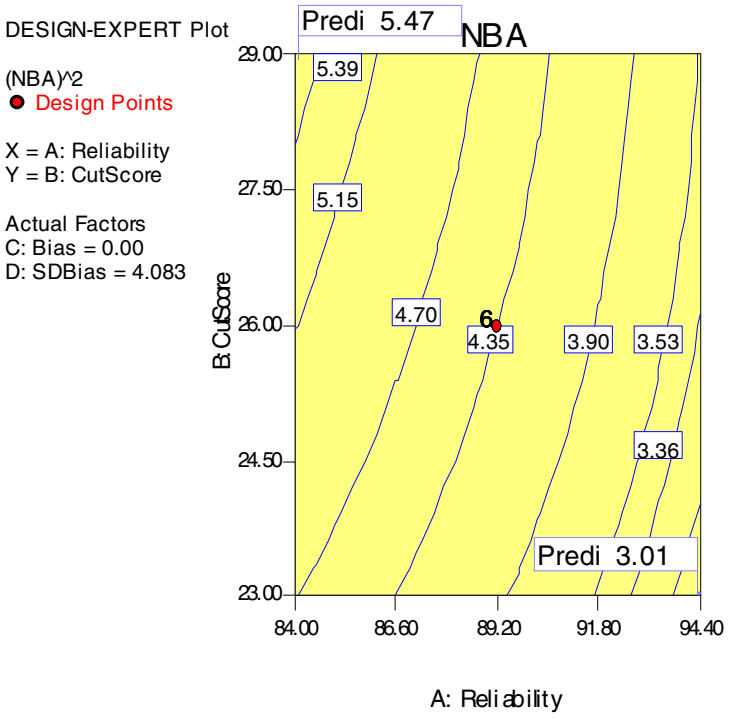


Figure 49: Contour plot of NBA from equation 33 by Reliability and Cut Score with Bias and SDBias at their centerpoints.

In Figure 49, Bias and SDBias were held constant at their centerpoints. Figure 50 below holds these two factors constant at two different levels. Again, NBA is the response of interest. Here levels of Bias and SDBias which are factorial points, the +1 levels for values for Bias (+15.0) and SDBias (6.0) are selected. These are the levels that result in a contour plot with smallest NBA

(within the levels of Reliability and Cut Score of interest). Here NBA is at least[30] 2.55 (at the bottom right corner, marked with a dot to indicate a simulation was run at this combination of the four factors) and is as high as 5.12 percentage points. The change in NBA as one moves along either the Reliability or Cut Score axis is approximately the same as when Bias and SDBias are at their centerpoints (in the prior Contour Plot, Figure 49). Thus, one can see that the relationship of interest (the change in NBA given changes in Reliability and/or Cut Score) is consistent across a wide range of values of these two aspects of the test.

[30] It is worth stressing that with these results NBA is always positive throughout the range of variables studied; thus a Bayesian approach will always improve classification by 2.55 percentage points in the region studied.

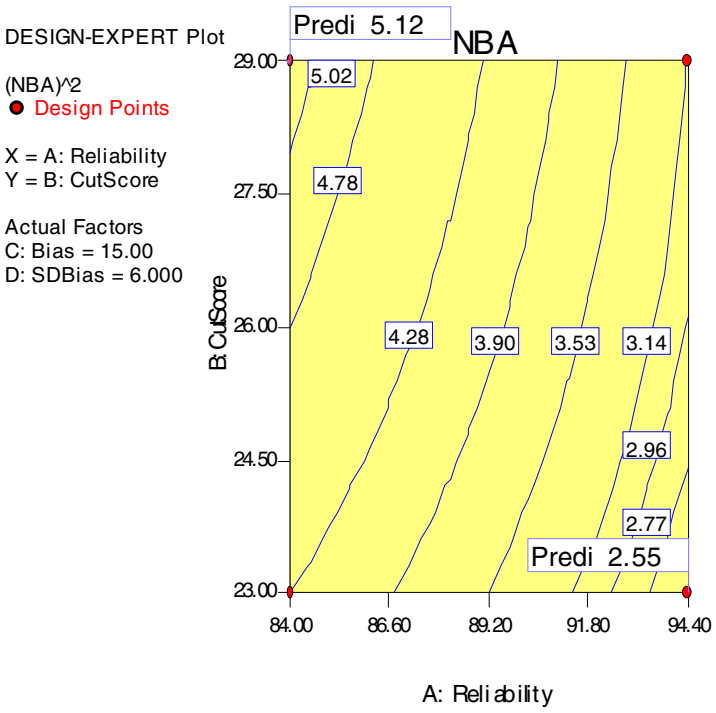


Figure 50: Contour plot of NBA from equation 33 by Reliability and Cut Score with Bias and SDBias at a factorial point.

Reliability has a statistically significant interaction with two factors, Cut Score ($p < 0.0001$) and Bias ($p = 0.0064$) (see ANOVA Table 3). The Reliability/Cut Score interaction (which is an interaction between two aspects of the test) was explored above. The discussion now moves to an interaction of an aspect of the test (Reliability) and an aspect of the Bayesian estimate (Bias). In the following 3D graph, Figure 51, the interaction of Reliability and Bias (with Cut Score and SDBias at their centerpoints of 26.0 and 4.083, respectively) can be seen in the fact that the Response Surface is a slightly curved and an even more slightly twisted plane. Again the relative impact of Reliability and Bias is easier to see in the Contour Plot.

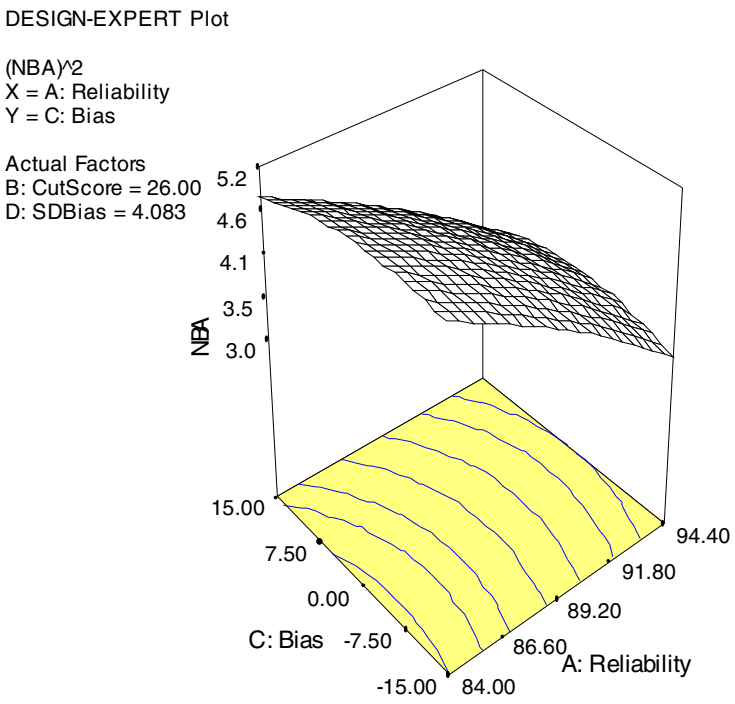


Figure 51: Contour plot of NBA from equation 33 by Reliability and Bias.

In the corresponding Contour Plot, Figure 52, one can see that the impact of changing Reliability is much greater than that of changing Bias. Starting at a Bias of 0 at low Reliability of 84.0, one passes from contours at NBA of 5.1 to 3.4 as Reliability increases to 94.4. On the other hand, the largest increase in NBA contours one can achieve traveling from Bias of –15 to +15 is only about 0.4 points.

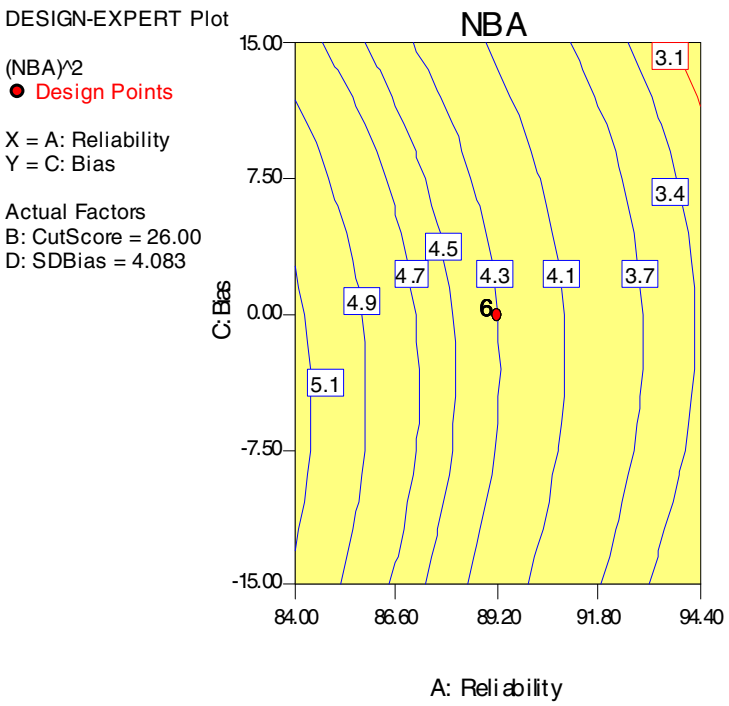


Figure 52: Contour plot of NBA from equation 33 by Reliability and Bias.

This graphical section on aspects of the tests concludes with the One Factor Plots for Reliability (Figure 53) and Cut Score (Figure 54). Again the software provides a warning that One Factor Plots should not be looked at alone unless the interactions are explored (as has been done above). These graphs, which have the same scale for NBA, indicate more impact on NBA by changes in Reliability. It is not surprising that more Reliable tests result in lower NBA, for a perfectly Reliable test would always classify the test takers the correctly. The impact of Cut Score may be significant because of the shape of the distribution of True Scores.

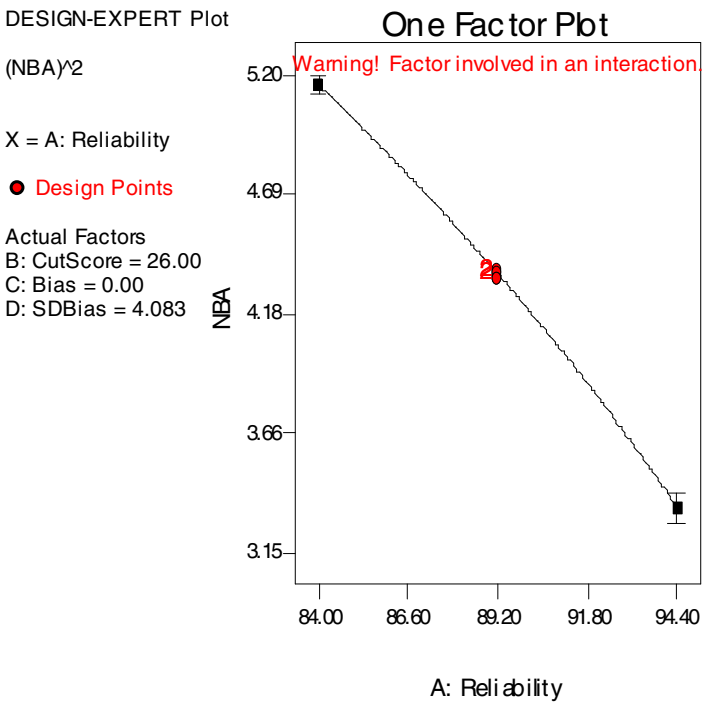


Figure 53: One factor plot for Reliability. Note the large change in NBA as one moves from lower (84.0, coded as -1) to higher (94.0, coded as +1) Reliability. (The computer generated warning is an indication that interactions should be studied, which has been done above).

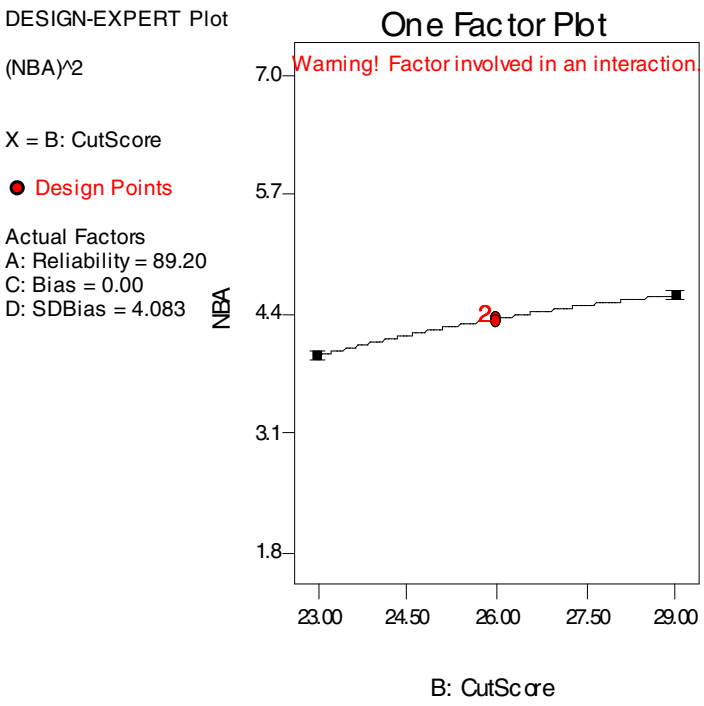


Figure 54: One factor plot for Cut Score. Note: The change in NBA as one moves from a lower (23.0, coded -1) to a higher (29.0, coded +1) level of this factor is smaller than the change in NBA for a corresponding change in Reliability, presented in Figure 53. (The computer generated warning is an indication that interactions should be studied, which has been done above).
.

The following graph, a Perturbation Graph (Figure 55), along with Figures 44 and 49, is one of the most important graphs in the study. It can be seen as a graphical summary of many of the results discussed to this point in the paper and it provides the motivation for the next Chapter. In this Perturbation graph for NBA, the quadratic impact of Bias (line C) is clear, as is the fact that its optimum is not at zero. SBBias (line D) is flat, consistent with its lack of statistical significance. The most dramatic impact if of Reliability (line A), which has a negative impact on NBA (more reliable tests resulting in a smaller advantage for a Bayesian approach). The positive impact of Cut Score (line B) is also

apparent. The next section explores the impact of Reliability and Cut Score on NBA (Net Bayesian Advantage).

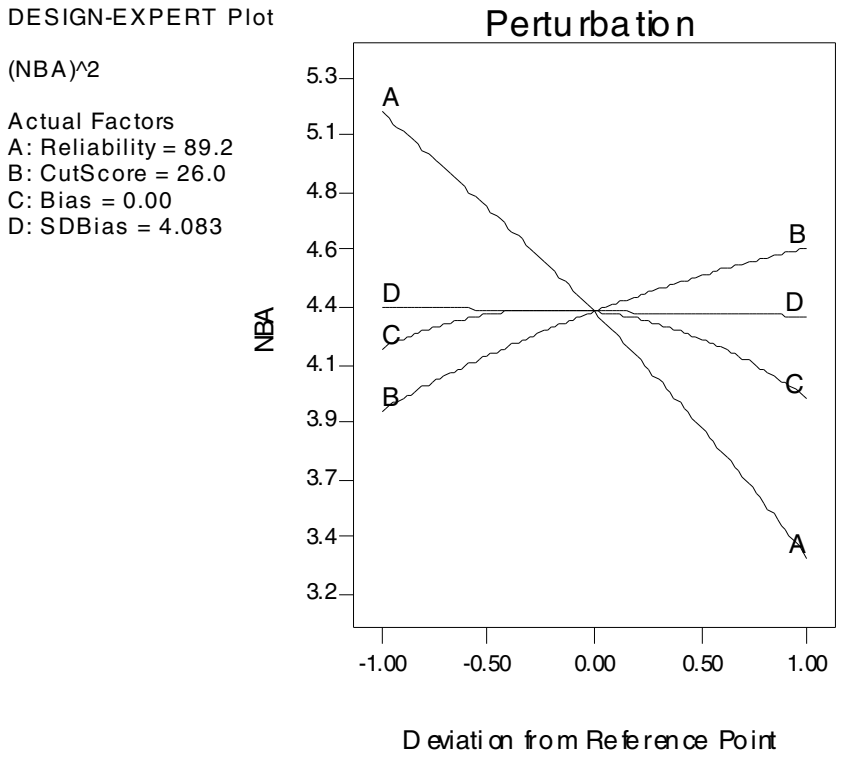


Figure 55: Perturbation graph for equation 33. This is a graphical summary of the relative impact of the main and squared effect of the four variables in the study.

A Note of the Costs of the Bayesian Approach: BWO (Bayesian Worse Than Observed)

The major results of the prior section are that, *for the data under examination* (Appendix A):

- Over the range studied of two aspects of the test (Reliability and Cut Score), a Bayesian approach would have improved the classification of test takers, and
- Over the range studied of two characteristics of the Bayesian Prior (Consistency [measured by the standard deviation around the Bias, SDBias] and Bias) a Bayesian approach would have improved the classification of test takers.

In addition, changes in Bias (Bias between –15.0 and + 15.0 and SDBias between 2.167 and 6.0 had little impact on NBA, Net Bayesian Advantage. On one hand this is good news because of the difficulty in evaluating the Bias and Consistency of a Bayesian Prior[31]. This good news might be taken as suggesting the surprising conclusion that little effort should be expended in attempting to develop an unbiased estimate which relates to the 'True Probability' that a test taker does or does not meet the standard (see equation 29 and related discussion) because it will have little impact on the *net* number of test takers that will be correctly classified. However, this conclusion could only be valid if NBA

[31] See chapter 5 for discussion regarding that this is not an uninformative prior.

were the only measure of adequacy of the Bayesian Prior in which policy makers and practitioners were interested.

To aid this discussion, Equations 30 and 31 are repeated below.

$$NBA = (tfOMbf + TmofBM) - (tfofBM + tfofBM) \quad (30)$$

$$BWO = tfofBM + tfofBM \quad (31)$$

BWO is part of NBA that adjusts NBA for classifying some test takers incorrectly when using a Bayesian approach while using the Observed Score alone would have resulted in the correct classification. One would wish BWO to be as small as possible for a given level of NBA. For example, a Net Bayesian Advantage of 2% could be achieved in a number of combinations of the factors in Figure 28 Specifically, it could result from a Gross Bayesian Advantage of 5 and a BWO of 3, or a Gross Bayesian Advantage of 2.1 and a BWO of 0.1. Assuming that it is preferable not to "fund" the NBA at the expense of test takers who would otherwise be correctly classified, the second case is far preferable because 2.9% more of students who would have been correctly classified with the Observed Score approach as compared with the second case (BWO of 3.0 – 0.1 = 2.9). While it has been shown in the previous section that Bias matters little with respect to NBA, this section will explore whether Bias impacts BWO.

BWO was modeled separately using Response Surface Methodology and the same independent variables that were used for NBA. As this section on BWO serves as a cautionary note for implementation, a full analysis of the adequacy of the resulting Response Surface Model (Box-Cox transformation,

ANOVA and residual analysis, leverage analysis, etc.) is beyond the scope of this paper. However, two of the contour plots will be presented in this section and their implications discussed in Chapter 5.

Figure 56 presents a contour plot for BWO (Bayesian Worse than Observed) as a function of Reliability and Bias when Cut Score and SDBias are held constant at their center points of 26.0 and 4.083, respectively. This is a graphical representation of the linear, square and interaction effects of Reliability and Bias on BWO. Over the range of interest for Reliability (84.0 to 94.4), BWO changes less than 0.1 (at Bias = 0). This contrasts shapely with NBA, where Reliability is the factor with the most impact.

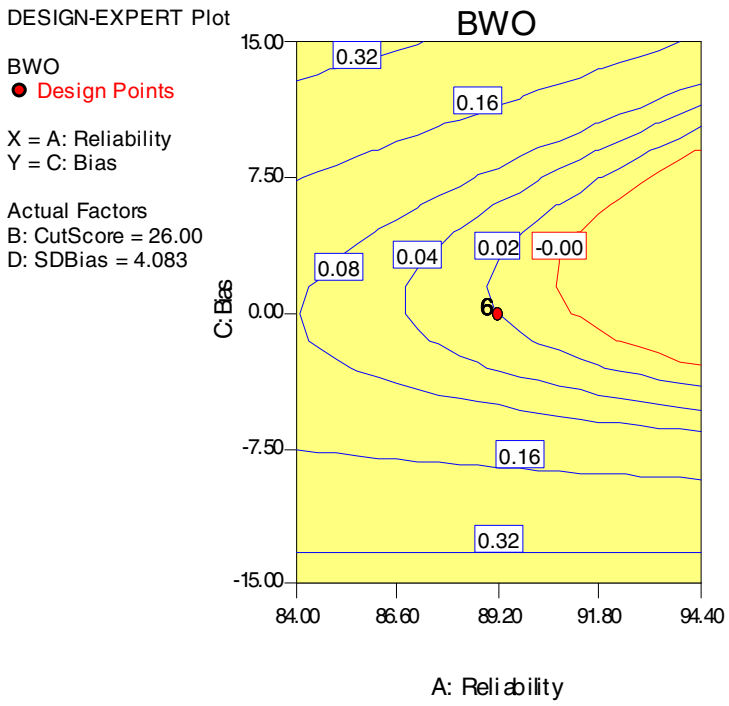


Figure 56: Contour plot of BWO as a function of Reliability and Bias with Cut Score and SDBias held constant at their centerpoints.

In Figure 57 the impact of Bias and Reliability is once more explored, the difference being that Cut Score is now held at 29.0, the high range of the area of interest (coded as +1 in a CCD). Here BWO reaches almost ½ of 1% when Bias is at -15.0, regardless of the level of reliability. In the data set under study, this would result in 375 (75,000 x 0.005 = 375) test takers being incorrectly classified using a Bayesian approach than would have been correctly classified if Observed Scores were used. This could be reduced to at most about 70 students with Bias near zero (75,000 x 0.0009 = 67.5).

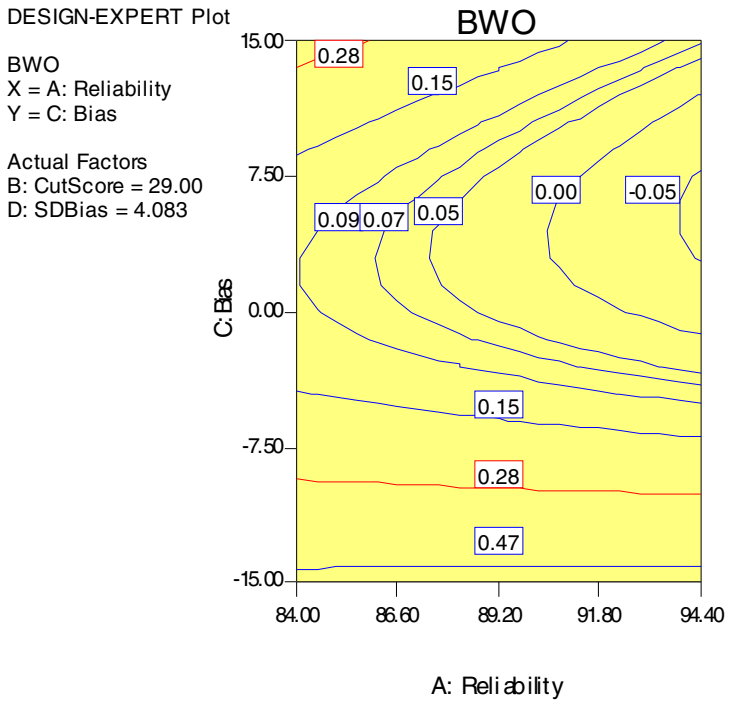


Figure 57: Contour plot of BWO as a function of Reliability and Bias with Cut Score held constant at 29.0 (coded +1) and SDBias held constant at its centerpoint.

From the above, it is clear that although use of Bayesian approaches is preferable to reliance on Observed Scores alone (for this dataset) on a net basis, regardless of Bias, reduction in Bias does result in a fewer test takers going from a correct classification to an incorrect classification. Thus, small bias in Priors is preferred.

"The final arbitrator in philosophy
is not what we think
but what we do."
Ian Hacking

CHAPTER 6

CONCLUSIONS, PRAGMATIC APPLICATIONS OF RESULTING PSYCHOMETRIC MODESTY, AND FURTHER RESEARCH

Synopsis of Chapters 1 through 4

A brief synopsis of the paper to this point is in order before discussing conclusions, possible pragmatic applications, and areas of further research. In the Introduction (Chapter 1) the role of high stakes educational tests was discussed. These tests have profound consequences for students, parents, teachers, schools, school districts, communities, and politicians. The consequences include graduation, teacher bonuses, pay and promotions, scholarships, and even property values. In the Review of Philosophical Foundations and Other Literature (Chapter 2), a Bayesian Epistemological Framework was demonstrated to be reasonable. The expanding role of Bayesian statistics in educational research, particularly IRT, was noted.

Chapters 3 and 4 explored the use classical statistical procedures (Monte Carlo Simulation, Logistic Regression, and Response Surface Methodology) to determine under what circumstances a Bayesian approach might result in better classification of high stakes test takers than Observed Scores alone for one

specific high stakes test[32]. The specific circumstances considered were aspects of a potential prior estimate of a student's chances of passing (Accuracy and Consistency, measured by their absences, Bias and SDBias) and aspects of the test (Reliability and Cut Score). The ancillary hypothesis of Classical Measurement Theory was employed to obtain a set of estimated true scores from a set of Observed Scores. The Priors were combined a Data Probability (based on a logistic regression which in turn was based calculation from the Observed Score) using a form of Bayes's Theorem favored by philosophers to produce a posterior probability and classification (meeting or not meeting standards). It was found in the case studied that over a wide range of Bias and Consistency of the prior as well as a wide range or Reliability and Cut Score for the test, between 2-½ % to 5-½ % additional test takers would be correctly classified (as having the same classification as the Estimated True Score) than would have occurred using Observed Scores alone.

Conclusions and Further Research

As only one test in one subject in one state for one year was used (Appendix A), the results by no means generalizable to all high stakes tests. However, a reasonable conclusion is that that the study provides strong evidence that there are circumstances in which a Bayesian approach can improve the classification of students *if our interest is classifying them according to their 'true' level of educational attainment rather than using a mere operational approach*.

---

[32] The Michigan Educational Assessment Program (MEAP) High School Test—Mathematics (Form E) for 11 Grade First Time Testers in Spring 1999 (Appendix A).

The operational approach defines meeting an educational requirement *as* receiving an Observed Score equal to or above the Cut Score on any permitted administration of a test. At minimum this study supports the contention that additional research in this area might be fruitful.

The Results of Chapter 4 are certainly motivation for further scientific research. This would include the impact of other functional forms of distributions of True Scores (compared with that used in the study, see Figures 12-14), the detailed empirical modeling of BWO (proportion of testers who would have been correctly classified using Observed Scores alone but are incorrectly a Bayesian approach), and the investigation of instances of several anomalous results (NBA not reaching at its maximum when Bias is zero).

In the present study only one distribution of scores was used (Attachment A). Future research should include collections of data (from the literature or by contacting state Departments of Education directly) on actual administrations of a number of high stakes tests, including distributions of Observed Scores, Cut Scores, and Reliability), for example, Cronbach's Alpha. This task is part of the present research programme. Second, research on predicting Observed Scores could be used in developing better information on the Priors used in future studies. This could draw on the work of other researchers on the prediction of success in high stakes tests. One would expect there is already research underway on estimating observed high stakes test outcomes from other student data (perhaps grades, absenteeism, difficulty of courses, etc.

With these two elements in hand, Response Surface Methodology can be used to develop generalizable models for NBA (Net Bayesian Advantage) and BWO (Bayesian Worse than Observed) and each of the eight classifications in Figure 15. In addition, although it is not the focus of this research program, these models would also be informative as to the potential improvement in correct classification from incremental improvements in Reliability. Another important question concerns bias. It is possible that more students of specific ethnic and SES (socioeconomic status) groups would be reclassified because of access to factors which might be in an equation which models anticipated probability of being proficient. This would include attendance, access to rigorous[33] and advanced placement courses, etc.

Pragmatic Applications and Further Research

Even if further research indicates that a Bayesian approach generally results in substantial classification improvement (which the present author considers likely in view of the results of Chapter 4), this in itself is not sufficient to recommend practical application of the approach. Keeping in mind "Hume's guillotine," the philosophical principle which states that 'one cannot derive an ought from an is,' it is clear that additional research *and* judgments are needed before specific actions *informed* by the anticipated confirming research on the superiority of a Bayesian approach can be specified. Moreover, as the very idea

---

[33] Names of courses are not sufficient. Anecdotally, I have heard it said that in some urban school districts a teacher might be teaching class called "Algebra II."

of meeting an educational standard is a social construct specified by a technopolitical process, political[34] considerations must be considered before using Bayesian reclassification.

A key—perhaps the most important—question is whether students who would be reclassified as being proficient by a Bayesian approach would be likely to be classified as proficient if retested. A similar question could be asked of students who are reclassified using alternate means, such as portfolios. Pilot studies could determine this through data analysis *without* actually changing the student's status.

Before implementation of such a policy, research is needed to determine if students who would be reclassified would be likely to study and learn a) more or b) less. In other words, what is the likelihood of students being better or more poorly educated? For example, it is possible that a student who is 'reclassified' as proficient by a Bayesian approach will 'coast,' whereas one who has been classified as not proficient by an Observed Score approach will work harder to learn what is necessary to 'pass' the test.[35]

Assuming future research produces data that increases our confidence[36] that a) a Bayesian approach will result in more accurate classification of testers as meeting educational requirements, and b) students who are reclassified will

---

when in fact the material is at a pre-algebra level.

[34] Politics has been defined as the art of the possible.

[35] It is possible that research by others will indicate students who face the prospect of a high stakes test are more likely to study—and thus learn more—than those hat do not.

learn no less, it is still not sufficient to use a Bayesian approach.[37] There are two types of questions remaining: Political and cost/benefit. (Figure 58 provides a flowchart including research and cost benefit factors. The path that the present author considers the most likely is in bold.)

---

[36] Note Bayesian phraseology.

[37] Note: The method for doing the reclassification is beyond the scope of this paper.

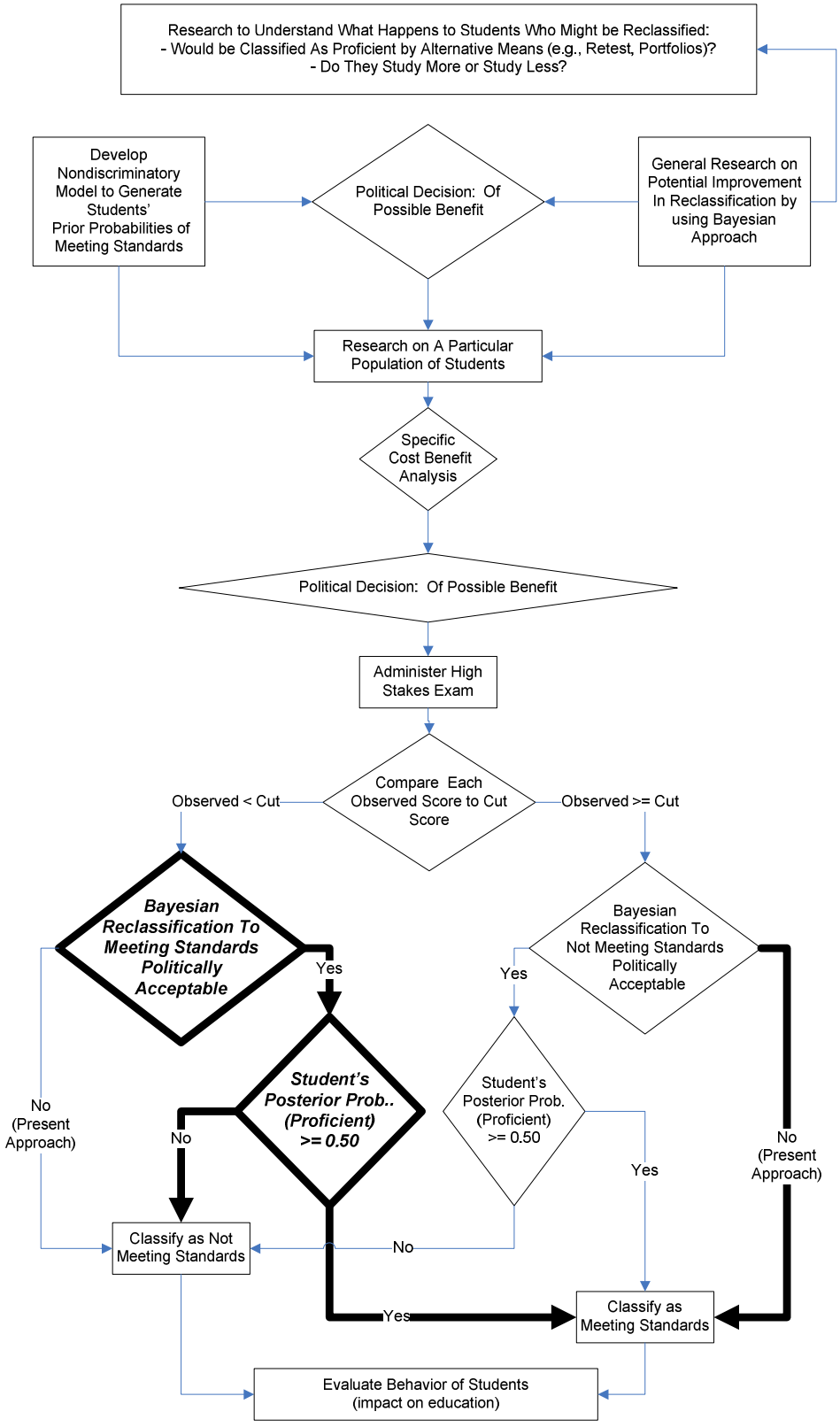


Figure 58. Potential pragmatic application

While it might be politically acceptable to seem to "give students credit" for their school performance in reclassifying them from 'not meeting standards' to 'meeting standards (this could be done, among other ways, with a logistic regression equation based on grades, attendance, and course rigor), it is unlikely that it would be acceptable for students to be reclassified as 'not meeting standards' if their Observed Score was above the Cut Score. Two "thought experiments" of a hypothetical state uses such Bayesian reclassification are sufficient to demonstrate the political minefield that classification from meeting standards (judged by Observed Score alone) to not meeting standards (combining Observed Score and Bayesian Priors).

Imagine the Speaker of the State Legislator, the head of the opposition party, and the State Superintendent of Education and Chair of the State Board of Education are in the Governor's office watching one of the following two news reports:

- Internationally known civil rights leaders are holding a news conference on the steps of the state capital with dozens of high school students—both minority and poor non-minority students—decrying the injustice of reclassifying students who have 'passed' the high stakes test as not passing, with the result that they are unjustly denied a high school diploma.
- One of the state's senators in Washington, who happens to be the Chair of the Senate Education Committee, is on the US capital steps announcing she is about to launch an investigation into adjustment of

scores of high stakes tests. (She had just received a letter indicating that although her son had an Observed Score equal to the Cut Score, he had been classified as not meeting state standards based on an adjustment based that includes opinion.)

As many politicians (and politically astute educational administrators) are adept at avoiding political embarrassment like those in the two thought experiments above, any system which uses prior information to reclassify testers who have Observed Scores at or above the Cut Score as 'not meeting standards' is unlikely to be adopted, whatever the scientific justification. This leaves the possibility of adopting systems which use prior information to reclassify testers who have Observed Scores below the Cut Score as meeting the standard.

While it may make intuitive sense to "give students credit" for other work and achievements in classifying them as proficient, there is the possibility that some journalists, business people, and parents (and educators who do not understand Classical Measurement Theory—which may be a majority) will see Bayesian reclassification as a lowering of standards. There are four approaches (which can be combined) to overcoming these objections:

1. Demonstrate that most students are really already proficient,
2. Demonstrate that most will eventually be classified as proficient,
3. In conjunction with either of the above, stress the ethical aspects of reclassification,
4. Demonstrate potential cost savings from reclassification.

The first approach, convincing people that most of the test takers are really proficient is the most technically compelling reason but also perhaps the hardest to accomplish. For many people, it would require a compelling illustrative example, like that given of the four students[38] in the last section of Chapter 3. For those technically inclined, a basic introduction to Classical Measurement Theory could be provided. Another approach would be to point out that in pilot studies of potential Bayesian reclassification—which must be done—students who would be reclassified form not proficient to proficient generally meet the requirements, either through retest or an alternative path. The third approach would be, after making one or both of the prior cases, to argue that it is immoral, given the high probability that the student does or will meet the requirements, to give them what may be a life long penalty—denial of a high school diploma, when the really deserve one.

The final argument is that Bayesian reclassification (from not meeting standards to meeting standards) will save school districts, the state, and ultimately the taxpayers' money. This can be used as a stand alone argument or as an argument in response to those who say given there are (in most states[39])

---

[38] The story had four students including one who hears answers during a radio news program on the way to the test and another who has just discovered she may be pregnant.

[39] "Examples of alternative paths for general education students (those who are neither special education students nor English language learners) include permitting students to meet the exam requirements by substituting scores on other tests like the SAT or ACT; taking a state—developed alternative assessment; pursuing a waiver or appeals process; receiving credit towards exam scores for satisfactory course grades; demonstrating competency by providing other evidence [the present author assume this includes portfolios]; and using various combinations of options.

alternative paths for a student to demonstrate competency, the Bayesian approach is not needed. However even if states continue to maintain alternative paths of demonstrating proficiency, a Bayesian approach can reduce the number of students subject to:

- Remediation,
- Retesting, and
- Alternative measures of proficiency (such as portfolios).

Figure 59 contains a histogram of the number of tests in each of the 25 states which require passing standardized tests for graduation, etc.. Given this data, it is reasonable to calculate the approximate number of students for whom the state would not have to have extraordinary costs for up to an average of 4 tests. Given about 15 million high school students, Figure 60 presents some orders of magnitude of the number of students who would not be required to go through alternative processes *if we assume 2% would be classified from not proficient to proficient by a Bayesian procedure.* This is not an unreasonable number, for it is half of NBA, which includes reclassification in both directions. Obviously, detailed research is needed, but the potential benefit, both in terms of cost to the taxpayer, to the students, and to society is great indicating the potential value of such research.

California, where policymakers have chosen not to allow alternatives for general educations students, is a notable exception (Center on Educational Policy, 2006. p 3.)

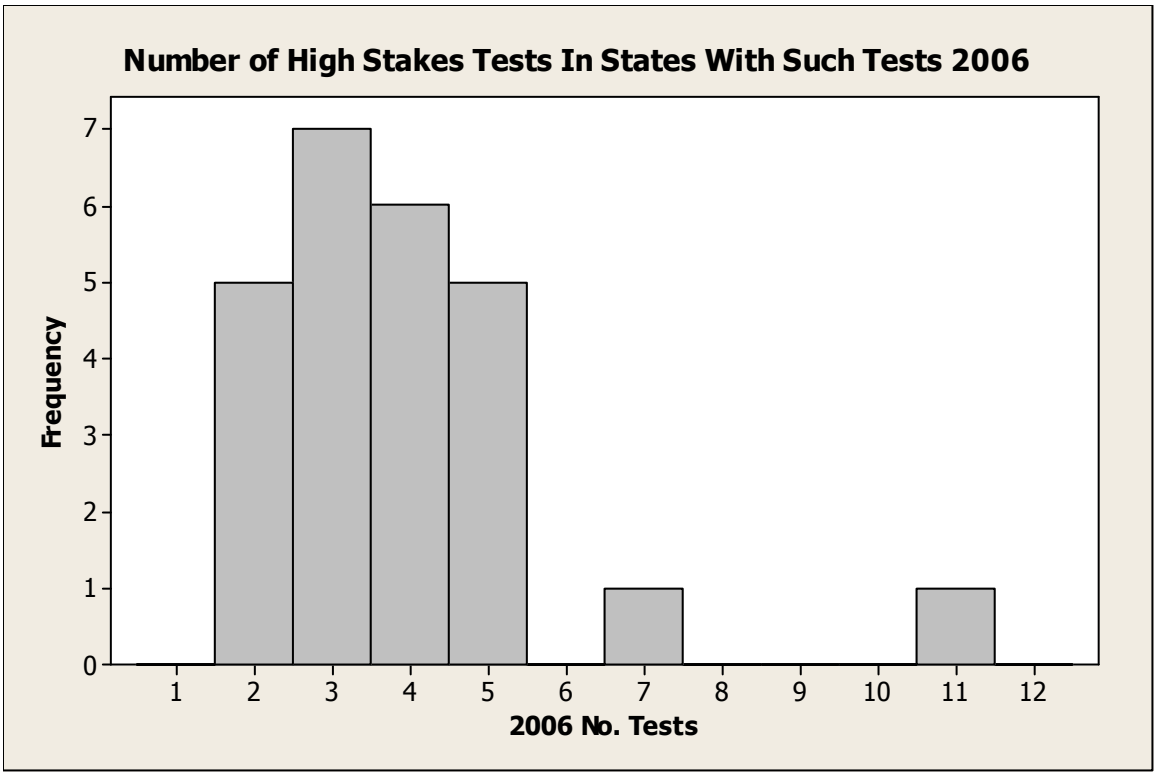


Figure 59: Number of high stakes tests in the 25 states with such tests, 2007 (Centre on Educational Policy, 2006, pp. 10),

	Number of High Stakes Tests			
	1	2	3	4
Total High School Students (Approximate)	15,000,000	15,000,000	15,000,000	15,000,000
Proportion of Students Subject ot High Stakes Tests	76%	76% #	76%	76%
Proportion First Time High Stakes Tests	25%	25% #	25%	25%
Number of First Time High Stakes Testers	2,850,000	2,850,000	2,850,000	2,850,000
Number of Tests	1	2	3	4
Approximate Percentage Reclassified to "Meet Standards"	2%	2%	2%	2%
Approximate Bayesian Reclassifications To "Meets Standards" Per Year	**57,000**	**114,000**	**171,000**	**228,000**

Figure 60: Illustration of order of magnitude of reclassifications from a Bayesian approach to reclassifying testers form not proficient to proficient only. Data of number of high school students, the world almanac and book of facts 2007, pp. 402-403.

A Minimalist Hope: "Psychometric Modesty"

The fact that large improvements in classification might be gained by a Bayesian approach should inspire what might be called "Psychometric Modesty," that is, that society (politicians, government officials, educational administrators, educators, parents, students and the public at large) should not claim for its standardized tests more power to discern the correct classification of test takers than these tests actually possess. It is ironic that this 'modesty' is inspired by a Bayesian approach as one of the arguments against Bayesians is the tolerance of outlandish prior probabilities—which could be seen by some as "epistemological immodesty." However, given there is the *possibility* for

substantial improvement in classification even with procedures that are modestly successful in generating Priors,[40] great care should be taken in evaluate the meaning of high stake test results. Indeed, while such tests may be useful for some purposes (for example, drawing attention to possible school systems or schools within systems in need of drastic improvement), using them for distributing societies profound rewards and punishments, particularly to America's youth, is a grave action that should only be taken with due care. It is minimally hoped that this paper, and the research programme of which it is the foundation, will contribute to a growing "Psychometric Modesty" where such high stakes tests are concerned.

[40] Where modestly successful means the priors can have relatively low accuracy and consistency (high Bias and SDBias).

APPENDIX A
MICHIGAN EDUCATIONAL ASSESSMENT PROGRAMS SCORES USED
Note: This is a Fax from the Michigan's MEAP Office

District: MICHIGAN DEPARTMENT OF EDUCATION
School: PUBLIC SCHOOLS
Codes: District- 99999
Run Date: 09/23/1999

Michigan Educational Assessment Program (MEAP)
High School Test: Mathematics (Form E)
Frequency Distribution Report
Grade 11 1st Time Testers
Spring 1999

	Scaled Score	Raw Score	Freq.
Level 4	214	0.0	4
	254	0.5	0
	293	1.0	5
	316	1.5	0
	333	2.0	9
	346	2.5	0
	358	3.0	17
	367	3.5	3
	375	4.0	33
	383	4.5	1
	390	5.0	56
	396	5.5	9
	402	6.0	131
	408	6.5	19
	413	7.0	173
	418	7.5	54
	422	8.0	264
	427	8.5	82
	431	9.0	331
	435	9.5	87
	439	10.0	461
	443	10.5	173
	446	11.0	635
	450	11.5	230
	453	12.0	686
	457	12.5	280
	460	13.0	757
	463	13.5	347
	467	14.0	838
	470	14.5	385
	473	15.0	1,018
	476	15.5	464
	479	16.0	1,101
	481	16.5	530
	484	17.0	1,128
	487	17.5	536
	490	18.0	1,233
	492	18.5	593
	495	19.0	1,362
	498	19.5	684
Level 3	501	20.0	1,322
	503	20.5	685
	506	21.0	1,389
	508	21.5	722
	511	22.0	1,495
	514	22.5	722
	516	23.0	1,469
	519	23.5	758
	521	24.0	1,584
	524	24.5	842
	526	25.0	1,574
	529	25.5	804
Level 2	532	26.0	1,609
	534	26.5	847
	537	27.0	1,653
	540	27.5	828
	542	28.0	1,637
	545	28.5	868
	548	29.0	1,643
	550	29.5	867
	553	30.0	1,726
	556	30.5	895
	559	31.0	1,690
	562	31.5	871
	565	32.0	1,705
	568	32.5	911
	571	33.0	1,694
	574	33.5	896
	577	34.0	1,636
	581	34.5	876
	584	35.0	1,669
	587	35.5	860
	591	36.0	1,554
	594	36.5	912
	598	37.0	1,537
	601	37.5	849
Level 1	605	38.0	1,429
	609	38.5	859
	613	39.0	1,425
	617	39.5	838
	621	40.0	1,287
	625	40.5	803
	630	41.0	1,248
	634	41.5	651
	639	42.0	1,066
	644	42.5	603
	649	43.0	991
	655	43.5	563
	660	44.0	847
	667	44.5	446
	674	45.0	703
	681	45.5	340
	690	46.0	556
	700	46.5	222
	713	47.0	375
	729	47.5	129
	752	48.0	287
	792	48.5	41
	835	49.0	119

Number included in summary: 74,146

A score of 605 through 835 is considered Level 1, Endorsed - Exceeded Michigan Standards.
A score of 532 through 601 is considered Level 2, Endorsed - Met Michigan Standards.
A score of 501 through 529 is considered Level 3, Endorsed - At Basic Level.
A score of 214 through 498 is considered Level 4, Not Endorsed.

APPENDIX B
MACRO FOR THE MAIN SIMULATION

```
GMACRO
DafRo06.mac

# master macro for all experiment runs for DAF's dissertation
Note  This Macro written September 14, 2003
Note
Note  This macro to generate observed scores,
Note  prior probabilities, and posteriors to perform one
Note  complete experimental run of DAF's dissertation

# EACH 'daf... macro' HAS DIFFERENT INPUT FOR
# K50, k30, k1, k2, k4-k8

#What is the Random Order
Let k50 = 6
#What is the Experiment Designator: Random Order . Reliability  and cut
score
Let k30 = 6.89226

# What are the Reliability (k1), SEM (k2), and Cutscore (k4)

Let k1 = 0.892
Let k2 = 3.1798581
Let k4 = 26

# What are the Logistic Regression Constant (K7) and Coefficient (K8)
# These are a function of the Reliability (and thus SEM) and the Cut
Score

Let k7 = -14.2156
Let k8 = 0.551401

# What are the Bias (k5) and SD Bias(k6)?
# These are aspects of the Bayesian estimates of the prior

Let k5 = 0
Let k6 = 4.083

# THIS COMPLETES THE UNIQUE INPUT FOR EACH MACRO
#How many lines (number of Estimated True Scores)
# 1038044 is 14 repetitions of the 74,146 true scores

Let k3 = 1038044
```

```
Name k1 = 'Reliability'
Name k2 = 'SEM'     # Standard Error of Measurement
Name k3 = 'lines'   # Number of Lines (number of Estimated True Scores)
Name k4 = 'cut'     # Cut Score

Name K5 = 'Bias'    # Bias of the Bayesian Estimate  (in percentage
points)
Name k6 = 'SDBias'  # Standard Deviation of the Bias (in percentage
points)

Name k7 = 'a'       # Constant from the Logistic Regression
Name k8 = 'b'       # Coefficient from the Logistic Regression

Name k10 = 'belo-cut'   # The number for categorization below the
custscore

Name k15 = 'check_no'   # [Square Root of([1-'Reliability'])]*9.676
# Where 9.676 is the standard deviation of the 74,146 Observed scores
# in the SPRING 1999 MATH PORTION OF THE HIGH SCHOOL MEAP


Name k25 = 'Const-LgstRegres'
Name k26 = 'Coeff-LgstRegres'
Name k30 = 'RDMord.RELcut'
Name k50 = 'Random_Order'

# C1 is ETS, the Estimated True Scores.  These have been calculated
previously
# taking 74,146 observed scores, calculating the deviation score by
subtracting # the average, attenuating using the reliability
coefficient, and then
# adding back the average

Name c2  = 'TM-Code' # Code for whether the True Score is Above the Cut
Score-                                  # TM (True Meets the Standard) or
not (tf)[True Fails to
                     # Meet the Standard]

Name c3  = 'e-O'     # error to add to ETS to get the observed
Name c4  = 'OO'      # Origional Observed (before trimming or rounding)
Name c5  = 'obs1'    # trim scores below 0 or above 49 to 0 and 49
respectively

Name c18 = 'DcmlPt obs1'   # The decimal part of obs1
Name c19 = 'DcmlPt Rndd'   # Round the decimal part to 0, 0.5, or 1.0
# Note: each gets about 1/3, producing digit preference distribution

Name c20 = 'obs'     # the observed score used in further calculations
(floor
                     # plus c19 in the formula)

Name c6  = 'OM-Code' # Code for whether the Observed Score is Above the
Cut
                     # Score-  OM (Observed Meets the Standard) or not
                     # (of)[Observed Fails      to Meet the Standard]
```

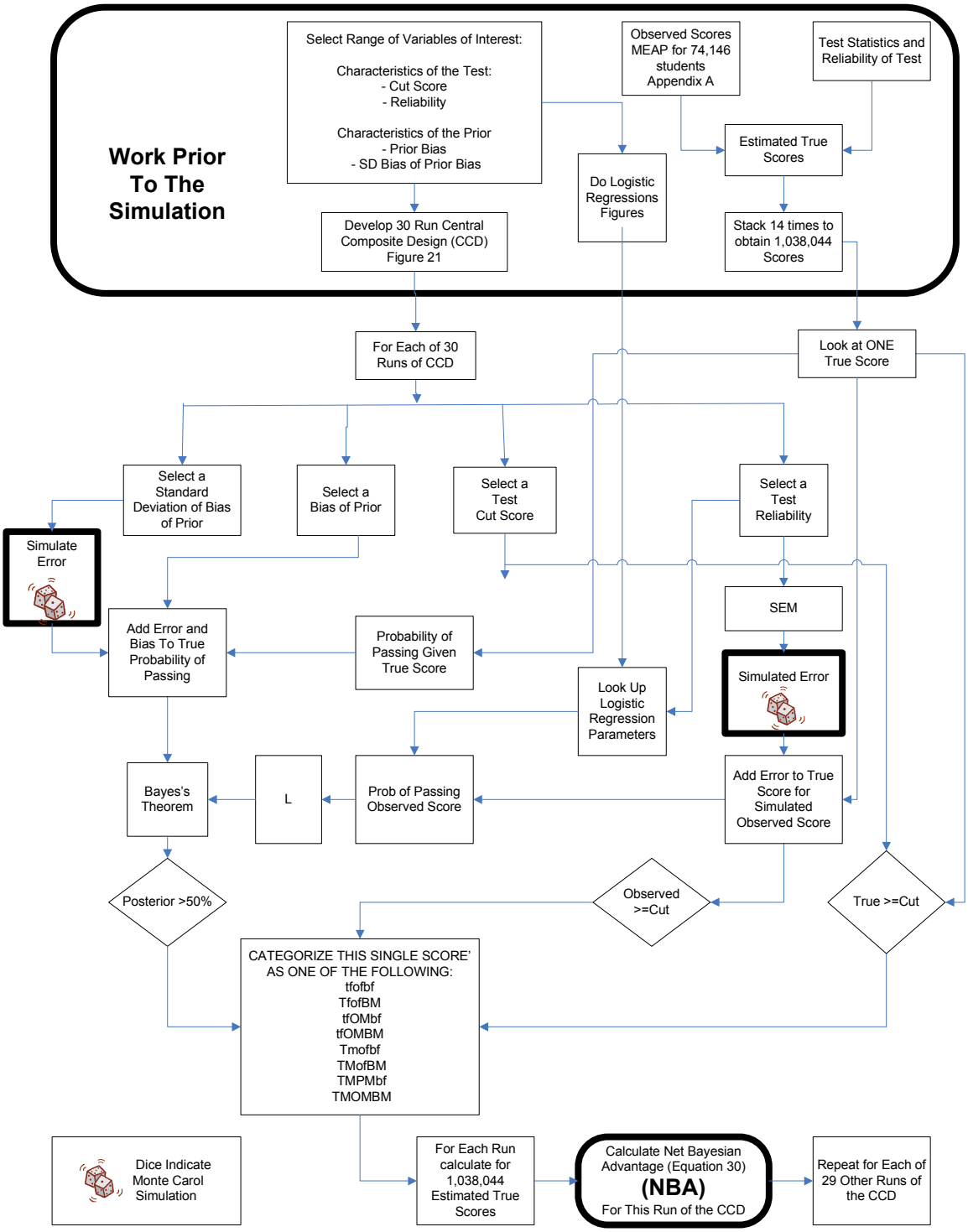
```
Name c7  = 'T-Prob'   # True probability:  P(Observed>Cut | ETS, SEM)
                      # Uses the ETS as point up to which the cumulative
goes
                      # Taking advantage of the fact that this is equal
to
                      # P(O<T | Cut, SEM)




Name c8  = 'e-bv-t-P' # error for Bayesian variability to add to true
                           # Probability (Based on Bias and SDBias)
Name c9  = 'ibe'      # initial Bayesian estimate [can be less 0, more
than 1]
Name c10 = 'Prior'    # Between 0.005 and 0.995
Name c12 = 'Dat Prob' # Probability of observed score being from a
student                           # with true > cut based on logistic
regression
                      # coefficients
Name c13 = 'Posterior'
Name c14 = 'BM-Code'  # Code for whether the Posterior is >= 0.5
                           # BM (Bayesian Meets the Standard) or <0.5
(bf)[True Fails                   # to Meet the Standard]


Name c15 = 'Result'   # Concatenate TM-Code, OM-Code, and BM-Code

Name C22 = 'Results'
Name C23 = 'Result Cnt'
Name c24 = 'Result Pct'


Let k10 = k4 - 0.000001
Let k15 = [SQRT([1-'Reliability'])]*9.676
Let k25 = k7
Let k26 = k8

# Following commands generate the Observed score from the ETS (Estimated
True    # Score) in C1

Random 'lines' C3;
Normal 0 'SEM'.

Let c4 = c1 + c3

Code (-100000:0) 0 (49:100000) 49  'OO' c5


Let c18 = c5-FLOOR(c5,0)
Code (-10000:0.333333333333) 0 (0.3333333333334:0.6666666666666) 0.5 &
 (0.6666666666667:10000) 1  C18 c19
Let c20 = FLOOR(c5,0)+C19

# Following Commands generate the Priors

CDF C1 c7;
Normal 'cut' 'SEM'.
```

```
Random 'lines' c8;
Normal 'Bias' 'SDBias'.
Let c9 = ('e-bv-t-P'/100)+'T-Prob'
Code (-10000:0.005) 0.005 (0.995:1000) 0.995  'ibe'  'Prior'

# Following commands generate the data probabilities and the posteriors
Note
Note  Macro is computing posterior
Note

Let c12 = 1/[1+E()**(-1*['a'+'b'*'obs'])]
Let 'Posterior' = ['Prior'/(1-'Prior')]/(['Prior'/(1-'Prior')] + &
[(1-'Dat Prob')/'Dat Prob'])

# The following lines code the True, Observed, and Bayesian and
# Concatenate them to the result

Code (-100000: k10) "tf" (k4:100000) "TM" C1 c2

# This is just a note to let me see the program is running well
Note
Note  Macro has completed TM-Code
Note

Code (-100000: k10) "of" (k4:100000) "OM"  'obs' c6

Code (-100000:0.4999999999) "bf" (0.5:1) "BM" 'Posterior'  'BM-Code'

Concatenate 'TM-Code' 'OM-Code' 'BM-Code' 'Result'

Note The following are the results for
Print k50, k30

Desc c1, c3-c5, c7-c10, c12-c13, c18-c20

Tally 'Result';
Counts;
Percents;
Store c22, c23, c24.

Note THE ABOVE TALLY IS FOR THE FOLLOWING INPUTS
Print K50, K30, k1, k2, k15, k3, k4, k5, k6, k25, k26


ENDMACRO
```

APPENDIX C
MACRO FOR LOGISTIC REGRESSION

```
MTB > Random 1038044   'fouth e';
SUBC>   Normal 0.0 3.1798581.
MTB > Let 'L4-89-26' = 'ETS'+'fouth e'
MTB > Code (-1000:0) 0 (49:10000) 49  'L4-89-26'  'L4t8926'
MTB > Describe  'ETS' 'D26' 'fouth e' 'L4-89-26' 'L4t8926'.


MTB > Name c6 = 'EPRO1'
MTB > BLogistic 'D26' = L4t8926;
SUBC>   Logit;
SUBC>   Eprobability 'EPRO1';
SUBC>   Brief 2.
```

APPENDIX D
DETAILED FLOWCHART OF RESEARCH PROCESS



REFERENCES

Achinstein, Peter (ed.) (2004) *Science rules: a historical introduction to scientific methods.* Johns Hopkins University Press. Baltimore, Maryland.

Albert, J. H. (1996). *Bayesian computation using Minitab.* Belmont, California: Duxbury Press.

American Statistician (1997) 51(2)

American Psychological Association (2001). *Publication manual of the American psychological association, Fifth Edition.* American Psychological Association. Washington, DC.

American Psychological Association (2005). *Concise rules of APA style.* American Psychological Association. Washington, DC.

Audi, R. (1998). *Epistemology: A contemporary introduction to the theory of knowledge.* London: Routledge.

Antelman, G. (1997). *Elementary Bayesian statistics.* (A. Madansky & R. McCulloch, Eds.). Cheltenham, UK: Edward Elgar.

Aquinas, T. *Suma Theologicia.*

Azzouni, Jody (2004). *Knowledge and reference in empirical science.* Routledge/Taylor and Francis Group. London and New York.

Bayes, T. (1940). An essay towards solving a problem in the doctrine of chances. In Deming, W. E. (Ed.). *Facsimiles of two papers by Bayes.* Washington D.C:

(Original published in 1673).

Berry, D. A. (1996) *Statistics: A Bayesian perspective.* Belmont, California: Duxbury Press.

Blair, J. (1998, October 21). Accountability measure kicks in: Pa. doles out $10 million to reward schools. *Education Week.* p. 5.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP [expected a posteriori] estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431-444.

Bolstad, William M. (2004) Introduction *to Bayesian statistics*. A. john Wiley & Sons, Inc.. Hoboken, New Jersey.

Box, G.E.P., Hunter, W.G., & Hunter, J. S. (1978). *Statistics for experimenters.* New York: Wiley.

Box, G.E.P., & Draper, N. R. (1987). *Empirical model building and response surfaces.* New York. Wiley.

Braun, Henry (2006). Empirical bayes. Chapter 14 in Green, Judith L., Camilli, Gregroy, and Elmore, Patricia B. 2006). *Handbook of complementary methods in education research.* American Educational Research Association. Washington, D.C. and Lawarece Erlbaum Associates, Publishers, Mahwah, New Jersey.

Brennan, Robert L. (ed) (2006). *Educational measurement, fourth edition.* American Council of Education / Prager. Publishers Westport, CT.

Center on Educational Policy (August, 2006). State High School Exit Exams: A Challenging Year. Center On Educational Policy. Washington, DC.

Chalmers, A. F. (1995). *What is this thing called science?: An assessment of the nature and status of science and its methods.* Indianapolis: Hacket Publishing Company.

Chow, S. L. (1996*). Statistical significance: Rationale, validity and utility.* London: Sage Publications.

Cotterell, Arthur and Storm, Rachael (2006). The ultimate encyclopedia of mythology. Anness Publishing Ltd., Hermes House, London, England.

Curd, M. & Cover, J. A. (1998). Commentary. In Curd, M, & Cover, J. A. (Eds.), *Philosophy of science: The central issues.* (pp. 627-674). New York: W. W. Norton.

Curd, M, & Cover, J. A. (Eds.). (1998*). Philosophy of science: The central issues.* New York. W. W. Norton.

David, F. N. (1998). *Games, gods, and gambling: A history of probability and statistical ideas.* New York: Dover.

Dwyer, Carol Anne, ed. (2005). *Measurement and research in the accountability era.* Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

Everitt, B. S. (2002). The Cambridge dictionary of statistics, second Edition. Cambridge University Press. Cambridge, UK.

de Finetti, Bruno (1980). Foresight: Its logical laws, its subjective sources. Kyburg, H. E., Jr., & Smokler H. E. (Eds.). *Studies in subjective probability.* (Rev. ed.). (pp. 193-224). (I. McGilvary, Trans). Huntington, New York: Robert E. Krieger Publishing Company. (Original published in 1976)

de Finetti, Bruno. (1980) Probability: Beware of falsifications. In Kyburg, H. E., Jr., & Smokler H. E. (Eds.). *Studies in subjective probability.* (Rev. ed.). (pp. 134-174). (I. McGilvary, Trans). Huntington, New York: Robert E. Krieger Publishing Company.

Deming, W. E. (1940). Introduction. In Deming, W. E. (Ed.). Facsimiles of Two Papers by Bayes. Washington D.C.

Dewey, John (1997). *Experience and education.* New York: MacMillian Publishing Company.

Daston, L. (1988). *Classical probability in the enlightenment.* Princeton: Princeton University Press.

DuMouchel, William H. (1992) . Introduction to Edwards, Lindman, and Savage (1963) Bayesian statistical inference for psychological research. In In Koltz, S. & Johnson N. L. (Eds.). *Breakthroughs in Statistics Volume I: Foundations and Basic Theory.* (pp. 517-530). New York: Springer-Verlag.

Earman, J. (1992). *Bayes or bust?: A critical examination of Bayesian confirmation theory.* Cambridge, Mass: A Bradford Book. The MIT Press.

Edwards, W., Lindman, H. & Savage L. J. (1992, originally published 1963). Bayesian statistical inference for psychological research. In Koltz, S. & Johnson N. L. (Eds.). *Breakthroughs in Statistics Volume I: Foundations and Basic Theory.* (pp. 531-578). New York: Springer-Verlag.

Elmore, P. B. & Woehlke P. L. (1996, April). *Research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1995.* Paper presented at the annual meeting of the American Educational Research Association (ERIC Document Reproduction Service No. ED 397122).

Fetzer, J. H. and Almeder, R. F. (Eds.) (1993). *Glossary of epistemology/philosophy of science.* New York: Paragon House.

Firestone, William A., Schorr, Roberta Y., and Monfils, Lora Frances, eds. (2004). *The ambiguity of teaching to the test: standards, assessment, and educational reform.* Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

FDA (Food and Drug Administration) (2007) Draft Guidance for Industry and FDA Staff: guidance for the use of Bayesian statistics in medical device clinical trials. U. S Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Division of Biostatistics, Office of Surveillance and Biometrics. Draft Guidance May 23, 2006. Rockville, Md.

Galison, P. & Stump D. J. (1996). *The Disunity of Science: Boundaries, Contests, and Power.* Stamford: Stamford University Press.

Gill, Jeff (2002). *Bayesian methods for the social and behavioral sciences.* Chapman & Hall/CRC. Boca Raton, Florida.

Gonick, Larry and Smith, Wollcott (1993). *The cartoon guide to statistics. HarperPerrennial.* New York. New York.

Good, I. J. (1993). Rational decisions. In Koltz, S. & Johnson, N. L. (Eds.). *Breakthroughs in statistics Volume I: Foundations and basic theory.* (pp. 365-377.) New York: Springer-Verlag.

Good, I. J. (1980). Subjective probability as the measure of a non-measurable set. In Kyburg, H. E., Jr., & Smokler H. E. (Eds.). *Studies in subjective probability.* (Rev. ed.). (pp. 133-146). Huntington, New York: Robert E. Krieger Publishing Company. (Original published in 1962)

Gould, Stephen Jay (1996). *The mismeasure of man (revised and expanded).* W. W. Norton & Company. New York.

Green, Judith L., Camilli, Gregroy, and Elmore, Patricia B. 2006). *Handbook of complementary methods in education research.* American Educational Research Association. Washington, D.C. and Lawarece Erlbaum Associates, Publishers, Mahwah, New Jersey.

Gymour, C. (1998). Why I am not a Bayesian. In Curd, M, & Cover, J. A. (Eds.). *Philosophy of science: The central issues.* (pp. 584-606). New York. W. W. Norton.

Gutherie, Will (October, 2006). "Hands-on Bayesian data analysis using Winbugs." ASA/ASQ Fall Technical Conference. Columbus, Ohio.

Hacking, Ian (2001). *An introduction to probability and inductive logic.* Cambridge University Press. Cambridge, UK.

Hacking, I. (1975). *The emergence of probability.* Cambridge, England: Cambridge University Press.

Hacking, I. (1997). Experimentation and Scientific Realism. In Tauber, A. I. (Ed.). *Science and the quest for reality.* (pp. 162-181). New York: New York University Press.

Hempel, Carl G. (1966) Philosophy *of natural science.* Prentice Hall. Upper Saddle River, New Jersey.

Hertz, David B. (1964, reprinted 1979). Risk analysis in capital investment. *Harvard Business Review* (reprint No. 79505). September-October 1979. pp 169-180.

Hertz, David B. (1968). Investment policies that pay off. *Harvard Business Review* (reprint No. 68107). January-February, 1968. pp. 96-107.

Helseth, T.J., et al, Design-Expert ®, Version 6 for Windows, Stat-Ease, Inc, Minneapolis, 2000, web site: www.statease.com.

Haig, B. D. (1996). Statistical methods in education and psychology: A critical perspective. *Australian Journal of Education 40(2),* 190-229.

Hoerl, Roger and Snee, Ronald (2001) *Statistical thinking: improving business. Performance.* Duxbury Press California.

Hoting, Jennifer A. (2004). Statistics 675—Bayesian statistics. Department of Statistics, Colorado State University. Colorado.

Howson, C. & Urbach P. (1993). *Scientific reasoning: the Bayesian approach.* (2nd ed.). Chicago: Open Court.

Horwich, P. (1998). Wittgensteinian bayesianism. In Curd, M. & Cover, J. A. Commentary. In Curd, M, & Cover, J. A. (Eds.). *Philosophy of science: The central issues.* (pp. . pp. 607-623.). New York: W. W. Norton.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics 23(1), 35-56.*

Iversen, G. R. (1984). *Bayesian statistical inference.* Newberry Park: Sage.

Jeffery, Richard C. (1980) Probable Knowledge. In Kyburg, H. E., Jr., & Smokler H. E. (Eds.). *Studies in subjective probability.* (Rev. ed.). (pp. 225-238). Huntington, New York: Robert E. Krieger Publishing Company. (Original work published 1968)

James, William. *Pragmatism*. (1981). (B. Kuklick, Ed.). Indianapolis: Hacket Publishing Company.

Kasser, Jeffrey L. (2006a*). Philosophy of science.* (CDs) The Teaching Company. Chantilly, Virginia.

Kasser, Jeffrey L. (2006b). *Philosophy of science, course guidebook.* The Teaching Company. Chantilly, Virginia.

Khuri, A. I. & Cornell, J. A. (1987) *Response surfaces: Design and analysis.* New York: Marcel Decker, Inc.

Kleinbaum, David G. (1994). *Logistic regression: a self learning text.* Springer-Verlag New York, Inc. New York, New York.

Kolen, Micael J. (2006). Scaling and norming, chapter 5 in Brennan (2006). Pp. 155-186.

Koltz, S. & Johnson, N. L. (Eds). (1992). *Breakthroughs in statistics volume I: foundations and basic theory.* New York: Springer-Verlag.

Kruger, L., Daston L. J., & Heldelberger, M. (Eds.). (1987). *The probabilistic revolution. Vol. I, ideas in history.* Cambridge, Mass: MIT Press.

Kuhn, T. S. The road since *structure*. In Tauber, A. I. (Ed.). (1997). *Science and the quest for reality.* (pp. 230-245). New York: New York University Press.

Kyburg, H. E., Jr., & Smokler H. E. (Eds.). (1980). *Studies in subjective probability.* (Rev. ed.). Huntington, New York: Robert E. Krieger Publishing Company.

Ladson-Billings, Gloria and Tate, William F. (eds.) (2006) *Education research in the public interest: social justice, action, and policy.* TEACHERS College Press, Teachers, Colllege, Columbia University. New York. New York.

Laplace, P.-S. (1951). *A philosophical essay on probabilities.* (Bell, E. T., Ed.). (Truscott, F. W. & Emory, F. L., Trans.). New York: Dover. (Original work published 1814).

Lakatos, I. Falsification and the methodology of scientific research programmes. In Lakatos, I. & Musgrave, A. E. (Eds.) (1970). *Criticism and the growth of knowledge: proceedings of the colloquium in the philosophy of science, London 1965.* Cambridge: Cambridge University Press.

Joftus, S. & Whitney, T. (1998, November 18). High standards without big lawsuits. *Education Week.* pp. 28, 32.

Matus, Ron (2007). How good is teacher? Bonus plan may tell: soon, we'll know which teachers get the rewards. But what ill parents do with that knowledge? *Tampa Bay Times.* February 4, 2007.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Manzo, K. K. (1998, November 4). How to Gauge Accountability Providing Ticklish for N. Carolina. *Education Week.* p.1, 26.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* (105), 156-166.

Michigan Department of Education (1998). Webpage.

Minitab® Statistical Software Version 13, Minitab, Inc. College Park, Pennsylvania.

Mises, R. von. (1981). *Probability, Statistics, and Truth.* New York. Dover. (Original work published 1957).

Mislevy, R. J. (1993). Some Formulas for use with Bayesian ability estimates. (ERIC Document Reproduction Service No. ED 384664)

Mislevy, R. J. (1994) Evidence and Inference in Educational Assessment. *Psychometrika*, 59(4), 439-483.

Moore, David S. (2001*). Statistics: Concepts and controversies. fifth edition.* W.H. Freeman and Company. New York. New York.

Moore, David S. and McCabe, George P. (1993). Introduction *to the practice of statistic, second edition.* W. H. Freeman and Company. New York. New York.

Montogmery, D. C. (1997) *Design and analysis of experiments.* (Fourth Ed.). New York: John Wiley & Sons.

Morton, Adam (2003). *A guide through the theory of knowledge, third edition.* Blackwell Publishing. Madden, MA.

Motulsky, Harvey (1995). Intuitive biostatistics. Oxford University Press. New York, New York.

Nickerson, Raymond S. (2004). *Cognition and chance: the psychology of probabilistic reasoning.* Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

Nunnally, J. C. & Bernstein, I. H. (1994) *Psychometric Theory* (Third Ed.) New York: McGraw-Hill, Inc.

Plson, Lynn (2007). School accountability system seen as unlikely to face major overall. *Education Week.* January 31, 2007.

Okasha, Samir (2002). *Philosophy of science: a very short introduction.* Oxford University Press. Oxford, UK.

Palisade Corporation (2002). @Risk. Palisade Corporation, Newfield, NJ.

Phelps, Richard P. (ed) (2005). Defending *standardized testing.* Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

Phelps, X. (1998, November 12). Standardized tests. *Education Week.* p. 30.

Plato, J. von. (1994). *Creating modern probability.* Cambridge: Cambridge University Press. 1994.

Popper, K. The Problem of Induction. (1998). In Curd, M, & Cover, J. A. (Eds.), *Philosophy of science: The central issues.* (pp. 426-443). New York: W. W. Norton.

Porter, T. M. (1988). *The rise of statistical thinking, 1830-1900.* Princeton: Princeton University Press.

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

Pojman, L. P. (1993). *The theories of knowledge: Classic & contemporary readings.* Belmont, California: Wadsworth Publishing Company.

Ramsey, F. P. (1980). Truth and Probability. In Kyburg, H. E., Jr., & Smokler H. E. (Eds.). *Studies in subjective probability.* (Rev. ed.). (pp. 25-52). Huntington, New York: Robert E. Krieger Publishing Company*. (Original work published1926)*

Reichenbach, H. (1995) The search for certainty. In Westphal, J. (Ed.). *Certainty.* (pp. 104-120). Indianapolis: Hacket Publishing Company, Inc. (Original work published 1951)

Robbins, Herbert E. An empirical Bayes approach to statistics. In Koltz, S. & Johnson N. L. (Eds). *Breakthroughs in statistics Volume I: Foundations and basic theory.* (pp. 388-394). New York: Springer-Verlag.

Roberts, Christian P. and Casella, George (2004). *Monte Carlo statistical methods, second edition.* Springer Science+Business Media, LLC. New York, New York.

Rossi, Peter E., Allenby, Greg M., McCulloch, Robert (2005*). Bayesian statistics and marketing.* John Wiley and Sons, Ltd. Crichester, West Sussex, England.

Salmon, W. C. (1993). Probabilistic Causality. In Sosa, E. & Tooley, M. (Eds.). *Causation.* (pp. 137-153).Oxford: Oxford University Press.

Salmon, W. C. Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In Curd, M, & Cover, J. A. (Eds.). *Philosophy of science: The central issues.* (pp. . 551-583). New York: W. W. Norton.

Salmon, W. C. (1996). *The foundations of scientific inference.* Pittsburgh, Pa: University of Pittsburgh Press.

Savage, L. J. (1972). *The Foundations of Statistics.* Second Revised Edition. New York: Dover Publications, Inc. 1972. (Original work published 1954)

Sawilowsky, S. Nonparametric tests of interaction in experimental design. *Review of Educational Research* (60)1, 91-126.

Shadham, X. (1998, November 14). Accountability. *Education Week.* p 82.

Shewhart, W. A. (1986). *Statistical methods from the viewpoint of quality control.* (W. E. Deming, Ed.). New York: Dover. (originally published in 1939).

Silberman, Todd (2007). Higher bar to graduates Hill trip up many students. *The News & Observer*. Florida.

Sosa, E. & Tooley, M. (Eds.). (1993). *Causation.* Oxford: Oxford University Press.

Sivia, D.S. (1996). Data analysis: A Bayesian tutorial. Oxford: Oxford University Press.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, Mass: Harvard University Press.

Stigler, Stephen M. (1999). *Statistics on the table: the history of statistical concepts and methods.* Harvard University Press. Cambridge, Massachusetts.

Thomas, R. Murray (2005). High-stakes testing: coping with collateral damage. Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

Tonn, Jessica L. (2007). Huston in uproar over teachers' bonuses: many highly valued teachers overlooked in test-driven system. *Education Week.* February 1, 2007.

Tauber, A. I. (Ed.). (1997). *Science and the quest for reality.* New York: New York University Press.

Turner, Tina. (artist) and Britten, Terry & Lyle, Graham (writers) and Britten, Terry (producer) (1984). "What's Love Got to Do with It?" From the album *Private dancer:* Los Angeles: Capital Records.

Yen, Wendy M. and Fitzpatrick, Anne R. (2006). "Item response theory, chapter 4. In Brennan, Robert L. (ed) (2006). *Educational measurement, fourth edition.* American Council of Education / Prager. Publishers Westport, CT. pp. 111-153.

Westphal, J. (Ed.) (1995). *Certainty.* Indianapolis: Hacket Publishing Company, Inc.

Winston, Wayne L. (2001). *Simulation modeling using @Risk: updated for version 4.* Duxbury/Thompson Learning. Pacific Grove, California.

World Almanac Books (2007). *The world almanac and book of facts.* World Almanac Books. New York, New York.

ABSTRACT

SIMULATION BASED SPECIFICATIONS FOR EVALUATING HIGH STAKES EDUCATIONAL TEST RESULTS FROM A BAYESIAN EPISTEMOLOGICAL PERSPECTIVE: PHILOSOPHICAL FOUNDATIONS, RESPONSE SURFACE INVESTIGATIONS, AND PRAGMATIC APPLICATIONS OF RESULTING PSYCHOMETRIC MODESTY

by

DAVID ARTHUR FLUHARTY

August 2007

Advisor: Dr. Shlomo S. Sawilowsky, Ph.D.

Major: Educational Evaluation and Research

Degree: Doctor of Philosophy

This dissertation explores the proposition that Bayesianism and High Stakes Test Results can be mutually illuminating. It employs a simple form of Bayes's Theorem, assumes Classical Measurement Theory, and uses classical (as opposed to Bayesian) statistical tools such as Monte Carlo Simulation, Logistic Regression, and Response Surface Methodology.

Generating Estimated True Scores of approximately 75,000 first time test takers from the Observed Scores from one administration of an actual High Stakes Test (Michigan MEAP Mathematics), it was found that a Bayesian approach could result in a net improvement in classification (meeting standard or not meeting standard) of between approximately 2-½ % to 5-½ % for a wide range Accuracy and Consistency of the Bayesian Prior as well as a wide range

of Test Reliability and Cut Scores. It was found that the factor which has the most impact was Test Reliability—the less reliable the test, the more a Bayesian approach improves classification. Thus, this dissertation provides a 'proof of concept' that an intellectual technology for Bayesian reclassification *might* be developed to improve the classification of students—and the distribution of high stakes consequences (graduation, passing to next grade, promotions of teachers, etc.) which such classifications entail.

Before implementation of a policy using this approach, considerable research is needed. First, it must be determined if the results hold for a wide range of distributions of test scores (a task which is part of this research programme). Second, it must be determined if students who might be reclassified will have better education outcomes. While it must be determined under what condition such a Bayesian reclassification would be politically acceptable, at minimum, it is hoped that this dissertation will contribute to increasing "Psychometric Modesty" at a time when profound societal rewards and punishments are associated with High Stakes Tests.

AUTOBIOGRAPHICAL STATEMENT
DAVID ARTHUR FLUHARTY

David Fluharty was born on February 28, 1951. An only child, his youth was spent in New Cumberland, Weirton, and Follansbee which are in what was then a thriving steel mill area of West Virginia's northern panhandle. His father, Ralph Fluharty (1917-1979), was a steel worker (roll turner/lathe operator/tool maker) and his mother, Grace Elaine Martin Fluharty (1915-1977) was a stay at home mother and later a sales clerk (although her position would now be called assistant manager of a woman's boutique). After attending local Catholic grade schools, he went to St. Joseph Preparatory Seminary in Vienna, West Virginia, a boarding school. His first two years at Wheeling College (now Wheeling Jesuit University) were also spent studying to be a priest. His generation was the first to attend college. He spent three summers during college as a laborer in Weirton Steel, including the coke plant, mason gang, and the hot mill (where his last position was Second Sweeper).

He completed his BA in Political Science at in three years Wheeling in 1972, the year he became an agnostic (returning to Catholicism in 1999). He went on to the University of Chicago, first in the Committee on International Relations (MA completed in 1978) and then the Graduate School of Business (MBA in Business Economics and Finance in 1975). It was at Chicago where his deep interest in statistics developed from efforts to understand quantitative models in economics, and international & nuclear strategy. He met his future wife, Mary Reiter of Pittsburgh, in the television room of their dormitory, International House. He held summer internships in the Chicago Office of Department of Housing and Urban Development and the Washington DC office of Senator Jennings Randolph (D-WV).

After spending one year helping to create capital market distortions as a Bond Guarantee Analyst at Maritime Administration in the US Department of Commerce in Washington DC, Fluharty joined Ford Motor Company in Dearborn Michigan as a Financial Analyst in 1977. Positions in Car Product Development Controller's Office included the path breaking "Team Taurus." In 1978 he married Mary Reiter. One day at Ford he saw a sign asking "Are you interested in statistics?" This led to membership in Ford's Statistical Methods Council, exposure to W. Edwards Deming, and a Graduate Certificate in Applied Statistics from Oakland University. In 1985 he moved to a Ford statistics group in Ford. In 1986 he and Mary became the proud parents of Margaret Rose Elaine Fluharty-Reiter, who would attend Waldorf nursery school, a progressive Kindergarten, and a Montessori elementary school. In 1988 Fluharty left Ford for what would become Alcoa Fujikura, Ltd. (AFL), where in worked as an individual contributor and manager in quality, warranty, and finance.

In 1995 Margaret died unexpectedly. Fluharty began studies in Education at Wayne State that year. In 2001 his position was eliminated at AFL. He spent two years at Continental Teves in Auburn Hills, Michigan. In 2004 he worked for two months at the American Statistical Association. In 2004 he joined Remy International in Anderson, Indiana, where he now is a Senior Statistical Analyst.

Fluharty participated in the Internship in Ignatian Spirituality at Manressa in Bloomfield Hills, Michigan. He has served in a variety of volunteer capacities in the statistics profession. He volunteered for the Jubilee 2000 (developing nation) debt relief effort. His biography is in *Who's Who in America*. His hobbies are reading, visiting art and other museums, and exploring the connection between ideas & action. He and his wife Mary enjoy dining, watching movies & the History Channel.