HUMAN DETECTION OF COMPUTER SIMULATION MISTAKES IN ENGINEERING EXPERIMENTS

by

TROY BRENDON SAVOIE

SUBMITTED TO THE DEPARTMENT OF MECHANICAL ENGINEERING
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MECHANICAL ENGINEERING

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Signature of Author . . . . . .
Department of Mechanical Engineering
May 18, 2010

Certified by . . . . . .
Daniel D. Frey
Associate Professor of Mechanical Engineering and Engineering Systems
Thesis Supervisor

Certified by . . . . . .
Brenan C. McCarragher
Charles Stark Draper Laboratory
Thesis Supervisor

Accepted by . . . . . .
David E. Hardt
Ralph E. and Eloise F. Cross Professor of Mechanical Engineering
Chairman, Department Committee on Graduate Theses

Human Detection of Computer Simulation Mistakes
in Engineering Experiments

by

Troy Brendon Savoie

Submitted to the Department of Mechanical Engineering
on May 18, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mechanical Engineering

Abstract

This thesis investigates the notion that the more complex the experimental plan, the less likely an engineer is to discover a simulation mistake in a computer-based experiment. The author used an *in vitro* methodology to conduct an experiment with 54 engineers completing a design task to find the optimal configuration for a device with seven two-level control factors. Participants worked individually using a prescribed design approach dependent upon the randomly assigned experimental condition – an adaptive one-factor-at-a-time plan for the control group or a resolution III fractional factorial plan for the treatment group – with a flawed computer simulation of the device.

A domain knowledge score was measured by quiz, and success or failure in discovering the flaw was measured by questioning during debriefing. About half (14 of 27) of the participants using the one-factor-at-a-time plan discovered the flaw, while nearly none (1 of 27) using the fractional factorial plan did so. Logistic regression analysis of the dichotomous outcome on treatment condition and domain knowledge score showed that flaw detection ability improved with increased domain knowledge, but that an advantage of two standard deviations in domain knowledge was insufficient to overcome the disadvantage of using the fractional factorial plan.

Participant reactions to simulation results were judged by two independent raters for surprise as an indicator of expectation violation. Contingency analysis of the surprise rating results showed that participants using the fractional factorial plan were significantly less likely (risk ratio ≈ 0.57) to appear surprised when the anomaly was elicited, but there was no difference in tendency to display surprise otherwise.

The observed phenomenon has ramifications beyond simulation mistake detection. Cognitive psychologists have shown that the most effective way to learn a new concept is to observe unexpected behavior, investigate the cause, then integrate the new concept into one's mental model. If using a complex experimental plan hinders an engineer's ability to recognize anomalous data, the engineer risks losing opportunities to develop expertise. Initial screening and sensitivity analyses are recommended as countermeasures when using complex experiments, but more study is needed for verification.

Thesis Supervisor: Daniel D. Frey
Title: Associate Professor of Mechanical Engineering and Engineering Systems

Thesis Supervisor: Brenan C. McCarragher
Title: Associate Director, Strategic Systems, Charles Stark Draper Laboratory

This page intentionally left blank.

Dedicated to the memory of my grandparents:

Ashton & Mabel Savoie Kenneth & Merle Elder

This page intentionally left blank.

Those who bring sunshine into the lives of others cannot keep it from themselves.

James M. Barrie

Acknowledgments

It is amazing to think of the many ways that my degree candidacy was nearly derailed, only to have key support – sometimes from the most unexpected places – exactly at the right moment to keep me on track.

I am especially grateful for guidance and inspiration from my advisor, Prof. Dan Frey. He gave me the freedom to find and pursue a topic that I am passionate about, then provided enthusiastic encouragement, punctual advice, and unwavering support through the end. Without question, the best decision I made as a graduate student was asking Dan to be my advisor.

All of the work in this thesis was performed at the Charles Stark Draper Laboratory, Inc., as part of an Internal Research and Development project. In addition to being an underappreciated national resource, Draper provides an ideal base from which to pursue graduate study on interesting and challenging projects. I thank my first technical supervisor there, Ed Lanzilotta, for bringing me into Draper, helping me to find an interesting topic, giving me pep talks when things weren't going well, and helping to smooth the transition when he left Draper before my degree was finished. I thank Brenan McCarragher for agreeing to take over my research supervision at Draper for the remainder of my doctoral program; his experience in supervising past doctoral students and his uncanny ability to quickly zero in on the nucleus of a problem were both assets to my development. I also thank Jasjit Heckathorn for her help in finding a great supervisor, for patience when I experienced setbacks, and for sustained confidence in me throughout my stay at Draper.

The Education Office at Draper was one of the most important factors in my success. I simply would not have prevailed without the help of its fantastic director, Linda Fuhrman. She went above and beyond by generously allowing me to conduct my human subjects experiment in the conference room adjacent to her office, providing crucial resources in support of that experiment, and ensuring that my funding was extended during the entire eighteen-month period of time when I was "six months away from graduating." Linda's assistant Gail DiDonato was also a great resource for me, from skillfully handling expired-badge emergencies to frequently locating and distributing free food to Draper's flock of hungry graduate students. I also appreciate the sound advice, administrative help and votes of confidence from the former director of the Education Office, George Schmidt, and the interim director, Milt Adams.

I would like to express my gratitude to Prof. Missy Cummings and Prof. Warren Seering for serving on my thesis committee. Missy's expertise in statistics and human factors was invaluable, and her presence in the committee meetings guaranteed spirited discussion. It was also a great pleasure to have the perspective of Warren's broad experience in engineering design methodology from equal parts academia and industry.

To the 64 engineers who volunteered for my experiment, I am most appreciative of your time and for being good sports when I revealed the true objective of the study during the debriefing. Special

thanks go to Javier Garcia and Morris Vanegas, the two independent judges of each of the 385 video recorded reactions. Their willingness to work odd hours in less than ideal locations allowed me to get the final analysis done on time.

I would not have dreamed of pursuing this degree without the encouragement of great teachers and mentors. At LSU, Profs. Warren Waggenspack and Robert Courter gave me early opportunities and guidance. At UT/Austin, Profs. Glenn Masada, Tess Moon, and Raul Longoria all influenced my decision to continue graduate study. Outside of academia, past supervisors Scott Fish of the Institute for Advanced Technology and John Vranish of NASA's Goddard Space Flight Center were both essential to my professional development.

One of the most pleasant surprises during my time here was the discovery that when things aren't going well, there is a support structure at the Institute filled with caring people, ready to help in any way possible. In the Course II Graduate Office, Leslie Regan and Joan Kravit were unflappable in handling crises. It would be difficult to overstate how important they are to the well being of the department's graduate students. In the Office of the Dean for Graduate Education, Dean Steve Lerman, Dean Blanche Staton and Dean Isaac Colbert were all wonderfully supportive in their handling of unusual circumstances threatening to block my path to graduation.

While living in New England and studying at MIT has been an incredible adventure, it came at great personal sacrifice, and I would be remiss if I didn't thank those closest to me who had to endure nearly the same level of sacrifice: my wife Jackie and our beloved daughter Gabrielle, who both spent countless evenings over the last few years accommodating my need for a quiet environment at home. Thank you, girls.

Troy Brendon Savoie

Cambridge, Massachusetts
May 2010

Contents

List of Figures

This page intentionally left blank.

List of Tables

This page intentionally left blank.

Chapter 1

Introduction

1.1 A Definition of Engineering Design Methodology

The goal of the design engineer is to specify the structure of a product that will satisfy the end user's functional requirements as completely as possible within the allocated budget and time. Engineering design methodology is the scientific discipline concerned with creating and refining systematic approaches that the design engineer can use to reach this goal.

In general, the engineering design process may be partitioned into three components: system design, parameter design and tolerance design. System design is the first stage, where concepts for a product and its architecture are considered. This is the design step normally associated with brainstorming and innovation. When the architecture is set, the settable parameters in the product are known. Parameter design is the second stage, where the designer determines which of these parameters are important to the product's performance. Tolerance design is the third stage, where the detailed design for the product is specified. The work in this thesis applies to the parameter design step in this view of the engineering design process.

In robust parameter design, one aims to configure a system[1] such that the response to system input is on average close to the ideal response, with deviation from the ideal minimized in spite of system noise. Figure 1-1 illustrates the main attributes of a system: control factors, noise factors, input signal and response. Control factors represent design options that one may specify; for example, in the design of an automobile control factors would include fuel type, engine size, tire shape, suspension linkage geometry, etc. Noise factors represent sources of variability over which one has little or no control. In the same example, this would include things like passenger and cargo weight, pavement-tire friction, road roughness, and wind load. The input signal in this example is the combination of steering wheel,

[1] Note that the definition of "system" that applies in most instances here is that of a single physical product; for example, a laptop is a computer *system*.

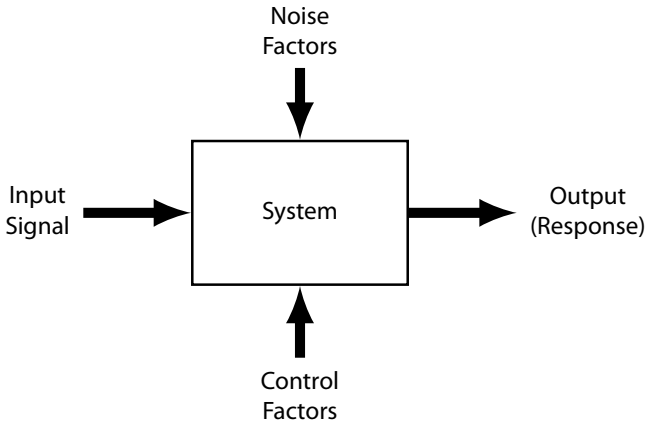


Figure 1-1: Generalized system attributes

accelerator pedal and brake pedal positions. The system response would include things like the "ride comfort" of the passengers, maximum acceleration, fuel economy, exhaust emissions, maximum speed, braking distance and aesthetic appeal. Through experimentation, the design engineer evaluates the performance of the system in order to make informed design decisions.

Before the advent of formal design methodology, one would build a limited number of candidate prototypes with design options set at one or more "best guess" configurations based on past experience and engineering judgment. Limits on development time and equipment spending meant that only a small portion of the design space could be explored, and the result was often a suboptimal design. Engineering design methodology applied to parameter design is about finding ways to explore the design space efficiently to evaluate the effect of system parameters on performance.

1.2 Motivation

1.2.1 The evolution of planned experiments

The origin of modern, structured approaches to experimental design is widely attributed to the agriculturally motivated work of Fisher (1926, 1935) more than 80 years ago, and development in this area remains active to this day (see for example Robinson et al. (2004)). Following Fisher's work, significant milestones in the theory of experimental design include Orthogonal Arrays (Plackett and Burman, 1946), Monte Carlo experiments (Metropolis and Ulam, 1949), Response Surface Methodology (Box and Wilson, 1951), Crossed Orthogonal Arrays and Signal-to-Noise Ratios (Taguchi, 1987; Phadke, 1989), Latin Hypercube Sampling (Mckay et al., 1979) and Hammersley Sequence Sampling (Kalagnanam and Diwekar, 1997). In general as experimental design theory progresses, more information about a system is obtained with fewer experiments; however, this efficiency typically comes with the price of increased complexity in the method.

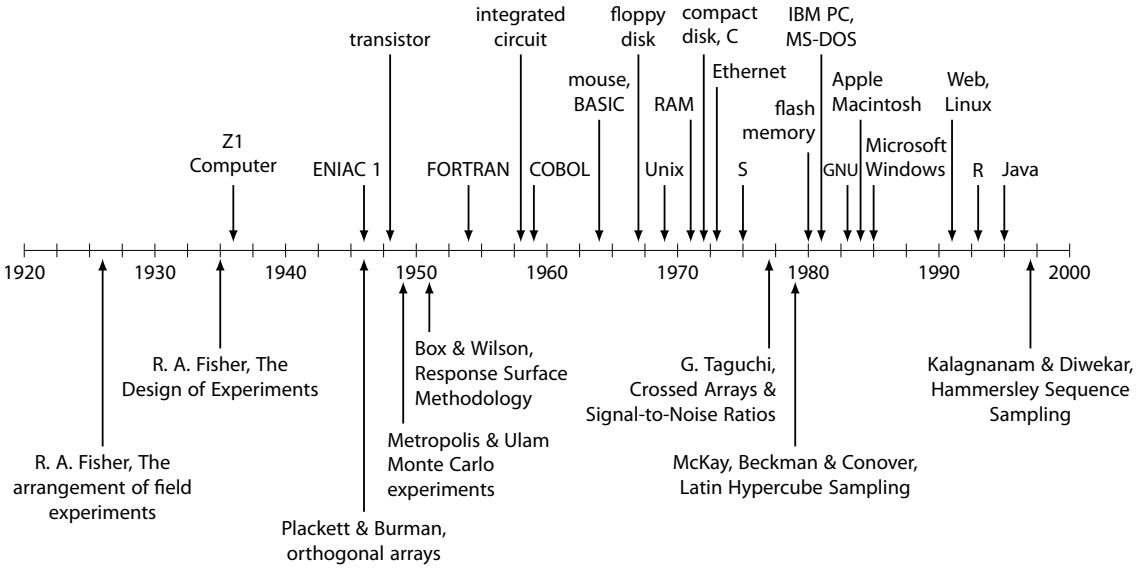


Figure 1-2: Milestones in the development of computer technology (above time line) and experimental design methods (below time line)

As one might expect, the physical experimentation at the beginning of this period gradually gave way to the computational methods widely used today, and the focus in the design research community has followed a similar trajectory. Figure 1-2 gives a time line showing the parallel development of computer technology and experimental design methodology. Of note is that the last two experimental design milestones shown on this time line were introduced specifically for use with computer simulation data.

In addition to the high cost of building prototype hardware, the combination of advances in computer technology and compressed product development schedules has made physical experimentation a rarely afforded luxury in design studies. A common theme among parameter design methods is the treatment of the system as a black box, where the goal is to build understanding of the relationships among the factors shown in figure 1-1 through targeted experimentation. The work in this thesis is focused on what happens in these sophisticated designed experiments when the black box is a computer simulation of a mathematical model of the system under study (Koehler and Owen, 1996; Booker, 1998; Simpson et al., 2001; Giunta et al., 2003; Kleijnen et al., 2005).

1.2.2 Cause for shaken trust in computer simulations

Implicit in the increasing reliance on mathematical models is an understanding that the models must accurately represent the physical system under study[2]. Indeed, a focus on the need for validity of

[2] This does not mean that there is no place for uncertainty in simulation, rather that any existing uncertainty in a system is accurately modeled as such.

computational models has led the American Institution of Aeronautics and Astronautics (AIAA) to publish the first such standard, on verification and validation of computational fluid dynamics simulations (AIAA Computational Fluid Dynamics Committee, 1998). The American Society of Mechanical Engineers (ASME) has formed committees to work on its own standards for both computational solid mechanics[3] and computational fluid dynamics and heat transfer[4]. A discussion of these and other worldwide efforts to create such standards is given in Oberkampf et al. (2004, §4.4). Additionally, some researchers have proposed teaching standard methods of commercial software development to physical scientists and engineers to help reduce errors in their codes through an organized approach to troubleshooting (Wilson, 1996; Dubois, 2005).

Despite these and similar efforts, there are indications that errors in computer models in engineering and science are fairly common. Hazelrigg (1999) contends that the majority of models used in support of engineering design are either erroneous or used incorrectly. He also presents a convincing argument that predictive models such as those used in engineering design can never really be validated (Hazelrigg, 2003). Hatton (1996) studied commercial software used in scientific and engineering analyses and concluded that errors in such software are more prevalent than previously thought. There are many infamous incidents whose causes were traced to errors in proprietary code. As a representative set of spectacular failures, one need look no farther than NASA's Moon and Mars programs. In the descent control algorithm for the lunar lander during the Apollo 11 mission, the sign of the gravitational force was incorrect, making the moon's gravity repulsive instead of attractive. The result was not a mission failure, but the lander descended faster than intended as a result of this error (Neumann, 1995). The Mars Climate Orbiter (MCO) is believed to have crashed into the surface of Mars due to a software error:

> The MCO [Mishap Investigation Board] has determined that the root cause for the loss of the MCO spacecraft was the failure to use metric units in the coding of a ground software file, "Small Forces," used in trajectory models. (JPL Special Review Board, 1999)

The Mars Polar Lander (MPL) suffered a catastrophic failure on approach to Mars for a similar reason:

> ...the probable cause of the loss of MPL has been traced to premature shutdown of the descent engines, resulting from a vulnerability of the software to transient signals. (JPL Special Review Board, 2000)

To be fair, such problems are not limited to NASA, and many examples may be found in the literature (Lin, 1985; Elmer-Dewitt, 1990; Mellor, 1998; Stevenson, 1999).

[3] ASME Performance Test Code 60 Committee
[4] ASME Performance Test Code 61 Committee

1.2.3 *Cognition and complexity*

It is well established that cognitive ability decreases with increasing task complexity (Miller, 1956; Pollack et al., 1959; Lloyd et al., 1960; Yntema, 1963; Venturino, 1997). Some have investigated different aspects of this phenomenon in the context of engineering design (Robertson et al., 1991; Hirschi and Frey, 2002; Ligetti and Simpson, 2005). This suggests a potential problem for today's computer-savvy design engineer; namely, that more of the engineer's faculties are devoted to learning and using the tools of the trade and less are available for thinking about the problem fundamentally. Lending credence to this viewpoint, Turkle (2004) proclaims that "The tools we use to think change the ways in which we think" and provides an anecdote from a personal experience at MIT in the 1970's:

> At a lunch for new faculty members, several senior professors in engineering complained that the transition from slide rules to calculators had affected their students' ability to deal with issues of scale. When students used slide rules, they had to insert decimal points themselves. The professors insisted that that required students to maintain a mental sense of scale, whereas those who relied on calculators made frequent errors in orders of magnitude. Additionally, the students with calculators had lost their ability to do "back of the envelope" calculations, and with that, an intuitive feel for the material.

Professional societies have taken notice and are starting to address this. The mission statement for the Design Process Subcommittee of AIAA's Design Engineering Technical Committee (DETC) is given by Briggs (2004) as follows:

> Pursue design integration as a systems oriented approach to improving the design engineering process. Develop cost modeling information suitable for use in projects and in teaching design engineering in the classroom. *Address the concern that increased computer use and reliance on analytical tools results in a loss of physical judgment in design outcomes.* (p. 1, emphasis added)

1.2.4 *Implications for computer experimentation in engineering design*

The preceding sections have established that the trend is to use more computer experimentation for design evaluation, but that simulation validity remains an area of concern, reflected in recent activities to establish verification and validation standards. High-profile failures traced to errors in engineering computer codes reveal that it is possible for such mistakes to go unnoticed throughout the entire design process. The ability to detect these types of blunders may be compromised when the finite cognition of the engineer is paired with experimental strategies of ever-increasing complexity. The confluence of these factors should lead one to expect such failures to occur with greater frequency in the future.

1.3 Proposed Study

Based on the preceding material, it seems reasonable to deduce that an engineer using a flawed computer simulation for design space exploration would be less likely to discover the flaw with a complex experimental plan than with a simple experimental plan. In plans that attempt to build a surrogate mathematical model of the system under study, the most efficient algorithms require many control factor changes between successive trials. As the number of factor changes increases, the amount of information in the engineer's working memory necessary to consider the changes will increase. It also seems reasonable to predict that comparing a simple one-factor-at-a-time plan to any of the more complex surrogate-model-building plans would elicit the phenomenon if it exists. To investigate this, I propose an experimental approach using engineers to perform a design task using a flawed computer simulation. The participants will not be told that there is a problem, they will be instructed to use either a one-factor-at-a-time plan in the control group or a surrogate-model-building plan in the treatment group, and the outcome of the experiment will be a measure of whether the subject identifies the flaw.

1.3.1 Special considerations for one-factor-at-a-time approach

In general, one-factor-at-a-time plans are not viewed favorably in the methodology literature. In a well-regarded textbook on experimental design, Wu and Hamada (2009) write:

> This mode of investigation is referred to as the one-factor-at-a-time approach. It is used explicitly or implicitly in many investigations, especially by those who have not been exposed to factorial experiments. By comparison with the factorial design method, it has the following disadvantages:
>
> 1. It requires more runs for the same precision in effect estimation.
> 2. It cannot estimate some interactions.
> 3. The conclusions from its analysis are not general.
> 4. It can miss optimal settings of factors.

These concerns are valid, but one-factor-at-a-time experiments do offer advantages over other approaches in certain situations. Others have specified conditions necessary to obtain advantages in optimization (Friedman and Savage, 1947; Daniel, 1973; Frey et al., 2003; Frey and Jugulum, 2006), but in this thesis the focus is on the cognitive effects. In fact, Wu and Hamada (2009) acknowledge:

> In spite of the criticisms made about the "one-factor-at-a-time" approach, it continues to be used widely in practice. One reason is that investigators are accustomed to thinking

> in terms of physical laws that govern phenomena which, from the viewpoint of cognitive psychology, are understood most readily when all factors but one are held at constant levels.

Finally, inspiration for including the one-factor-at-a-time approach in this study comes from Frey and Wang (2006):

> It is also possible that adaptive OFAT will prove useful in computer experiments in which some model errors may be present, because physically reasonable predictions are made more easily when only one factor is changed.

If the hypothesized phenomenon exists, it may turn out that effectively countering it requires only minor modifications to the more complex methods. For example, Kleijnen et al. (2005) advocate checking the signs of main effects and using preliminary analyses for screening and validation in any experimental design using simulation models.

1.4 Literature Review

As in many contemporary engineering problems, this study lies at the intersection of several disciplines. The application is engineering design methodology, but most of the prior applicable works are in the realms of experimental cognitive psychology and human factors research.

1.4.1 Theoretical Basis

The process by which an engineer might detect an anomaly in computer code can be described in an abstract sense as follows. The engineer presumably has subject-matter expertise regarding behavior of the components of a device under design consideration. This expertise might be only at the level of first-order physics, but it is assumed that there is a basic threshold of competence required for one to be compensated for performing such a task. Under this assumption, the engineer is likely capable of forming a mental model of probable behavior of the device. When observing the behavior of the device, whether a physical prototype or a computer simulation of it, the process of integrating the observed results into the framework of the existing mental model may be considered a continuous internal process of hypothesis generation and testing. When observed results are in opposition to the mental model, the attentive engineer experiences an *expectation violation* and must resolve the discrepancy before trust in the mental model is restored and work can continue. A central tenet of this thesis is that expectation violation, used interchangeably throughout this work with the terms surprise or *expectancy disconfirmation*, is the trigger to starting an attributional search that will result in discovery of the source of the flawed observation.

The concept of a mental model was introduced by Craik (1943) in 1943, but lay dormant until it was revisited in the early 1980's, separately by both Johnson-Laird (1983) and Gentner (1983). In these works, the mental model is discussed as a general internal representation of some external reality. In the 1990's, Chinn and Brewer (1993, 1996) specifically studied mental models in the context of scientific decision making. They ultimately proposed a taxonomy of responses by a scientist receiving anomalous data and found eight possibilities (Chinn and Brewer, 1998). In an abstract sense, the situation of a scientist reacting to anomalous data is equivalent to that of a designer reacting to unexpected simulation results.

Klahr and Dunbar (1988) investigated scientific reasoning in a discovery task and found evidence to support their hypothesis that such an activity can be described as alternating searches in two internal spaces: one for generating hypotheses and the other for testing them. If one accepts the assertion that an engineer thinking about a mathematical model of a physical device falls under the guise of scientific reasoning, then this mode of operation applies here as well.

The last bit of theory necessary to support the process model mentioned above is whether an engineer would experience an expectation violation and what the external manifestation of this emotion would entail. Gorsky and Finegold (1994) provides a clue to the first question in a study of high school students learning scientific concepts. In this study, cognitive dissonance caused by the "juxtaposition of opposing explanatory frameworks" was found to create disequilibrium in the subjects that was resolved upon acceptance of the new concept. The manifestation of such a disequilibrium was addressed by Stiensmeier-Pelster et al. (1995) who provided strong experimental evidence that surprise is caused by expectancy disconfirmation. Furthermore, this group of studies found that "the data support our hypothesis that surprise is ... an affective reaction to unexpectedness that precedes the attributional process or more precisely, stimulates causal thinking." (p. 29)

In summary, the literature cited above provides a theoretical basis of support for the major assumptions in the posited process model. This is important, as one must trust this logical sequence of events to accept the findings in the study. In particular, monitoring surprise as an indicator of the start of an attributional search is key.

1.4.2 Related Prior Work

This section identifies prior studies that are similar to the situation of a design engineer reacting to unexpected simulation results. In many cases, there are parallel efforts to describe the human behavior in a computer model for usage in *artificial intelligence* applications (Davis, 1984; Simon, 1986). Here the focus is primarily on the literature with actual human subjects testing.

One topic area rich in literature is that of a student learning about concepts that contradict his or her *a priori* knowledge. This is largely aimed at studying a student's ability to comprehend ideas

that are counterintuitive or at least not normally encountered outside of formal schooling. Baker (1985) describes the general field of *reading comprehension monitoring* as consisting of two distinct phases: evaluation (when the student recognizes he or she does not comprehend something in the text) and regulation (when the student proceeds with some action or thought to deal with the problem). According to Baker, the specific type of reading comprehension problem that would be applicable to our case would be that of a *semantic* type with a subtype of *external consistency*, described therein as "checking that the ideas in the text are consistent with what one already knows."

In a recent study by Rapp (2008), subjects' reading speeds were recorded for text passages containing historically inaccurate facts. One notable technique used in this work was the use of a *norming study*. The researchers assumed that the knowledge necessary for subjects to recognize external inconsistencies was held by all. To validate this assumption, the norming study was run on a separate population after the main experiments, in which subjects were asked specifically about the facts in question. The researchers set a lower limit of 70% correct responses to consider an item as being a commonly known fact. Previously collected data related to those facts that did not meet this requirement were discarded in the analysis.

In the 1970's, human subjects testing was used to study the ability of human operators to diagnose and troubleshoot failures in automated systems, from pilots monitoring the autopilot (Gai and Curry, 1977) or attempting an instrument landing (Ephrath and Curry, 1977) to simply maintaining control in spite of stability augmentation failures Sadoff (1962). Ruffle-Smith (1979) performed studies with three-man civil air crews and found that increasing workload increased the number of errors made by the crew. Lest one think that pilots had a monopoly on cognitive psychologists' attention during this time period, van Eekhout and Rouse (1981) performed similar experimental research using engineering crews in the setting of a supertanker engine control room.

Another class of human subjects research addresses diagnosis of equipment problems by workers at the level of electronics technicians (Rasmussen and Jensen, 1974) or automobile mechanics (Besnard and Cacitti, 2001). The conceptual difference between this group of studies and the work in this thesis is that in those studies, the subject begins with knowledge of the problem symptoms, whereas the current study assumes that the fundamental issue is recognizing that a problem exists in the first place.

A group of research that is conceptually close to the work in this thesis is experimental research on scientists detecting anomalies in new data sets (Alberdi et al., 2000; Trickett et al., 2001). As is typical of these types of studies, the focus was on case study of a few experts, with no possibility of drawing statistical inferences in such a limited data set.

Finally, given the financial incentives involved, it is no surprise that there has been much expenditure of energy and resources to investigate methods of debugging in software development. On the surface, it may seem similar to this problem. However, the software engineer knows that defects in code are inevitable, so much so that there is a metric specifically for it: defect density, typically expressed in

defects per kLOC[5] (McConnell, 1997). Finding and repairing software defects is accepted as one of the primary roles of the software development engineer, whereas this is not a central consideration of the engineer approaching a computer experiment as a user of the tool.

1.5 Research Summary

1.5.1 *Hypothesis*

The hypothesis under investigation here may be stated succinctly as follows.

> When an engineer uses a flawed computer simulation of a physical device together with an experimental plan for parameter design, the likelihood that the engineer identifies the flaw is decreased when the complexity of the experimental plan is increased.

The goal of this research is to perform a direct statistical test of the hypothesis, as described below.

1.5.2 *Methodological Approach*

The approach in this work is to design and execute an experimental plan to directly test the main hypothesis. Taking a case-study based approach with a small number of subjects would be highly unlikely to achieve statistical significance. The study must necessarily be large ($N > 30$) to accomplish this.

In planning the experiment, acceptable levels of Type I and Type II errors are first chosen. Next, appropriate methods for analyzing the results are selected. With a few assumptions regarding effect sizes and uncertainty of potential rare event observations, a power analysis can be performed to identify the minimum number of subjects needed to satisfy the statistical requirement.

One of the objectives of this work is to begin the process of establishing external validity. Satisfying this requires the use of human subjects from the population to which the results would apply: working full-time engineers. Many of the decisions regarding key aspects of the experiment are informed by this choice.

Finally, a direct test of the hypothesis implies that any pair of engineering design methods used as a treatment variable must necessarily be asymmetrical in complexity. However, measures are taken where possible so that this asymmetry is minimized.

[5] kLOC = one thousand lines of code

1.5.3 Main Contribution

The main contribution of this thesis is to provide the first direct experimental evidence supporting the notion that the possibility of mistakes in computer simulations should be included in the discussion when selecting a method for parameter design.

On the periphery, there are key lessons learned in this work that should, if applied, increase the productivity of investigators taking a similar tack to this problem in future experimental work.

1.6 Structure of the Thesis

- This chapter has provided a common frame of reference with which to consider the remaining material.
- **Chapter 2** presents a detailed exposition of each consideration in the approach to solving the problem. A direct test of the hypothesis using a human subjects experiment is proposed, and all of the key decisions required in the design of this experiment are supported: analytical methods, statistical power analysis, design task, experimental methods, and randomization.
- **Chapter 3** covers the experiments performed to test the main hypothesis. There are self-contained formal experimental protocols for a two-part pilot study and the full main study included.
- **Chapter 4** is a technical analysis of the experimental data, including both confirmatory and exploratory components.
- **Chapter 5** discusses the experimental results and also addresses limitations of the experimental approach.
- **Chapter 6** summarizes the important findings, discusses the implications, and gives suggestions for future related work.
- **Appendix A** provides the mathematical model and simulation results for the catapult device that is the target of the design task in the experiment.
- **Appendix B** includes copies of all graphical aids and numerical tables necessary to attempt an exact replication of each experiment.

This page intentionally left blank.

Chapter 2

Research Strategy

2.1 Overview

At its core, this research aims to study the behavior of engineers performing a parameter design task. This aligns with the general class of behavioral experimentation that uses subject-matter experts as participants to perform a task requiring this expertise. There are two approaches to performing this type of study, each with its own strengths and weaknesses.

In the first approach, the research takes place in the expert's environment, with the investigator observing the participant's behavior working on real problems. This tactic is preferred in case studies for its authenticity, but it is obviously not suited to controlled experimentation in which the investigator wishes to make conclusions based on statistically significant results. Using terminology adopted from Dunbar (1995), this type of behavioral experiment is labeled *in vivo*[1] for the parallel with biological experiments of this kind if one substitutes "living organism" with "authentic task."

In the second approach, the research takes place in the investigator's laboratory, with the subject-matter expert performing a simulated task designed by the investigator to be a controlled approximation of an authentic problem faced by the expert. This tactic sacrifices a degree of realism for more control by the investigator, so it is better suited to larger scale studies where the goal is a statistically significant result. Again adopting terminology from Dunbar, this type of behavioral experiment is labeled *in vitro*[2] for the parallel with biological experiments of this kind if one substitutes "laboratory work" with "simulated task."

To test the main hypothesis, I propose a controlled *in vitro* experiment with engineers and engineering school upperclassmen as test subjects. The objective for the test subject in this experiment is to perform a parameter design optimization of a physical system using a computer model to evaluate

[1] *In vivo* is Latin for "within the living," and is used to refer to experimentation on whole, living organisms.
[2] *In vitro* is Latin for "within the glass," meaning the test tube, a metaphor for laboratory work in general.

the performance of each configuration according to a prescribed scheme for exploring the design space. There are many factors that could be controlled in such an experiment: The qualifications of engineers selected as test subjects (area of specialization, earned degree(s), amount of post-degree experience, etc.), the *energy domain* of the physical device (mechanical, fluidic, electromechanical, etc.), the number of configurable design parameters for the device, the number of possible settings for each design parameter, the fidelity of the computer model of the device, whether the computer experiment contains an undisclosed flaw and the type of flaw if applicable, whether training in the use of the computer model is provided and the type of training if applicable, and the specific design space sampling method to be used. To minimize confounding in the experimental data, I shall simplify in each of these areas where possible.

In much of the research referenced in the preceding chapter, human subjects testing was performed with a relatively small number of subjects. Such case-study approaches can reveal valuable qualitative information, but typically there are too few subjects to claim statistical significance. These are called *hypothesis-generating* experiments.

The aim of this work is to test the main hypothesis, and that will be done with a rigorous statistical approach in a *hypothesis-driven* experiment. To do so requires *a priori* specification of predictor (independent) and response (dependent) variables, a mathematical relationship between them, a method for fitting experimental data to the model and methods to verify the statistical significance of effect(s) and goodness of fit of the model.

However, experiments that study the cognitive behavior of engineers using formal design methods are rare. There is an opportunity here to discover other aspects of this situation that merit further study. Thus, there are really two components of this work. First, the rigor described above is implemented only insofar as it applies to the hypothesis. Outside of the main analysis the experiment tends toward a hypothesis-generating mode, where the data are explored to elicit possible alternative models of behavior.

The most obvious example would be for the test subject characterization *technical ability*. It seems reasonable to assume that this will have a strong impact on the outcome of the experiment. It would be great to account for this in the model, but there are problems in doing so. As in any human subjects experiment, there is likely to be a wide range of technical ability in the sample. This could be addressed by identifying ability before the experiment using demographic data or a screening quiz, then using the result to assign treatment conditions to ensure balance. However, how exactly should technical ability be assessed? It is the opinion of this author that specific domain knowledge is most important, and demographic data such as degrees earned and years of experience may not correlate with this. Under such ambiguity, it does not seem wise to incorporate this into the criteria for experimental success, but it would be well suited for exploratory data analysis.

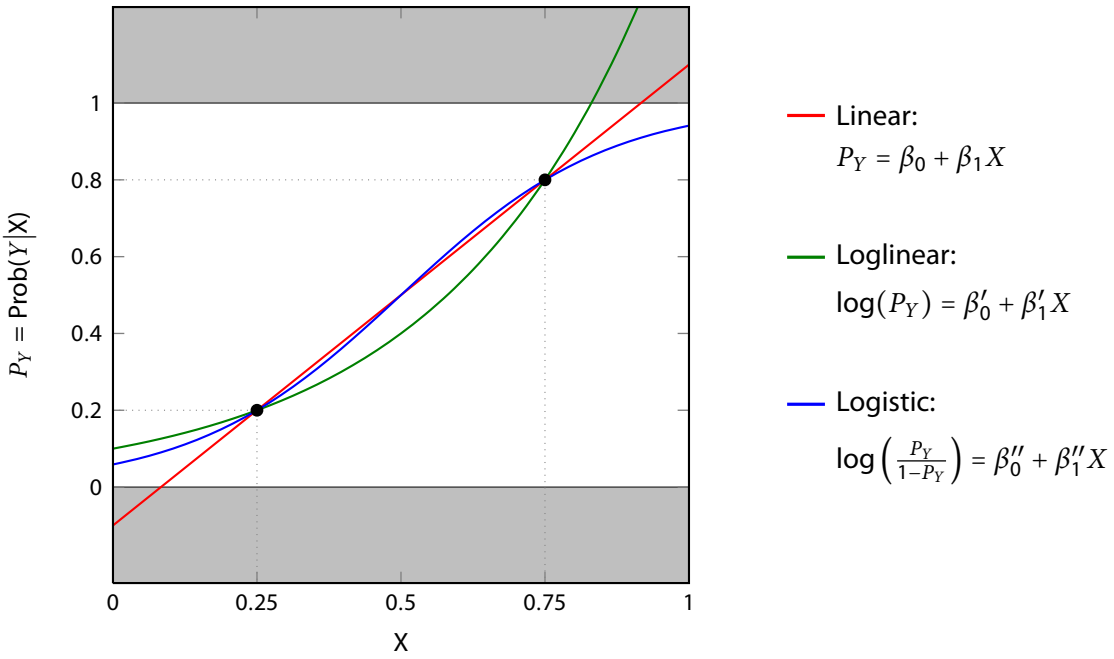


Figure 2-1: Linear, loglinear and logistic regression on proportions

2.2 Analytical Methods

This section discusses the methods to be used in this study for both confirmatory analysis and exploratory analysis. Logistic regression is presented as a means to perform a null hypothesis significance test . For exploratory analysis, methods for use with 2×2 contingency table data are given.

2.2.1 Logistic Regression

Ignoring possible nuisance variables for the moment, the goal is to test the assertion that the independent variable (design method) is a statistically significant predictor of the dependent variable (test subject response). The independent variable may be coded as $X = 0$ for the simple design method and $X = 1$ for the complex design method. The dependent or response variable is a dichotomous variable – described further in the next chapter – coded as $Y = 0$ for failure to recognize a problem and $Y = 1$ for successfully recognizing a problem. Thus the predictor variable and response variable are both dichotomous. There are various approaches one may take to analyze such data; the two most prevalent in the literature seem to be proportional comparison and logistic regression. When analyzing data from a dichotomous process, and a choice between proportional analysis with the arcsine transformation and logistic regression can be made, the "current recommendations lean toward logistic regression" (Cohen et al., 2003, p. 244).

The motivation for using logistic regression may be explained with the help of the illustration in

figure 2-1 on the preceding page, where X is the abscissa and P_Y, the probability of event $Y = 1$ occurring, is the ordinate. In this example, both the independent variable X and the dependent variable P_Y are continuous and shown for the range 0 to 1. An example pair of data points is shown at $X = 0.25$, $P_Y = 0.2$ and $X = 0.75$, $P_Y = 0.8$. If P_Y is a variable that is defined *only* from 0 to 1 (e.g. the probability of some event Y given X), then the grayed-out areas at the top and bottom of this illustration represent unrealizable outcomes.

With the pair of data points shown as black dots, one can see that a linear regression fit would protrude into the gray areas for very low or very high values of X. This would be of particular concern if the regression were used to fit a model for prediction. A partial solution to this would be to fit instead the logarithm of P_Y to a line. In this *loglinear* approach, values of P_Y less than zero are prevented, but values of P_Y greater than one are still possible. Finally, in the *logistic* approach one may fit the logarithm of the ratio $P_Y/1-P_Y$ to a line. For $P_Y = 0$ this value is $\log 0 = -\infty$, and for $P_Y = 1$ it is $\log \infty = \infty$. Thus, the transformation allows an unconstrained linear regression while ensuring that $0 \leq P_Y \leq 1$. Note also that if P_Y is the probability of an event occurring, then $P_Y/1-P_Y$ is the *odds* of the event occurring.

Logistic regression is a special case of the generalized linear model, where a transformation on the data is used to linearize the model. Here the transformation, or link function, is called the *logit* of Y.

$$logit(\hat{Y}) = \ln\left(\frac{Pr(Y = 1|X)}{1 - Pr(Y = 1|X)}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{2.1}$$

The expression in the parentheses for the logarithm is the odds of event Y occurring, so this equation is for the *log odds* of the event as a linear function of n predictor variables X_i. The regression constants β_0 and β_i are found using maximum likelihood estimation. If the equation is exponentiated, it becomes

$$\text{odds} = e^{\beta_0} \prod_{i=1}^{n} e^{\beta_i X_i} \tag{2.2}$$

This is a particularly enlightening expression, as the effect of each independent variable X_i is a multiplier $e^{\beta_i X_i}$ that is directly proportional to the odds. If X_i is not a predictor of Y, then its associated regression coefficient β_i is zero (or close to it), which means the multiplication factor $e^{\beta_i X_i}$ is one (or close to it) and the odds is not changed by the presence of that factor in the regression equation.

One often hears the term *odds ratio* in discussion of logistic regression. In the case of a dichotomous independent variable coded 0/1, the exponentiated regression coefficient is the ratio of the odds when the variable is 1 to the odds when the variable is 0. This is simple to prove. From (2.2), if the

dichotomous predictor variable of interest is X_c, then the odds when $X_c = 1$ is

$$\text{odds}|_{X_c=1} = e^{\beta_0} \prod_{i=1}^{c-1} e^{\beta_i X_i} e^{\beta_c} \prod_{j=c+1}^{n} e^{\beta_j X_j} \tag{2.3}$$

and the odds when $X_c = 0$ is

$$\text{odds}|_{X_c=0} = e^{\beta_0} \prod_{i=1}^{c-1} e^{\beta_i X_i} \prod_{j=c+1}^{n} e^{\beta_j X_j} \tag{2.4}$$

Taking the ratio and canceling terms leaves

$$\frac{\text{odds}|_{X_c=1}}{\text{odds}|_{X_c=0}} = e^{\beta_c} \tag{2.5}$$

Note that the odds ratio and the exponentiated logistic regression coefficient are only equivalent for a dichotomous independent variable coded 0/1.

For a single-input, single-output regression, (2.1) reduces to

$$logit(\hat{Y}) = \beta_0 + \beta_1 \cdot X_1 \tag{2.6}$$

The null hypothesis here is that the predictor variable X_1 has no influence over the response variable Y. This null hypothesis

$$H_0 : \beta_1 = 0 \tag{2.7}$$

is tested against the alternate hypothesis

$$H_1 : \beta_1 > 0 \tag{2.8}$$

using the *likelihood ratio* test.

In the likelihood ratio test, the test statistic is calculated as the ratio of the maximum likelihood of the null hypothesis model

$$logit(\hat{Y}) = \beta_0 \tag{2.9}$$

to the maximum likelihood of the alternate hypothesis model in (2.6). The test statistic G (Hosmer, Jr. and Lemeshow, 2000) is then calculated as

$$G = -2 \ln \left[\frac{\text{(likelihood without the variable)}}{\text{(likelihood with the variable)}} \right] \tag{2.10}$$

The likelihood values are usually given at the end of the iterative maximum likelihood estimation algorithm in most software packages that have the capability to perform logistic regression. Under the

null hypothesis, the test statistic G approaches the chi-square distribution with one degree of freedom, and a one-tailed test on this distribution may be used to decide to accept or reject the null hypothesis.

2.2.2 *Contingency Table Analysis*

The generalized 2×2 contingency table is shown in table 2.1. It is possible to calculate odds ratios directly from the frequency data in this table. Morris and Gardner (1988) provide equations for estimating the odds ratio with a 95% confidence interval. The odds ratio is

$$OR = \frac{\text{odds(Success}|\text{Treatment)}}{\text{odds(Success}|\text{Control)}} = \frac{n_{11}/n_{10}}{n_{01}/n_{00}}. \tag{2.11}$$

The error associated with this value is estimated by assuming that the logarithm of the odds ratio is normally distributed. The equation for the standard error of the logarithm of the odds ratio in that case is

$$SE(\log OR) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{00}}}. \tag{2.12}$$

The 95% confidence interval for the odds ratio may be calculated by combining (2.11) and (2.12) to get

$$OR_{95\%\text{ CI}} \in \left(e^{\log OR - 1.96 \cdot SE(\log OR)}, e^{\log OR + 1.96 \cdot SE(\log OR)}\right) \tag{2.13}$$

Although it is mathematically advantageous to work with the odds ratio, particularly in cases with one or more continuous predictor variables as described in §2.2.1, there is a similar descriptive parameter that may be calculated from 2×2 contingency data: the *risk ratio*. This parameter – also called the *relative risk* – is a ratio of probabilities rather than odds. It is a more intuitive alternative to the odds ratio, since the concept of probability may seem more natural than that of odds. The risk ratio is estimated from contingency table data (Morris and Gardner, 1988) as in table 2.1 according to

$$RR = \frac{\text{Prob(Success}|\text{Treatment)}}{\text{Prob(Success}|\text{Control)}} = \frac{n_{11}/(n_{11}+n_{10})}{n_{01}/(n_{01}+n_{00})}. \tag{2.14}$$

The error associated with this value is estimated by assuming that the logarithm of the risk ratio is normally distributed. The equation for the standard error of the logarithm of the risk ratio in that

Table 2.1: Contingency Table for Calculating Risk Ratio

	Outcome		
Group	Success	Failure	Total
Treatment	n_{11}	n_{10}	$n_{11} + n_{10}$
Control	n_{01}	n_{00}	$n_{01} + n_{00}$

case is

$$SE(\log RR) = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{11}+n_{10}} + \frac{1}{n_{10}} - \frac{1}{n_{10}+n_{00}}}. \tag{2.15}$$

The 95% confidence interval for the risk ratio may be calculated by combining (2.14) and (2.15) to get

$$RR_{95\%\text{ CI}} \in \left(e^{\log RR - 1.96 \cdot SE(\log RR)}, e^{\log RR + 1.96 \cdot SE(\log RR)}\right) \tag{2.16}$$

Although this section addressed calculation of the risk ratio specifically for a 2×2 contingency table, this property may also be calculated for the more general case of more than one predictor variable, one or more of which may be continuous. This is exactly analogous to calculating the odds ratio by exponentiating the logistic regression coefficients, but the risk ratio comes from the exponentiated coefficients obtained through loglinear regression. For reasons discussed in §2.2.1, it is possible to generate a physically unrealizable model using loglinear regression (see fig. 2-1 on page 29). However, the risk ratio is a useful parameter if the data are dichotomous, it is not used to create a model for predicting behavior, and the analysis is exploratory in nature.

2.3 Experiment Size

As discussed previously, the main hypothesis in this study will be assessed using null hypothesis significance testing. In this approach, one assumes that the null hypothesis is true, then performs the experiment. If the observed results are highly unlikely to be obtained under the null hypothesis, it is rejected in favor of the alternate hypothesis; otherwise, the null hypothesis should not be rejected. In general, there are three interconnected factors in a null hypothesis significance test: the effect size, the criterion for rejection of the null hypothesis, and the number of samples. The interdependence among these factors is illustrated in figure 2-2 on the following page. In this figure, each of the two plots shows the probability density functions for the null hypothesis H_0 and the alternate hypothesis H_1. If H_0 is true and the experiment is repeated a large number of times, the density function on the left obtains for the histogram of the results. If H_1 is true and the experiment is repeated a large number of times, the density function on the right obtains for the histogram of the results. However, the objective is to infer from a single experimental result whether the null hypothesis may be rejected. This is done by comparing the result to the *criterion*, shown in the figure as the thick, downward-pointing arrows. If the observed result is on the left side of the criterion, there is not sufficient evidence to rule out the null hypothesis. If the observed result is on the right side of the criterion, it is very unlikely that the null hypothesis is true. An error in inference occurs when the observed result and the true value are on opposite sides of the criterion. Referring to the shaded areas in figure 2-2, the Type I error rate α is equal to the normalized area under the H_0 probability density curve for all values greater than the criterion, and the Type II error rate β is equal to the normalized area under the H_1 probability density curve for all values less than the criterion.

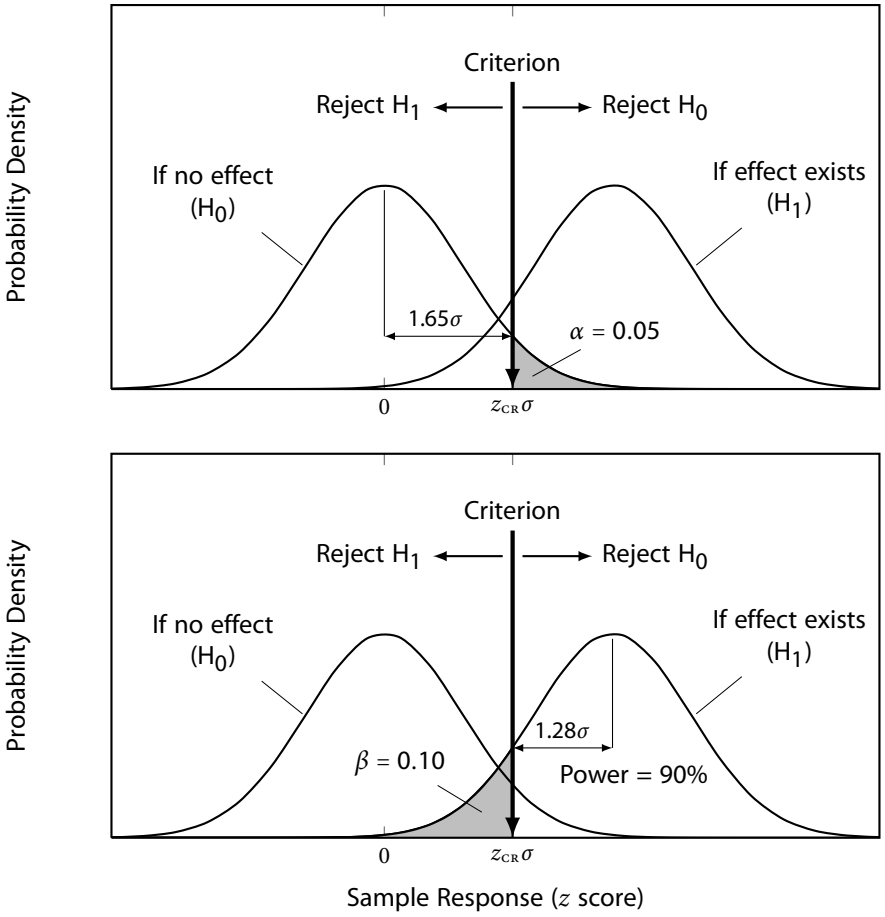


Figure 2-2: Parameters in null hypothesis significance testing

Many researchers stop here in reporting the results, and this is the basis for much criticism of null hypothesis significance testing. The central issue is that there may be a high likelihood of a false positive result, unless it is explicitly considered during planning. This criticism may be effectively countered by specifying the maximum tolerable Type II error rate β in calculating the required experiment size and by disclosing the observed effect size with confidence limits in reporting the results, and this is the approach taken here.

2.3.1 Sample Size Requirements for Statistical Power

In planning for the number of samples to be collected, one should consider the risk of making both Type I and Type II errors. The Type I error risk is typically specified by $\alpha = 0.05$, meaning that the rate of making the error of incorrectly rejecting a true null hypothesis is at most 5%. The Type II error risk is typically specified by $\beta = 0.20$, meaning that the rate of making the error of failing to reject a false null hypothesis is at most 20%. This is a trade off between a 1-in-20 risk of finding effects that do

Table 2.2: A Convention for Effect Size Indices Suggested by Cohen (1988).

Effect Size	h	Examples of Proportion Pairs with this Difference
small	0.2	(0.05, 0.10); (0.20, 0.29); (0.40, 0.50); (0.60, 0.70)
medium	0.5	(0.05, 0.21); (0.20, 0.43); (0.40, 0.65); (0.60, 0.82)
large	0.8	(0.05, 0.34); (0.20, 0.58); (0.40, 0.78); (0.60, 0.92)

not exist with a 1-in-5 risk of not finding effects that do exist.

The complement of the Type II error rate, $1 - \beta$, is called the *power* of the experiment. It is the rate of correctly rejecting a false null hypothesis. For the widely used value of $\beta = 0.20$, the experimental power is thus 0.80 or (more commonly) 80%. There is some criticism in the literature that this is a barely acceptable value, and that an experimental power of 90% or even 95% should be used (Muller and Benignus, 1992). Here, I shall adopt the moderate approach to plan this experiment for 90% power.

In addition to Type I and Type II errors, when planning for sample size one must consider the expected size of the effect. Obviously, this is not known before the experiment, so it is simply a best guess. Many researchers follow the recommendations given by Cohen (1988) for a general "effect size index" h based on the arcsine transformation of the proportion of successes in an observed group. Based on a literature survey to study trends in reported effect sizes, Cohen suggests the values in table 2.2 as a basis for convention. In this study, I expect that roughly 3-in-5 subjects will detect a problem when using the simple design ($X = 0$) method and 1-in-5 subjects will detect a problem when using the complex design ($X = 1$) method. Such a pair of observed values of $(0.2, 0.6)$ would mean $h \approx 0.8$, which would be a "large" effect according to Cohen.

There have been few studies done on power analysis in logistic regression, and no consensus on the best approach to the task. Whittemore (1981) proposed a method of calculation that is based on using the Wald statistic for testing the regression coefficients, but it assumed low overall response rates. A more recent work by Hsieh et al. (1998) compared several methods for accuracy and concluded that a relatively simple equation gives good results for univariate logistic regression. For a design that is balanced in the independent variable X, it is

$$n < 4P \cdot (1 - P) \left(\frac{\mathcal{Z}_{1-\alpha} + \mathcal{Z}_{1-\beta}}{P_1 - P_2} \right)^2 \tag{2.17}$$

where P is the overall event rate for a balanced design

$$P = \frac{P_0}{2} + \frac{P_1}{2} \tag{2.18}$$

and P_0 and P_1 are the measured event rates for the population samples corresponding to $X = 0$ and

$X = 1$, respectively. The minimum required sample size for the entire range of possible response rates is shown as a contour map in figure 2-3a on the next page.

2.3.2 *Sample Size Requirements to Minimize Separation Artifacts*

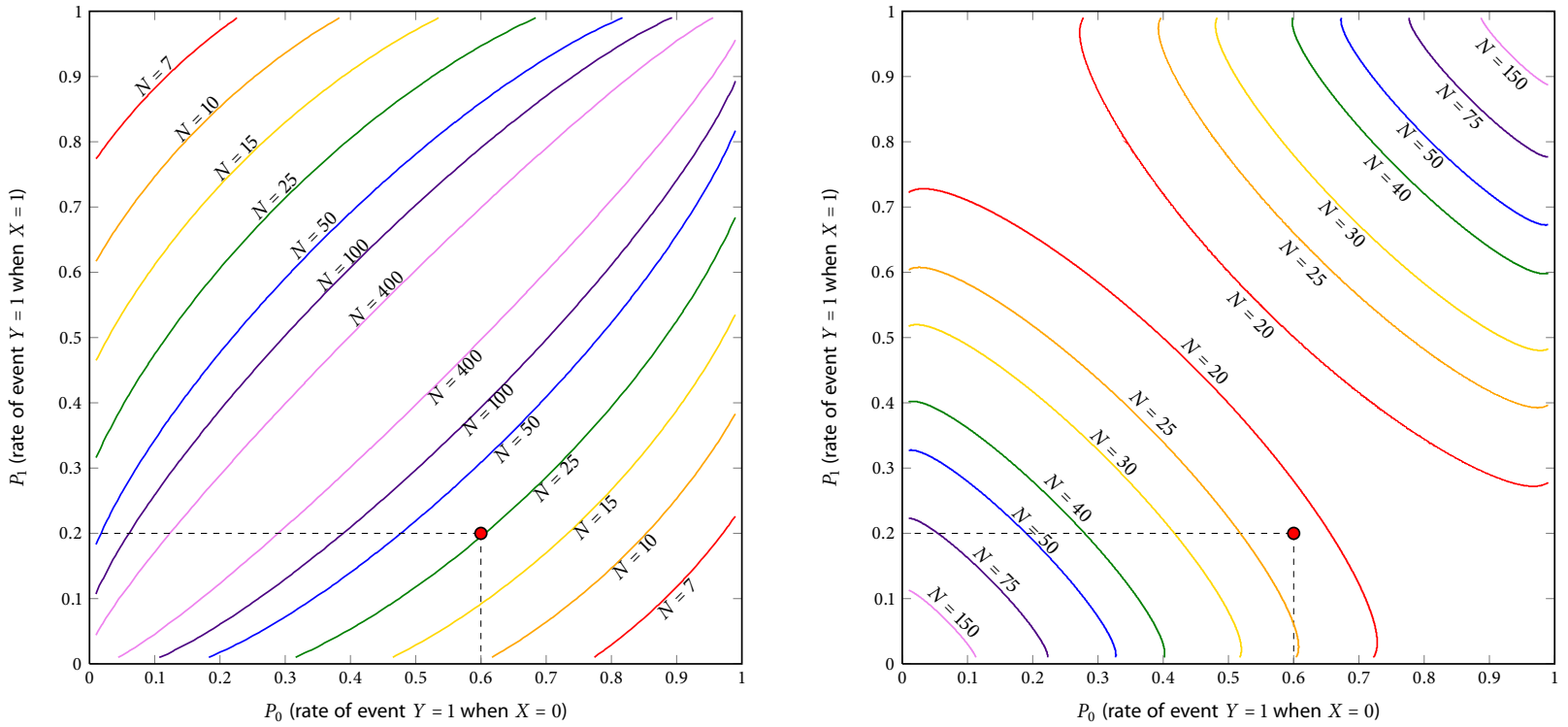
Another issue in using logistic regression is the problem of *separation*, where for instance the observed experimental data produces a model in which the effect of one or more independent variables is completely predictable. For example, $Y = 0$ is observed for all $X = 0$, then the likelihood calculation would find $Pr(Y = 0 | X = 0) = 0$. This means that the log odds is $-\infty$ and the logistic regression model will not work since a finite solution for β_0 does not exist. There are various methods of dealing with this situation in practice, but it would be better to avoid it if possible.

A similarly troublesome situation may arise when the response rate is low. In a widely cited article on this topic, Peduzzi et al. (1996) found through Monte Carlo simulation that the accuracy of the logistic regression model is compromised when the minimum event count *per covariate* for any population is lower than 10. Vittinghoff and McCulloch (2007) recently revisited this problem and found no appreciable degradation in performance for the logistic regression model when the event count per variable is in the range of 5 to 9. However, I shall follow the more conservative advice of a minimum of 10. For the proposed univariate model, this requires observing at least 10 responses of $Y = 0$ and at least 10 responses of $Y = 1$.

Addressing this requires careful thought. The initial instinct may be to take the number of events $Y = 0$ to be equal to $N \times Pr(Y = 0 | X = 0)$ plus $N \times (Pr(Y = 0 | X = 1)$, where N is the number of test subjects assigned to each of the two treatment groups. Assuming the conditional probabilities are correct, this is in fact the *expected value* of the number of events $Y = 0$. However, there is uncertainty associated with these values given by their binomial probability distributions. To be reasonably certain that the minimum event count will be reached, one must take into account this uncertainty. One *ad hoc* method for doing so would be to choose the sample size such that the sum of the lower boundaries of the 95% confidence intervals for the expected number of successes in each treatment group is at least 10. Using the expected value and variance from the binomial distribution, each asymmetric interval is

$$n \cdot p - \mathcal{Z}_{0.95} \cdot \sqrt{n \cdot p \cdot (1 - p)} \quad < \quad N_{s,95\%} \quad < \quad \infty \tag{2.19}$$

Solving for n such that the minimum value of N_s in the 95% confidence interval is at most 10 gives the values shown as a contour map in figure 2-3b for all possible event rates.



(a) Required for $\alpha = 0.05$, $\beta = 0.10$

(b) Required for minimum event count of 10

Figure 2-3: Minimum sample size required for all possible effect sizes

2.3.3 *Conclusion*

The ethical consequences of improper sample size selection may be summarized as follows:

- If the experiment is sized such that it is overpowered, whereby an investigator plans for a small effect but actually expects and ultimately observes a large effect, more human subjects were subjected to testing than were actually necessary. This is considered to be ethically questionable (Bacchetti et al., 2005).

- If the experiment is sized such that it is underpowered, whereby the effect being tested exists and is important but smaller than the investigator estimated in planning, human participants have again been subjected to unnecessary testing because the important effect was not found (Halpern et al., 2002).

The pair of contour maps in figure 2-3 illustrate the statistical consequences of sample size selection for the conservative values of $\alpha = 0.05$, $\beta = 0.10$ and $N_{s,min} \geq 10$. The point $P_0 = 0.6$, $P_1 = 0.2$ is shown on each map and is a rough guess of the expected outcome of this experiment. For the simple design method $X = 0$, the guess is that 3-in-5 subjects will recognize anomalous results in the simulation data. This guess assumes that not all human subjects will be experts. Those with lesser domain knowledge may have a difficult time recognizing aberrant behavior even with the simple design method. For the complex design method $X = 1$, the guess is that 1-in-5 subjects will recognize anomalous results in the simulation data. It may be that a few test subjects are particularly good at recognizing patterns, or that a few test subjects take a lucky guess at the location of the problem, or that the method by which test subjects are judged is slightly flawed. For whatever reason, the author believes that there will be a few test subjects that are judged to locate the problem successfully.

The minimum sample size of 25, required to find the estimated effect with 90% power, exceeds the minimum sample size of about 22 required to be reasonably assured of meeting the minimum event count requirement for a good logistic regression. Therefore, the goal is to have 25 valid data points for each of the two treatment groups, for a total of 50. It would be prudent to plan for the necessity of some data points to be discarded. For example, a test subject could choose to withdraw, or the test administrator could make an error in executing the protocol. It is also good practice to pilot test the protocol with a small number of human subjects. Altogether, the total number of test subjects required for this experiment will be planned as 60: 50 valid points, six potentially discardable points and four pilot test points.

2.4 Conceptual Framework for Experiment

2.4.1 Introduction

This section discusses the specific concepts that must be addressed before detailed design of this human subjects experiment can be completed. According to the analysis in the preceding section, a fairly large number of technically trained volunteers is required to be reasonably certain of a successful experiment. This requirement drives many of the decisions that follow in this section.

2.4.2 Eligibility of Human Subjects

As the statistical power analysis in §2.3 showed, approximately 60 volunteers are needed for this experiment. These should come from a pool of candidates that have, at a minimum, the domain knowledge necessary to have engineering intuition about the physical system under design. Given the large number of test subjects required and the relatively sparse dissemination of formal design of experiments knowledge among engineers, it is not feasible to require test subjects to have this as well.

The author has limited access to a pool of approximately 800 technical personnel at an engineering research and development company with current projects in nearly every industry in which engineers are employed. The benefit of soliciting volunteers from this population is that there are a wide variety of technical specialty areas, and a range of experience levels from current engineering undergraduate students to those with doctoral degrees and 40 years of experience. There is some criticism (e.g. Wintre et al., 2001) that too many human subjects experiments rely on volunteers exclusively from university student populations, making broad interpretations questionable. By recruiting volunteers from the company described above, this potential problem is avoided.

Since there is such a wide array of technical expertise in the candidate population, required domain knowledge should be limited to the "greatest common denominator." Regardless of discipline, practically all candidates should have taken a college-level course in general physics for engineers and scientists. To enable the largest number of candidates from this population to participate, it seems reasonable to require domain knowledge to be at the level required to pass such a course.

2.4.3 Physical Device to be Designed

The device described here is a physical object that is the target of the design task for human subjects in this experiment. As the domain knowledge of test subjects is limited as described above, so should the complexity of this device be limited. A source of inspiration for this comes from the following devices used to teach design of experiments using in-class demonstration: ball in funnel (Gunter, 1993), paper

airplane (Sarin, 1997), paper helicopter (Box, 1992), tabletop hockey (Anderson, 2009) and catapult (Antony, 2002). The selected device should be relatively simple to model mathematically yet not so simple as to be a trivial problem. A device that meets this criteria is the catapult.

In many catapults used in teaching design of experiments, there are strong interactions between some of the control factors. Although instructive in that context, such strong interactions are perhaps not necessary for this experiment and would only add confounding to the results. One catapult appears to be designed such that only main effects are important if the configurations of interest are carefully selected: the Xpult catapult shown in figure 2-4. According to the manufacturer,

> The Xpult catapult is an instructional aid, developed by Professors Christian Terwiesch and Karl Ulrich at the University of Pennsylvania, for teaching engineering, science, design of experiments, Taguchi methods, and problem solving.(Peloton Systems LLC., 2009)

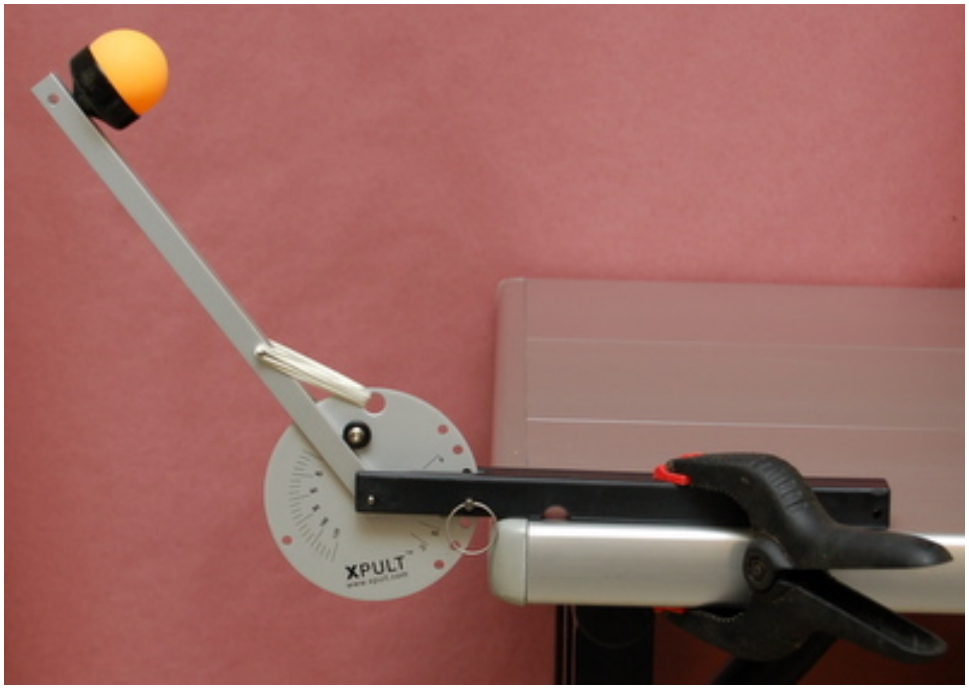


Figure 2-4: The Xpult catapult (©Peloton Systems LLC. Used with permission.)

Operation of the catapult is simple: one clamps the base to a table, pulls back the metal arm, places a table tennis ball on the holder at the end of the arm, then releases the arm to propel the ball into a ballistic trajectory. In its intended role as a tool for teaching statistical design methods, the device has four control factors:

1. *Type of ball.* Table tennis or perforated golf.
2. *Number of rubber bands.* 1, 2, or 3.

3. *Pullback*. Any position is possible on the continuum from 20° to 120°. Practically speaking, pullback can be limited to a much smaller range for high launch angles.

4. *Launch angle*. Indexed at 0°, 15°, …, 90°.

The canonical design task for this device is to find a combination of the four control factor settings that will reliably hit a target or go into a bucket placed 96 inches horizontally from the catapult pivot.

2.4.4 Design Task

To avoid experimental bias introduced by an imbalance in resources required, and to satisfy the requirement of a nontrivial design task, the following virtual modifications were made to the device:

1. The perforated golf ball was not used, due to the unnecessary complexity it adds to the system regarding both the mathematical model and human predictive ability. Instead, another smooth table tennis ball was used – one that is 10% larger in diameter but about 15% lighter in weight.

2. Choice of rubber bands was limited to 2 or 3.

3. Pullback was limited to two discrete positions: 30° or 40°.

4. Launch angle was limited to two indexed positions: 45° or 60°.

5. Material comprising the arm was introduced as a control factor: Aluminum or Magnesium.

6. Two ambient environmental conditions were introduced as controllable factors: temperature (32°F or 72°F) and relative humidity (25% or 75%).

The three additional control factors, and the limitation of each control factor to two possible settings, makes this 2^7 system an appropriate choice for the design task. Another benefit to using this configuration over the standard one is that it more closely reflects the tendency of systems to have sparse effects (see *effect sparsity principle* in Wu and Hamada, 2009). The original four control factors each affect the outcome significantly, while two of the three added factors have very small effects on the outcome.

The design task for the participants to complete is to find the combination of control factor settings, from the set of $2^7 = 128$ possible configurations described above, that results in the table tennis ball landing nearest the target distance of 96 inches from the catapult pivot point. The method used to perform this task is the treatment variable in this experiment, and both treatments are described in §2.4.6.

2.4.5 *Flaw in the Computer Simulation Results*

In carrying out the design task, participants are required to evaluate the response of the catapult for each configuration of control factor settings prescribed by the specific design algorithm that has been assigned. This evaluation is completed by recording the outcome of a computer simulation of the device in the appropriate configuration. In general, there are several conceivable situations in which an engineer should suspect a problem with a simulation of a physical device:

1. *Accuracy*. The magnitude of the response is not as expected. For example, if designing a conventional automobile and using the simulation to estimate the fuel economy, one should expect the calculated result to be somewhere in the range of 10 to 40 miles per gallon (mpg) depending on the size of the vehicle and its engine. If the simulation instead returns a value of 2 mpg or 200 mpg, one should be highly suspicious of the result and investigate further.

2. *Scale*. In comparing the simulation results of two configurations of the device, the change in magnitude of the response is not as expected. For example, if designing a conventional automobile and using the simulation to estimate the range of the vehicle (i.e., total possible distance traveled on a full tank of fuel) and only changing the size of the fuel tank, one should expect the calculated range to be directly proportional to fuel tank size. If doubling the size of the fuel tank results in no appreciable change in the range from the simulation, one should suspect a problem in the calculation.

3. *Sign*. In comparing the simulation results of two configurations of the device, the *direction* of the change in magnitude of the response is not as expected. Using the preceding example, if doubling the size of the fuel tank results in a *reduction* in calculated range, one should suspect a problem in the calculation.

The theory postulated in this thesis is that comparison between successive trials in the simpler design algorithm is more likely to cause an engineer to recognize anomalous data from a computer simulation. In the first situation above, one does not need to compare two configurations of the device for the problem to become evident; thus, this type of problem is not of interest in this study. The second and third situations require a comparison of configurations to become evident, and this is relevant since the number of configuration changes between successive simulations is dependent upon the choice of design space sampling technique.

There are conceivably many ways in which scale and sign problems can appear in simulation results. In fact, this would be an interesting treatment variable in an experiment similar to the one described here. One could insert a misplaced sign or decimal point, use inappropriate integration algorithm parameters (e.g. step size too large), or have an error in calculating contact between two physical objects (e.g. ball goes through a wall instead of bouncing off). Such a treatment variable would be secondary to the variable that directly tests the main hypothesis, so only one scale or sign problem

will be selected for this experiment. Furthermore, it should be obvious that the scale or sign problem must be observable upon changing control factor settings.

The mathematical model of the catapult used in this task is presented in detail in Appendix A. For the control factors and settings described in §2.4.3-2.4.4, the full factorial results are given in table A.4 on page 123. The "main effects" calculated from these data are shown graphically in figure A-7 on page 120 and as a Pareto chart in figure 2-5 on page 45. Reasoning for each of the effects is as follows:

- *Type of table tennis ball.* The choice is a "regulation" table tennis ball or a "large-ball" table tennis ball. High quality table tennis balls are constructed of highly pressurized, gas-filled celluloid and standardized with tight tolerances for both diameter and weight. The regulation ball has a diameter of 40 mm and a mass of 2.67 g, and the large-ball version has a diameter of 44 mm and a mass of 2.30 g. Since the mass of either of these is only 4% to 8% of the mass of the catapult arm, the launch speeds should be nearly equal. However, during ballistic flight the aerodynamic drag force on the larger ball is 20% greater at launch and is acting on a mass that is 14% lower than for the regulation ball. According to Newton's Second Law of Motion, each of these factors will cause the larger ball to decelerate more than the regulation ball. The full factorial results show that, on average, the regulation ball travels about 7 inches farther than the large ball.

- *Number of rubber bands.* The choice is 2 or 3 rubber bands to be used as the propulsion for the catapult. Rubber bands are typically difficult to model because of creep, temperature effects and large manufacturing variations. Including these effects in the simulation model would add unnecessary complexity for this experiment, so the rubber band is modeled here as a simple ideal spring governed by Hooke's Law. Since the energy put into the system is directly proportional to the equivalent spring constant, and the spring constant is directly proportional to the number of rubber bands, one should expect a change from 2 to 3 to increase the ball's landing distance by about 50%. The full factorial results show that, on average, using 2 rubber bands the ball lands at about 76 inches and using 3 rubber bands it lands at about 108 inches. This is slightly less than a 50% increase, which is reasonable since the aerodynamic drag force on the ball will be greater at the higher velocity attained using the extra rubber band.

- *Arm material.* The choice for the arm material is Magnesium or Aluminum. For the same geometry, the former is about 34 g while the latter is about 53 g – an increase of just over 50%. All else being equal, a higher mass moment of inertia about the catapult pivot will result in a lower launch velocity for the ball. The full factorial results show that, on average, using a Magnesium arm the ball lands at about 101 inches or using an Aluminum arm the ball lands at about 82 inches.

- *Launch angle.* The choice for launch angle is 60° or 45°. Simple physics states that a 45° launch angle results in the longest distance a ballistic object will travel *in a vacuum*. The farther away from this launch angle in either direction, the shorter the landing distance for the object. The

effect of drag on a ballistic trajectory is to decrease this optimal launch angle slightly, typically between 45° and 40°. Thus, one can safely assume that changing the launch angle from 45° to 60° will result in a shorter landing distance. The full factorial results show that the average landing distance goes from about 98 inches to about 85 inches.

- *Pullback*. The choice for pullback is 40° or 30°. One can deduce from the geometry of the device that a greater pullback will result in the rubber bands stretching more. The additional kinetic energy that this imparts will cause the ball to travel farther. The full factorial results show that the average landing distance goes from 105 inches for a 40° pullback to 78 inches for a 30° pullback.

- *Ambient temperature*. The choice for ambient temperature is 32°F or 72°F. Had it not been excluded from consideration, the greatest effect from temperature change would be in the rubber band response. With the rubber band modeled as a spring, temperature most strongly affects the density and viscosity of air (shown in fig. A-5 on page 115) used in the aerodynamic force calculations. The full factorial results show that the average landing distance goes from about 90.5 inches at 32°F to about 92.5 inches at 72°F.

- *Relative humidity*. The choice for relative humidity is 25% or 75%. Changing humidity also affects the air density and viscosity (shown in fig. A-5), but the effect is even more slight than for the change in ambient temperature. The full factorial results show that the average landing distance goes from slightly more than 91 inches at 25% humidity to slightly less than 92 inches at 75% humidity.[3]

Referring to figure 2-5 on the facing page, at the control factor levels for this experiment there are two large effects (number of rubber bands, pullback), three medium effects (arm material, launch angle, type of ball) and two nearly negligible effects (ambient temperature, relative humidity). One of these control factors must be tied to the intentional flaw in the simulation. There are several good reasons to choose the arm material for this:

1. Control factors related to mechanical components are more likely to be intuitive than those related to the aerodynamic model.

2. There is potential for confusion between launch angle and pullback, since one is absolute and one is relative.

3. Considering the type of ball requires thinking about at least three aspects of the simulation: launch speed, aerodynamic drag and momentum.

[3] This is not a typographical error. Increasing humidity decreases both the density and viscosity of air. It may help to understand this counterintuitive result by thinking of humid air as a mixture of dry air and water vapor. Water vapor (H_2O = atomic weight $2\times1+1\times16 = 18$) is less dense than dry air (78% N_2 + 21% O_2 = atomic weight $0.78\times2\times18+0.21\times2\times16 \approx 35$) so the mixture decreases in density as the component of water vapor increases when humidity increases.

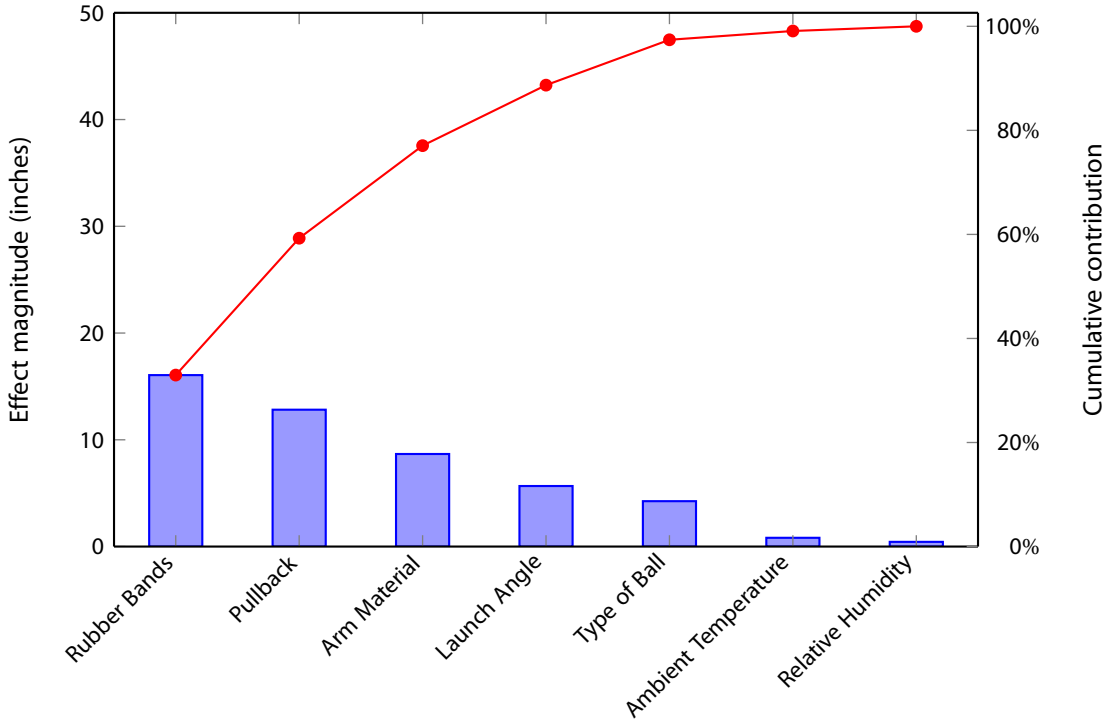


Figure 2-5: Pareto chart of the control factor main effects in the catapult device

4. The rubber band is such an obvious and dominating effect that it is difficult to imagine even a non-engineer failing to recognize a problem tied to it.

5. The effect of the catapult arm material is the largest of the three moderately-sized effects.

There is more than one way to tie an intentional flaw to the arm material. Here it is done by switching the numerical value of the mass of the Aluminum arm with that of the Magnesium arm. Thus, in the erroneous simulation, the result provided to the test subject as the simulation result for a system with the Aluminum arm will actually be the correct value for the system with the Magnesium arm, and vice versa. This type of usage mistake could occur in an otherwise valid simulation model if the model constants are not input correctly.

2.4.6 Experimental Treatment: Design Algorithms

Introduction

This section concerns the two design space exploration methods selected for use as the treatment variable. There are several criteria that any design method used in this study should meet:

1. The method should be easy to understand. There are many existing formal methods of design

space exploration, with varying degrees of difficulty in understanding them. Typically, as the effectiveness of the method increases, so does the complexity involved. This study requires testing a rather large number of human subjects, individually and in person to ensure the study's integrity. This in effect limits the amount of time that may be spent with each subject. Further, in general one cannot assume the test subjects have preexisting knowledge of the design methods. Since each subject must be trained on the design method to be used, the method must be relatively simple to allow both training and testing to be completed in a reasonable amount of time. I anticipate each test subject's available time for this test to be at most one hour.

2. The method should be easy to implement. Although there are many useful techniques that could be used when stepping through the design, one must balance between utility and complexity. For example, an often-used technique in basic screening designs is to plot main effects and two-factor interaction effects so that a quick visual analysis is possible. To do that here, the test subjects would have to construct such graphics by hand, or the process could be automated using the computer. In the first case, this would add to both the method complexity and time required for both training and testing, and in the second case this would be removing part of the process from the test subject and possibly inducing complacency effects into the experiment.

3. The method should be capable of meeting a design goal of finding the control factor settings that result in system performance close to a specified target response. This is a pragmatic requirement resulting from the choice of physical system used for the design task, the Xpult catapult. In this system, if the goal were, say, to maximize the distance of the ball's first landing point, one could use a knowledge of basic physics to determine the required configuration without any testing at all. This is due to the decoupled nature of the control factors in the system. The design goal in this case is to hit (or get as close as possible to) a specified target position. Incidentally, this is the way in which the developers of the Xpult (professors of business administration teaching experimental methods) intended the device to be used.

The methods to be used in this study are the adaptive one factor at a time (aOFAT) and the Plackett-Burman L_8 orthogonal array (PB-L_8) approach. The algorithm for each is described below. It is often helpful when discussing such algorithms to visualize a system with three two-level control factors as a three-dimensional space as shown in figure 2-6 on the next page. In this spatial abstraction, each factor is assigned to an axis, and position along the axis corresponds to the setting for the factor – minus signs and plus signs denote the nominal and alternate settings, respectively. Each of the corners of this cube represents one possible configuration of the system. The algorithms will demonstrate methods for finding a set of optimal control factor settings without the brute-force approach of evaluating every possibility.

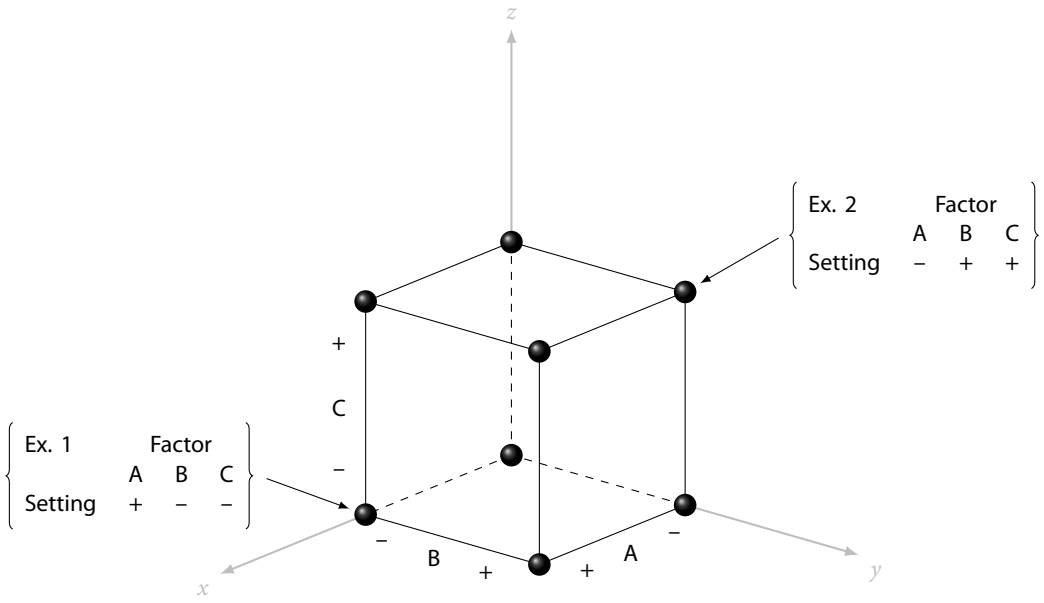


Figure 2-6: Spatial abstraction of 2^3 system for algorithm illustration

Adaptive One Factor at a Time

In the adaptive one factor at a time (aOFAT) method, given the set of control factors and their possible settings, the algorithm is:

1. Select a starting configuration either at random or based upon *a priori* knowledge of the system.
2. Evaluate the system response at the starting configuration.
3. For each control factor in the system
   (a) Select a new configuration by using the previously evaluated configuration with the best performance, changing only this factor's setting to its alternate value.
   (b) Evaluate the system response at the new configuration.
   (c) If the performance improves at the new setting of this factor, keep it at this setting for the remainder of the experiment; otherwise, keep it at the original value for the remainder of the experiment.
4. The configuration obtained after stepping through each control factor exactly once is the optimized result for this design approach.

This is illustrated graphically for a system with three two-level factors in figure 2-7 on the following page. The catapult to be configured in this design task has seven control factors, each with two level settings of interest (i.e., a 2^7 system). Using the aOFAT algorithm in this case requires eight trials

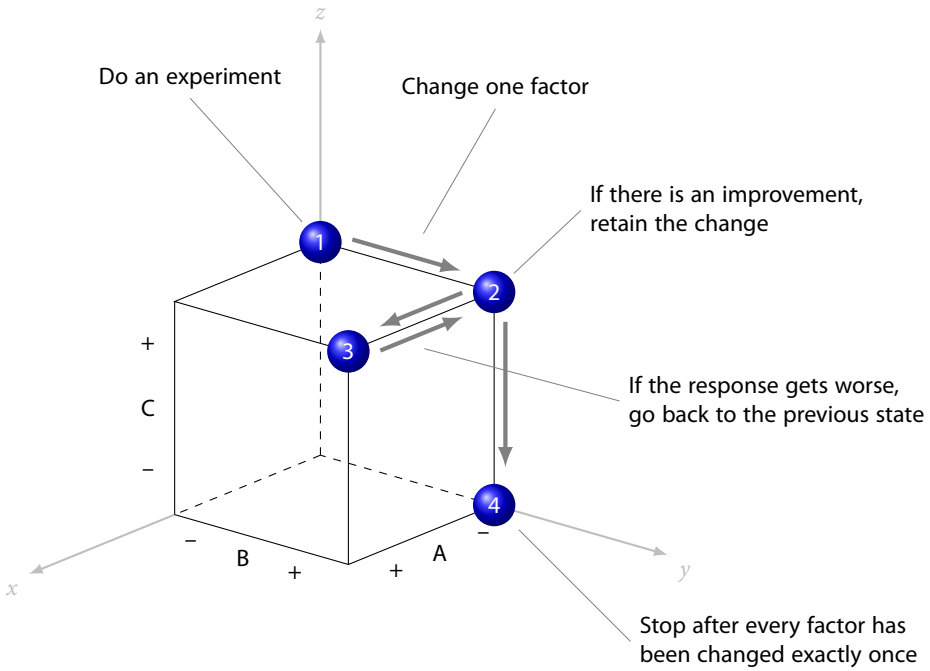


Figure 2-7: Illustration of the adaptive one-factor-at-a-time (aOFAT) algorithm, adapted from Frey and Wang (2006)

(initial configuration plus one iteration for each control factor). The performance of this algorithm can depend upon the specific starting configuration chosen and the order in which control factors are tested. Table 2.3 shows an example of using the algorithm on the catapult system. In this example, the target landing position of the ball is 96 inches from the pivot, and the starting configuration and control factor order were chosen at random with uniform probability. In the table, shading indicates control factor values that have been selected using the aOFAT method. The optimized result is the configuration arbitrarily labeled in the full factorial results as number 44, in which the ball's landing position is about 2.5 inches away from the target. From the full factorial results, the globally optimal

Table 2.3: Example Usage of the aOFAT Design Algorithm

Trial	Launch Angle	Relative Humidity	Rubber Bands	Type of Table Tennis Ball	Pullback	Ambient Temperature	Arm Material	Landing Position	Error[1]
1	60°	75%	3	Regulation	30°	72°F	Magnesium	100.81 in	−4.81 in
2	45°	75%	3	Regulation	30°	72°F	Magnesium	115.89 in	−19.89 in
3	60°	25%	3	Regulation	30°	72°F	Magnesium	100.68 in	−4.68 in
4	60°	25%	2	Regulation	30°	72°F	Magnesium	70.17 in	25.83 in
5	60°	25%	3	Large-Ball	30°	72°F	Magnesium	93.44 in	2.56 in
6	60°	25%	3	Large-Ball	40°	72°F	Magnesium	117.78 in	−21.78 in
7	60°	25%	3	Large-Ball	30°	32°F	Magnesium	90.55 in	5.45 in
8	60°	25%	3	Large-Ball	30°	72°F	Aluminum	75.87 in	20.13 in
Result	60°	25%	3	Large-Ball	30°	72°F	Magnesium	93.44 in	2.56 in

[1]The target landing position is 96 inches. Error is defined here as the target minus actual landing position.

configuration is number 96, where the landing position is at 95.13 inches – slightly under one inch from the target. This example may lead one to expect reasonable results using the aOFAT algorithm. This is addressed with a simple Monte Carlo analysis following the introduction of the Plackett-Burman L_8 method described next.

Plackett-Burman L_8 Orthogonal Array

The adaptive nature of the aOFAT algorithm prevents us from being able to specify *a priori* all configurations to be tested. However, in the Plackett-Burman approach this is precisely what is done. For seven two-level control factors, the appropriate design matrix is the Plackett-Burman L_8 orthogonal array shown in table 2.4 on the following page. In the nomenclature of experimental design methodology, this is a 2^{7-4}_{III} design.

Here one collects data for all eight trials before analyzing the results to develop a mathematical model relating the input (control factor settings) to the output (system response). The mathematical model can then be used as a surrogate in exploring the design space of the system. The main benefit is a great reduction in the size of the experiment, and the main drawback is that a good result will not be obtained if the mathematical model is not a good fit for the system.

For simplicity in this experiment, test subjects are instructed to assume that a model containing only

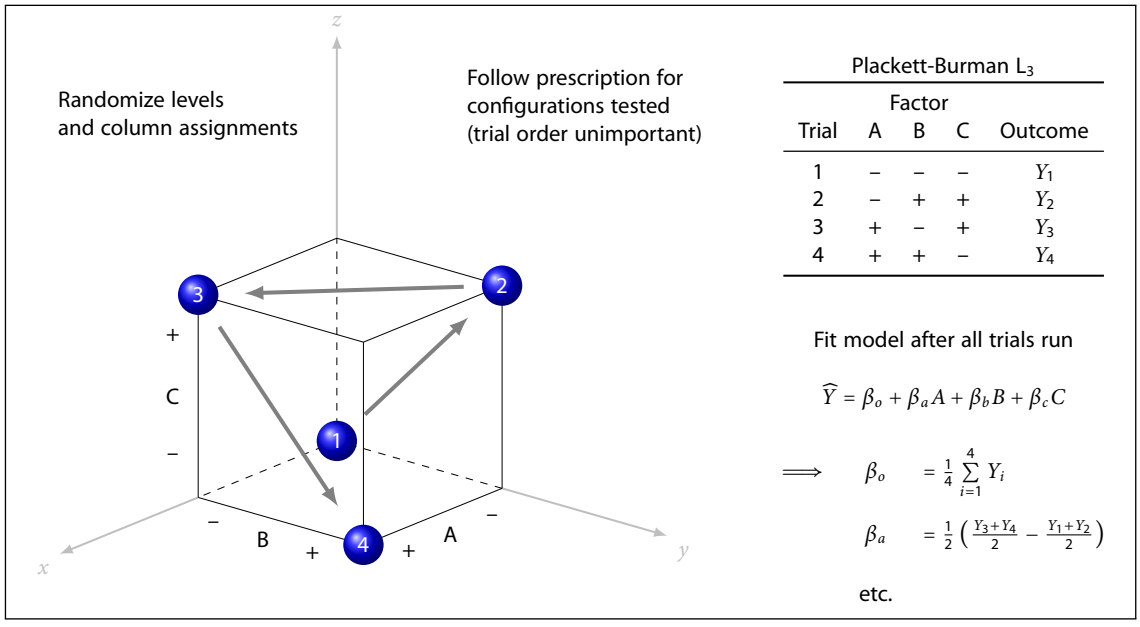


Figure 2-8: Illustration of the orthogonal-array-based, fractional factorial algorithm

Table 2.4: Plackett-Burman L_8 Orthogonal Array

Trial	Control Factor						
	A	B	C	D	E	F	G
1	−1	−1	−1	−1	−1	−1	−1
2	−1	−1	−1	+1	+1	+1	+1
3	−1	+1	+1	−1	−1	+1	+1
4	−1	+1	+1	+1	+1	−1	−1
5	+1	−1	+1	−1	+1	−1	+1
6	+1	−1	+1	+1	−1	+1	−1
7	+1	+1	−1	−1	+1	+1	−1
8	+1	+1	−1	+1	−1	−1	+1

main effects is sufficient to represent this system. In this case, the system response is approximated by

$$\widehat{Y}(X_i) = \beta_0 + \sum_{i=1}^{7} \beta_i X_i \tag{2.20}$$

where X_i represents the control factor settings for nominal $(X_i = -1)$ or alternate $(X_i = +1)$, and the constants β_0 and β_i are obtained by analyzing the Plackett-Burman experimental results as follows:

$$\beta_0 = \frac{1}{8} \sum_{j=1}^{8} Y_j \tag{2.21}$$

$$\beta_i = \frac{1}{2} \left(\underbrace{\frac{1}{4} \sum Y(X_i = +1)}_{\text{alternate mean}} - \underbrace{\frac{1}{4} \sum Y(X_i = -1)}_{\text{nominal mean}} \right), \quad \forall \ \ i \in (1, 2, \ldots, 7) \tag{2.22}$$

As with the aOFAT algorithm, the performance of this algorithm can depend upon the specific starting configuration chosen and the order in which control factors are tested. Table 2.5 shows an example of using the algorithm on the catapult system.

Table 2.5: Example Usage of the PB-L_8 Design Algorithm

Trial	Launch Angle	Relative Humidity	Rubber Bands	Type of Table Tennis Ball	Pullback	Ambient Temperature	Arm Material	Landing Position	Error
1	60°	75%	3	Regulation	30°	72°F	Magnesium	100.81 in	−4.81 in
2	60°	75%	3	Large-Ball	40°	32°F	Aluminum	95.13 in	0.87 in
3	60°	25%	2	Regulation	30°	32°F	Aluminum	53.01 in	42.99 in
4	60°	25%	2	Large-Ball	40°	72°F	Magnesium	88.24 in	7.76 in
5	45°	75%	2	Regulation	40°	72°F	Aluminum	85.54 in	10.47 in
6	45°	75%	2	Large-Ball	30°	32°F	Magnesium	76.15 in	19.85 in
7	45°	25%	3	Regulation	40°	32°F	Magnesium	147.45 in	−51.45 in
8	45°	25%	3	Large-Ball	30°	72°F	Aluminum	87.47 in	8.53 in

As in the example for the aOFAT algorithm, the target landing position of the ball is 96 inches from the pivot. The starting configuration and control factor order were chosen to be the same as in the aOFAT example. If the coefficients are calculated with (2.21) and (2.22), the resulting equation for estimating the system response as a function of control factor settings is given by

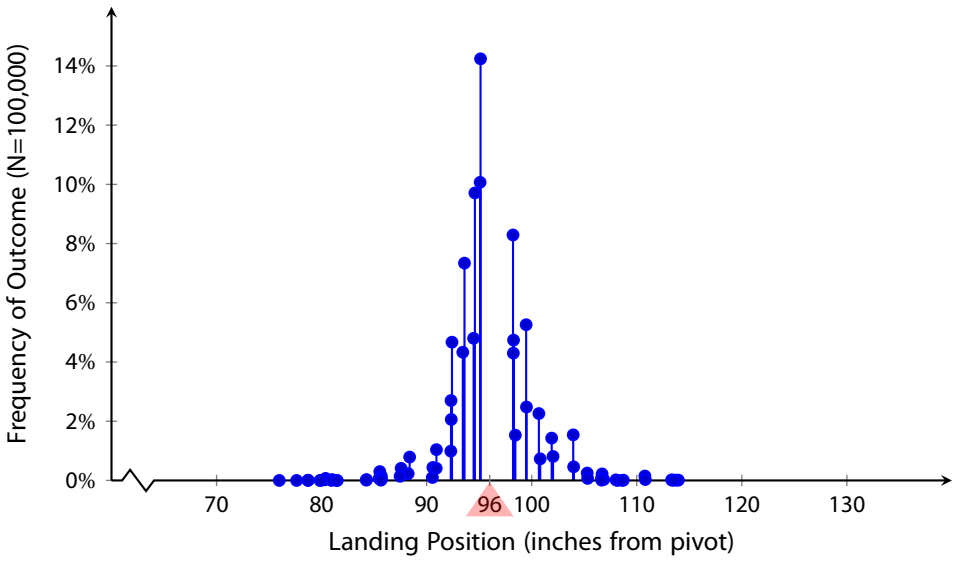
$$\widehat{Y}(X_i) = 91.73 + 7.43X_1 + 2.32X_2 - 15.99X_3 - 4.98X_4 + 12.37X_5 + 1.21X_6 - 11.44X_7 \qquad (2.23)$$

Using this equation to find the optimal configuration settings results in choosing configuration 96, which coincides with the globally optimal configuration in the full factorial. In fact, the predicted response $\widehat{Y}(X_i)$ is identical to the simulated response for configuration 96, since by chance it was one of the eight trial configurations explicitly tested in this example and $\widehat{Y}(X_i)$ is an exact fit through the eight trial results. It must be noted that in general, there is error in the predicted response $\widehat{Y}(X_i)$, the model may not return the true optimal configuration, and the optimal configuration will probably not appear in the small fraction of the full factorial that is tested. These three coincidences aside, the performance of the PB-L_8 algorithm is investigated further with the following Monte Carlo analysis.

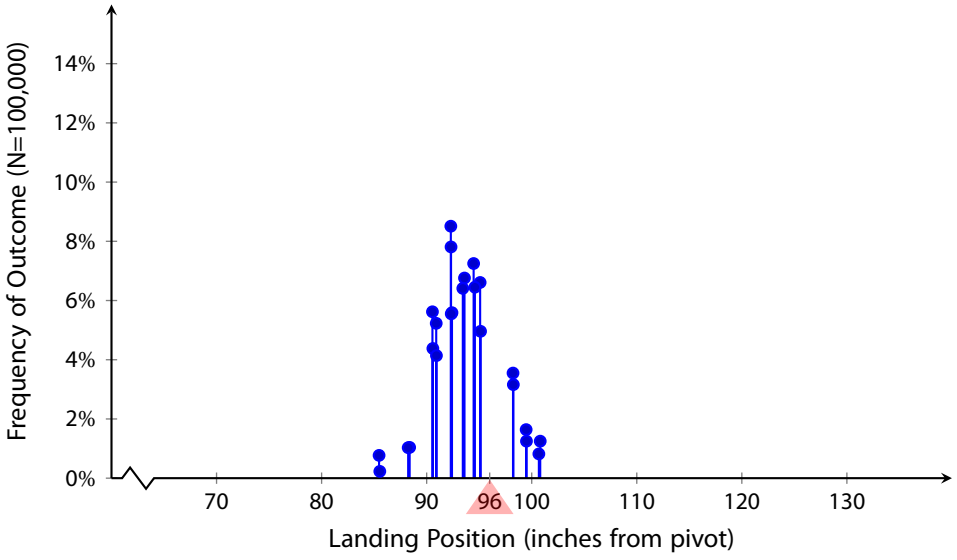
Feasibility of the Proposed Algorithms

The two examples discussed in the previous section performed well for that randomly chosen starting configuration. However, both of the algorithms are quite simple and not particularly well suited to optimization around a target value. To check the performance over a wider range of starting values, I performed a Monte Carlo simulation with 100,000 iterations. For each iteration, one of the 128 possible configurations of the 2^7 catapult is chosen at random with equal probability, then one of the two design methods described above is applied to find the optimal configuration for the chosen starting point. The results from this analysis are presented graphically as frequency plots in figure 2-9 on the following page, which shows distribution around the target landing position of 96 inches for each algorithm. There appears to be a gap between 95 inches and 98 inches, in which no optimal configurations were found. This is because no configuration exists in the full factorial with a landing position in this range.

The two methods discussed in this section appear to meet the criteria for use in this study. They are relatively simple while still remaining effective in exploring the design space for an optimal value.



(a) aOFAT method without simulation error



(b) PB-L_8 method without simulation error

Figure 2-9: Monte Carlo analysis of the proposed design methods

Randomization

Experimental bias in the two selected algorithms is ideally countered as follows:

1. For each control factor, identification of one level as nominal and the other as alternate is set at random. Note that this is probabilistically equivalent to choosing the nominal configuration at random from the 128 possible permutations of this 2^7 system.

2. For the aOFAT approach, the order in which control factors are considered is random.

3. For the PB-L_8 approach, assignment of control factors to specific columns in the L_8 design matrix is random. The order in which the trials are evaluated could also be randomized, but it would not change the result in a computer experiment with a time-invariant simulation model.

In general, the effect of randomization in experimental protocols is to trade off reduced bias with increased variance in the response. The randomization strategy described above would do this, but there is one source of bias with the potential to dominate the results: a *learning* effect. This is a well-known phenomenon, in which a human subject performing a task becomes more proficient at the task with each iteration. This implies that the reaction to the intentional flaw will likely vary with the order of the trial in which it is encountered. Intuition here suggests that if the flaw is encountered during the first trial, the participant would be less likely to recognize aberrant behavior than if it occurred during a later trial after the participant has gained experience with the design process and knowledge about the physical device from earlier trials.

The purpose of this human subjects experiment is to introduce an intentional flaw into the computer simulation results, then to assess the response of each participant to this stimulus. Each participant completes eight trials using the simulation data. For the aOFAT treatment condition, the participant has one opportunity to observe *isolated* anomalous behavior: when the control factor related to the intentional flaw is changed. Depending upon whether the setting is kept at the new value or reverts to the nominal value, there are also either 15 or 5 unique pairings of trials, respectively, in which the flawed control factor is changed along with at least one other factor. For the PB-L_8 treatment condition, there is no opportunity to observe isolated anomalous behavior, since any pair of trials in this algorithm have four control-factor changes between them. There are 16 unique pairings of trials in which the control factor related to the intentional flaw is changed; in all 16 cases, there are three additional factors changing along with the flawed factor.

In both treatment conditions, it would be beneficial to delay introduction of the flaw until the latest trial possible. For the aOFAT group, the flaw could be introduced as late as the final trial. However, the PB-L_8 group could encounter evidence of the flaw during the fifth trial at the latest. This situation would occur if the flawed control factor is assigned to the first column in the L_8 design table (see table 2.4 on page 50). To reduce the influence of the learning effect, trials for the aOFAT group should

Table 2.6: Design Tables with Column Assignments for Flawed Control Factor

(a) aOFAT

Trial	Control Factors A	B	C	D	E	F	G
1	−	−	−	−	−	−	−
2	+	−	−	−	−	−	−
3		+	−	−	−	−	−
4			+	−	−	−	−
5				+	−	−	−
6					+	−	−
7						+	−
8							+

(b) PB-L_8

Trial	Control Factors A	B	C	D	E	F	G
1	−	−	−	−	−	−	−
2	+	−	+	−	−	+	+
3	+	+	+	−	+	−	−
4	−	+	−	−	+	+	+
5	+	−	−	+	+	−	+
6	−	−	+	+	+	+	−
7	−	+	+	+	−	−	+
8	+	+	−	+	−	+	−

be ordered so that the flaw is also first encountered during the fifth trial.

There are two other phenomena to take into account in this situation, both related to cognitive ability when recalling items encountered serially. The *primacy effect* says that items at the beginning are more easily recalled. The reasoning for this is that at the start, one needs to consider a small number of items, whereas further into the task, earlier items are still being considered and are contributing to an overload of working memory. In contrast, the *recency effect* says that items at the end are more easily recalled due to a disproportionate emphasis on recent observations. Taken together, the net effect is that items in the middle are the least likely to be recalled. Thus, introducing the flaw in the middle of the trials, as proposed above, should not result in undue influence from either of these biases.

Finally, it would be beneficial to match as many common elements between the two control groups as possible. These algorithms and the modifications to the system were chosen so that the number of trials would be equal. Assignment of control factor levels as nominal or alternate is random, but held the same for all participants. For simplicity, the order of control factor consideration for the aOFAT group is from left to right across the design table. Aside from the requirement that the flawed control factor be in the fourth column, assignment of control factors to columns is random, but held the same for all participants. In a design table based on orthogonal arrays, columns may be rearranged without affecting the orthogonality of the rows. The columns in the L_8 design table are reordered such that the first column replaces the fourth column and the other columns are shuffled at random. The resultant design tables for the aOFAT and PB-L_8 algorithms to be used in this experiment are shown in table 2.6.

Table 2.6a shows a partial design table, as the aOFAT algorithm's adaptivity precludes knowledge of the full design table prior to experimentation. Since the progression through factors is specified as left to right in the design table, the empty lower left triangular region represents all settings in the table that will be determined based on experimental feedback.

The shaded columns – shown as column D in each design table – are reserved for the flawed control factor. As shown, the first change in this control factor will occur during the fifth trial for both designs. During the first four trials, the factor will be set at its nominal value. For the fifth trial, it is set at its alternate value. For trials six through eight, it can be at either value for the aOFAT design, but it will remain at the alternate value for the PB-L_8 design.

2.4.7 Summary of Conceptual Framework

The major decisions in specifying the conceptual framework for this experiment are as follows:

1. The minimum eligibility for human participants is successful completion of a basic college-level physics course required of engineering and science undergraduates.
2. The physical device to be configured in this design task is a 2^7 system derived from the Xpult catapult, a commercial product used to teach experimental design and related concepts.
3. The design task is to find a configuration, out of 128 possibilities, that will result in the ball landing nearest 96 inches from the catapult pivot, using a specified design algorithm and results from a computer simulation of the physical device.
4. The computer simulation used to generate these results contains a flaw that is not revealed to participants: the mass property of the Aluminum arm is switched with that of the Magnesium arm.
5. The treatment variable in this experiment is the design algorithm, where the options are an adaptive one-factor-at-a-time (aOFAT) approach or a simplified orthogonal-array-based approach using the Plackett-Burman L_8 design matrix (PB-L_8).

2.5 Replication, Randomization and Blocking

This section describes the fundamental methods of addressing experimental variation in the context of this human subjects experiment. *Replication*, in which the entire experiment is performed again with a new group of participants, would allow estimation of the experimental error with higher confidence. Given the large amount of resources required for one execution of this experiment, replication is a luxury that is not feasible here. *Randomization* in multiple aspects of the experimental protocol is a cornerstone in countering unknown sources of bias or variation. It will be used where practical, as described below. *Blocking*, in which experimental units (here, one human participant) are arranged into groups sharing some characteristic believed to affect the outcome, is a technique that may be used when sources of bias or variation are known. Its effectiveness is dependent upon the within-block variation being much less than the between-block variation. Since this is a new experimental design

and there are no existing studies from which to estimate potential blocking factors, it will not be used in this experiment.

A fully randomized study requires the application of randomization to the following: (1) assignment of treatments to experimental units, (2) order of application of treatments, and (3) order of measurement of responses. The first requirement is met by generating a randomized list of identification numbers to be assigned to subjects in order of participation. The first digit in the number is 0 for the aOFAT treatment condition or 1 for the PB-L_8 treatment condition. The second requirement is met by scheduling participants in the order in which they respond to the solicitation of volunteers. The notice will be sent by electronic mail to approximately 800 engineers and scientists simultaneously, and the order in which they choose to respond (if at all) is assumed to be random. The third requirement for randomization concerns the order in which responses are measured. This would apply if the experimental subjects all participated at the same time, but here the complexity of the experiment and high degree of involvement of the test administrator limits the experiment to one participant at a time. In this context, the order of measurement has no practical meaning.

2.6 Summary of Research Strategy

In this chapter, an approach to testing the main hypothesis in this study is described. To summarize in broad terms, this approach proposes a human subjects experiment using 60 working engineers to perform a design task *in vitro*. The design task is selecting one of 128 possible configurations of a catapult with seven two-level control factors in order to hit a specified target. The treatment condition in this experiment is the experimental design used by participants, with half using an adaptive one-factor-at-a-time method and the other half using a fractional factorial, orthogonal-array-based method. Unknown to the participants, the computer simulation used to evaluate the performance of the catapult contains an intentional flaw. Each participant's domain knowledge is assessed with a simple exercise before starting the design task, and the participant's ability to discern the location of the intentional flaw is assessed during debriefing after the design task. Finally, the predictor variables representing design method and domain knowledge score are used in a logistic regression of the debriefing outcome to generate estimates of the effect of each as odds ratios. To be consistent with the hypothesis, the ratio of odds of flaw discovery as a function of design method used should be different from one in the sense of statistical significance.

Chapter 3

Experiments

3.1 Introduction

This chapter describes the experimental protocol used to test the main hypothesis that choice of design method affects the ability to notice counterintuitive behavior in the results of a computer simulation. The initial protocol was executed as a pilot study using eight (8) test subjects. No statistical analysis was performed with such a small sample size; rather, the qualitative results of this pilot study were used to modify the protocol for the main study. Thus there is much redundancy between the two protocols; however, both are included in their entirety to accommodate readers who wish only to get this information from this document.

3.2 Pilot Study

3.2.1 Experimental Protocol

The format used below follows the American Psychological Association's Journal Article Reporting Standards (JARS).

Participant characteristics.

To be eligible for this study, participants must have the domain-specific knowledge necessary to understand mathematical models of simple mechanical dynamics and aerodynamics. It is assumed that one who has passed a first college-level course in physics for scientists and engineers has this knowledge.

Sampling procedures.

To recruit, the author sent e-mail invitations to a small group of technical staff members of a nonprofit, engineering design and development company that operates in a broad range of industries. Many in this group participated in a prior human subjects experiment that the author conducted in March-April, 2007. The author specifically chose candidates that he thought would be likely to volunteer. Under other circumstances, this would compromise experimental integrity; however, the purpose of this pilot study is to identify and eliminate flaws in the procedure, rather than to generate data for statistical analysis.

Eighty percent of those contacted agreed to participate in the pilot study. Since it was not publicly advertised, there was no opportunity for self selection by participants. The experiments were conducted during normal business hours in a small conference room in the main building of the company where the participants are employed. Only the participant and the author were in the room during the experiment. This pilot study was conducted on nonconsecutive dates starting on February 20, 2009, and ending on March 18, 2009. There was no promise of compensation for participating in the experiment. As an expression of gratitude, the author gave each of the participants a snack upon completion.

To meet ethics standards, the protocol for this experiment was reviewed and approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES), 77 Massachusetts Avenue, Room E25-143B, Cambridge, MA 02139. The protocol number assigned by COUHES to identify it is 0709002385. Administration of the experiment was monitored for safety by Dr. Robert Najjar, Director of Environmental Health and Safety, Charles Stark Draper Laboratory, 555 Technology Square, Cambridge, MA 02139.

Sample size, power, and precision.

The results of this pilot study were considered only qualitatively; sample power and precision are meaningless in this context. Sample size is the smallest number required to identify and fix problems in the procedure. From past experience, this number was assumed to be eight (8).

Measures and covariates.

The nuisance variable "technical ability" is measured through a survey administered after the participant grants informed consent. The dependent variable "response to evidence of error" is measured via protocol analysis (Ericsson and Simon, 1993) on the written results obtained with the Predict-Observe-Explain technique introduced by White and Gunstone (1992, p. 44):

> Prediction-Observation-Explanation, which we abbreviate to POE, probes understanding by requiring students to carry out three tasks. First they must predict the outcome of some event, and must justify their prediction; then they describe what they see happen; and finally they must reconcile any conflict between prediction and observation.

There was no explicit method for enhancing the quality of data measurement, aside from using the same test administrator for all collected data.

Research design.

Each participant is randomly assigned to a treatment condition using the following simple method. A sequence of 5-digit identification numbers is created using a random-number generator. The first digit is either 0 or 1 with equal probability and indicates whether the participant uses the aOFAT design (for 0) or the PB-L_8 design (for 1). The second through fifth digits comprise an (possibly zero-padded) integer selected with uniform probability from $[0, 9999]$ without replacement. Participants are assigned an identification number from this sequence in the chronological order in which they participate in the experiment. The author generated the list of random identification numbers and coordinated the testing schedule. Masking was not possible since the treatment variable determines which training is required.

Experimental manipulations or interventions.

All participants are first trained on the operation of the physical device under consideration, by listening to a description of the device including the name of each component of the device, the intended operation of the device, and the factors that may be changed in the design task. The graphical aids used by each participant in this pilot study include:

1. the component diagram and modeling assumptions table on page 147,
2. diagrams of the device illustrating pullback and launch angle on page 148,
3. an engineering sketch of the device with all variables labeled and a table of numerical values and units for these variables on page 151, and
4. plots of humid air viscosity and density as functions of ambient temperature and relative humidity on page 152.

The participant is then instructed to provide, for each of the seven control factors, a prediction of what will happen to the response of the device if the factor is changed from its nominal value to the alternate setting, the rationale supporting this prediction, and a level of confidence in the prediction

on a scale of 1 to 5 in order of increasing confidence. Next the participant is trained on the design method to be used, by being told the assumptions and level of fidelity of the computer simulation, then listening to an explanation of the design task.

Participants with identification numbers starting with 0 use the adaptive one-factor-at-a-time (aOFAT) method:

1. Select a starting configuration either at random or based upon *a priori* knowledge of the system.
2. Evaluate the system response at the starting configuration.
3. For each control factor in the system
   (a) Select a new configuration by using the previously evaluated configuration with the best performance, changing only this factor's setting to its alternate value.
   (b) Evaluate the system response at the new configuration.
   (c) If the performance improves at the new setting of this factor, keep it at this setting for the remainder of the experiment; otherwise, keep it at the original value for the remainder of the experiment.
4. The configuration obtained after stepping through each control factor exactly once is the optimized result for this design approach.

As an aid in understanding and implementing the aOFAT method, participants were also provided with the sheet shown on page 155. The top half is a concise summary of the design algorithm, and the bottom is the design table that is used as a worksheet while stepping through the algorithm.

Participants with identification numbers starting with 1 use the Plackett-Burman L_8 (PB-L_8) method:

1. Evaluate the system response at each of the 8 configurations prescribed by the L_8 design matrix.
2. Calculate the coefficients in the linear model used to approximate the relationship between the control factors and system response.
3. Using the linear model, find the configuration that results in the best system response.
4. Optionally, check the system response for this configuration using the simulation results.

In either case, evaluating the system response means getting the computer simulation result from a lookup table of all possible results. For the PB-L_8 design, the corresponding predicted response using the linear approximation is also provided in tabular form. This approach is taken to reduce the complexity of this experiment and the time required of each participant.

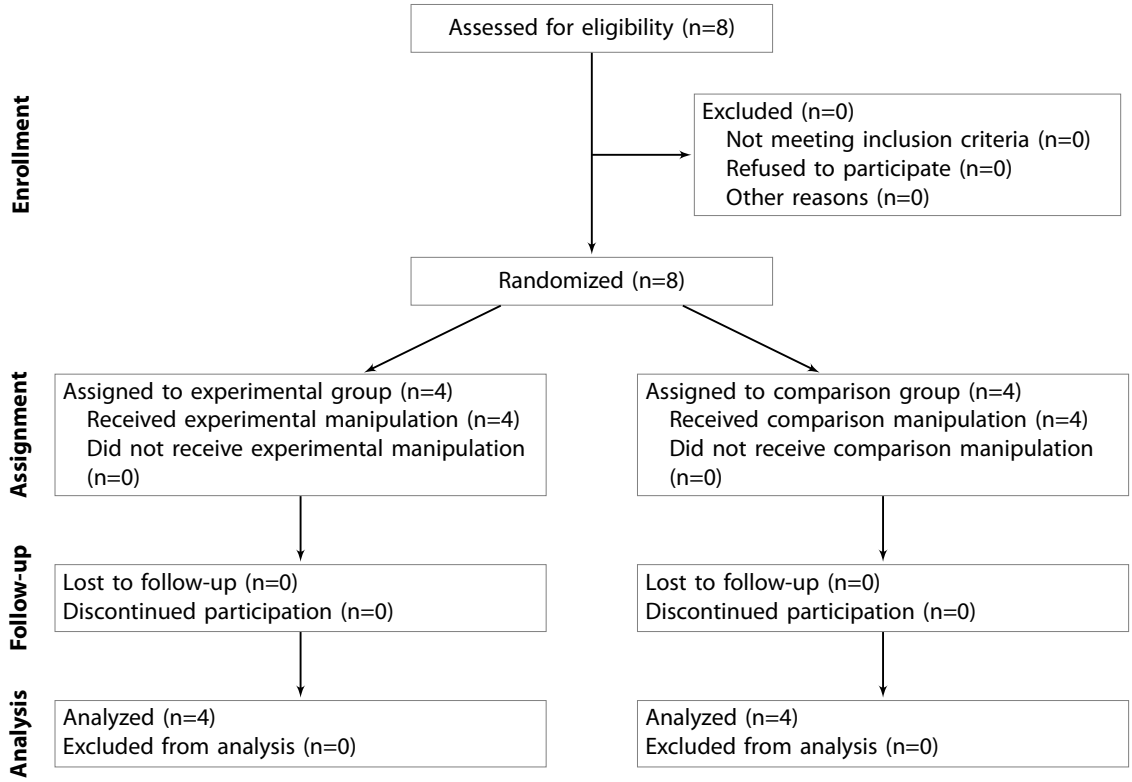


Figure 3-1: Participant flow diagram for the pilot study

Units of delivery and analysis.

Participants completed the design task individually. The qualitative analysis for this pilot test was based on both individual and group results. An example of using an individual result is where an unanticipated reaction occurs. An example of using the group result is in assessing the average time required for each participant.

3.2.2 Results

Participant demographics

The participant flow diagram for the pilot test is shown in figure 3-1. The format of the diagram is a hybrid of the participant flow diagrams proposed in the JARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) and CONSORT (Moher et al., 2001) guidelines. Eight participants successfully completed the pilot study and have the following characteristics:

1. *Gender*. 7 male, 1 female.

2. *Age*. All in the age group 25-35 years old.

3. *Education*. All hold bachelor's degrees, seven also hold master's degrees, and one is a doctoral candidate.

4. *Work experience*. Two have 2-5 years experience, five have 5-10 years experience and one has 10-15 years experience. This is full-time technical work experience beyond completion of a bachelor's degree.

5. *DoE experience*. Seven identified themselves as novices in using design of experiments methods, and one self identified as an intermediate.

6. *Simulation experience*. One self identified as a novice in using computer simulation, two were intermediates, two were experts and three were developers. Years of experience using simulation tools ranged from less than one to 10-15 years.

Data collection

One of the primary aspects of the protocol to be tested here was the collection of written data using the Predict-Observe-Explain (POE) method. The first four subjects were instructed to write down the prediction, confidence level and reasoning before being told the outcome of each trial. After the disclosure, they were asked to write down any reasons that might explain the difference between the predicted result and the observed result.

A surprising phenomenon seen in the first four subjects was that the reaction to the anomalous result witnessed by the test administrator was sometimes quite different from what could be interpreted by the written reaction. For example, one subject upon learning the anomalous result, said

> Really, huh, so it decreased?…Huh?…Huh?…What did I miss?…Huh! I'm not sure what I missed. …Hmm!…I really can't explain the discrepancy, I'd like you to explain it to me afterwards.

However, this is what the subject wrote down after expressing the above:

> Nonlinear drag effects? (No, I can't explain this.)

In this case, the verbal comments were far more revealing that the subject had indeed been surprised by the result. The written comments seem like they could apply to a reaction to a slight error in prediction, rather than a large discrepancy in an unexpected direction.

Another negative consequence of using the POE written forms is that subjects were not fully complying with the instructions to fill out the form for each trial. In some cases, no reasoning was written for the discrepancy at all. For the test administrator, having to check for compliance after each trial was

disrupting the natural flow of the subject's design experiment and adding to the time required to complete it.

Finally, having the additional paper forms for written POE responses (4 double-sided sheets) added to a cluttered work surface that was impeding a smoothly running experiment.

For the above reasons, the written approach to collecting POE data was abandoned after the first four subjects. For the second four subjects, the entire session (starting after informed consent was granted) was recorded on video. This change was approved by the Institutional Review Board and required changing the consent form signed by each participant.

Changes to graphical aids provided

Some participants attempted to calculate the length of the rubber bands as a function of pullback. While this *is* possible with the information provided, the complex geometry makes it a nontrivial calculation. Since there are no losses modeled in the system prior to launch, a good estimate of the launch velocity of the ball may be calculated if one knows the length of the rubber band at pullback and at launch. Points in between are not required for this estimate. To allow for graphical estimation of the key rubber band lengths, the six 1:5 scale drawings shown on page 187 were provided as a late addition to the pilot study (for the last four participants). The rubber band length depends only upon pullback; however, the launch angles were included here to convey more clearly the operation of the device at these control factor settings. With the 12-inch ruler ($1/32$-inch resolution) provided, participants have everything necessary to determine graphically the length of the rubber band at release and at launch.

Participants referring to the plots for humid air properties seemed to have a difficult time distinguishing between the lines of constant relative humidity. For the second group of four subjects, the lines for 0%, 50% and 100% humidity were removed from the plots since these humidity points are not included in the chosen control factor settings, the color coding of lines was removed, and the two remaining lines were coded as solid and dashed.

3.2.3 Discussion

Modifications to the experiment, based on lessons learned in this pilot study, are discussed in this section and can be divided into the areas of improving data quality, improving experimental flow, and reducing participant time.

Improving data quality

Before this pilot test, the dependent variable "test subject response to anomalous data" was thought to be best measured by protocol analysis on the written response just after first exposure to the anomalous behavior as part of the Predict-Observe-Explain technique of probing understanding. The intention was to have protocol analysts judge each response according to the Chinn & Brewer taxonomy of responses to anomalous data. Based on participants' responses, it seems that such an approach may not work well.

The first problem with the approach is that many participants, upon debriefing at the end, indicated that they did not consider the possibility that the simulation could be wrong, despite the administrator advising them to treat the simulation as if they had received it from a colleague and were running it for the first time. In nearly every case, those that displayed signs of expectation violation assumed that their own reasoning must have been incorrect. In the proposed dichotomous organization of meta response categories, this response would be classified the same as if the test subject had received results that were expected.

The second problem with this approach is that there was a discrepancy between the verbal response and the written response. This was addressed halfway through the pilot testing by adding video recording of the participant to the protocol. Any protocol analysis intended to be performed on the written response would be performed on the video recording instead.

A key reason for judging participant response to every simulation result received was that it was thought to be better to obtain some measure of suspicion right after the anomalous data is first encountered, rather than waiting until the end when the response may be forgotten. However, for the pilot participants this seemed not to be the case. During the debriefing, there were two questions asked before revealing the location of the flaw:

1. The administrator asked the participant, "Did you think the simulation results were reasonable?"
2. Those who expressed doubt were asked to pinpoint the area of concern: "Which of the control factors do you think the problem is tied to?"
3. To those who did not express doubt, the administrator then said, "There is a problem in the simulation. Knowing this now, which of the control factors do you think the problem is tied to?"

A dependent variable based on the responses to the above questions would seem to be less ambiguous than results based on verbal protocol analysis. Each participant would be classified according to the following:

1. Participant provided the correct location of the flaw without being told there is a problem.

2. Participant provided the correct location of the flaw after being told there is a problem.

3. Participant did not provide the correct location of the flaw at any time.

Moving forward to the main experiment then, there will be data used to generate two different dependent variables recorded:

1. The verbal reaction to learning the result of each trial. This is worth analyzing in the future in an exploratory manner to determine if the Chinn & Brewer taxonomy may be useful in future experiments.

2. A categorical variable "point at which the participant correctly identifies the location of the flaw" with categories "without being told there is a problem", "after being told there is a problem", and "never". This variable will be used to judge the success or failure of the hypothesis test. It does not change the sample size required.

Another benefit of moving the success criteria to the end of the experiment is that it measures participant response based on the entire design exercise rather than a small snippet. One of the probable criticisms of judging participant response based solely on the reaction to a single trial is that it biases the results in favor of the aOFAT algorithm. Participants using aOFAT have an opportunity to observe the anomaly in isolation: when comparing the two trials that differ by only the problematic control factor. Participants using PB-L_8 cannot observe the anomaly in isolation during progression through trials; however, when the main effect coefficients are calculated after all trials are completed, an astute participant may notice that the sign of the coefficient for the problematic control factor effect is opposite the value implied by the expected behavior. To repeat, moving the success criteria to the end of the experiment would be a fairer comparison.

Improving experimental flow

Experimental flow in this context means how quickly and completely does the participant understand the physical device operation and the design task, how easily does the experiment transition from one trial to the next, etc. Improvement here is attained through a simplified and clearer presentation of graphical aids.

In the pilot study, there were five separate reference sheets:

1. A component diagram and modeling assumptions table, held by the administrator.

2. Two diagrams showing the difference between launch angle and pullback, held by the administrator.

3. An engineering sketch, a modeling parameters table and a control factors table, held by the test subject.

4. Six scale drawings of the catapult in positions of interest to determine graphically the rubber band stretch, held by the test subject.

5. Humid air property plots, held by the test subject.

These were condensed into one 17×11-inch sheet of paper by making the following changes:

1. On one side is the component diagram – a sketch with all components labeled according to the nomenclature to be used throughout the experiment. This side of the reference sheet is shown on page 211 in Appendix B.

2. On the other side is the table of modeling assumptions, a table showing the control factor settings, and a ½-scale drawing showing control factor options and all non-geometric modeling data. This side of the reference sheet is shown split over pages 212 and 213 in Appendix B.

Having all of the reference material on one oversized, double-sided sheet avoids much of the clutter and subsequent confusion in the experiment.

Some of the information presented in the pilot study should really be coming from the participant's own domain knowledge instead. The following changes were made to remedy this:

1. The humid air property plots are excluded.

2. The ball nomenclature was modified. The "regulation table tennis" ball is now referred to as the "white" ball, and the "large-ball table tennis" ball is now referred to as the "orange" ball. This change makes it easier to identify which ball is being discussed, and it removes a potential emphasis on the size of the ball over the mass of the ball.

3. The six scaled drawings for graphical estimation of rubber band length are excluded. The same information can be estimated from the provided ½-scale drawing with the proper domain knowledge.

Reducing participant time

One of the primary objectives of this pilot study was to determine how to conduct the experiment, particularly the more complex PB-L_8 treatment condition, in a reasonable amount of time. The experiment requires a relatively large sample size ($n \approx 50$) from a relatively small population ($N \approx 800$), and one may reasonably assume that recruiting would be easier for a shorter time commitment. In addition, the complexity difference between the methods – an unavoidable consequence of the

main hypothesis – implies that the time commitment will be asymmetrical depending upon treatment condition.

A measure of the time required to teach naïve subjects how to perform a simple Taguchi method using the well-known paper helicopter experiment is claimed by Antony and Antony (2001) to be about six hours. For this experiment, each participant will have about one hour. Although the difference between the six-hour estimate and the one-hour requirement is large, the gap can be closed by focusing on bottlenecks identified in the pilot run. Also, Antony teaches a method that considers variance through replication, the subjects are in a group as opposed to one on one, and the subjects must physically construct the prototype for each trial.

Four main categories of bottlenecks were identified during the pilot run.

1. *Assisting table lookup.*

   During the pilot test, some of the test subjects frequently referenced the L_8 design table (i.e. the table filled with "−1" and "+1") while working through the design table filled with the actual control factor settings. These were on opposite sides of the same sheet of paper, and it became a distraction to the subject. This problem was addressed by superimposing the "−1" and "+1" values onto the working design table with a diagonal offset, smaller font, and different color. The resulting table is included on page 222 in Appendix B.

   While the Predict-Observe-Explain worksheets were removed for another reason as previously discussed, this also solved the problem of subjects frequently switching between these worksheets and the working design table to compare configurations.

   In providing simulation results to the subject, the administrator looks up the values in a table of all possible results. During the pilot testing, this was done by asking the subject to state verbally each configuration option for the administrator to find the correct value. This turned out to be time consuming and prone to error. To solve this problem, a marker with the design method and trial number was placed in the margin next to the corresponding result, as shown in the table starting on page 225 in Appendix B.

2. *Providing timely guidance.* Although they were not specifically instructed to do so, most subjects talked out loud as they were stepping through the design. In some cases, the administrator overheard facts regarding the modeling assumptions that were incorrect. For example, one subject was considering the temperature dependence of the rubber band response, even though the administrator explained during training that it would be modeled as an ideal spring. To prevent a mistake due to failure to recall the assumptions, the administrator will repeat the relevant modeling assumption when necessary.

   One calculation that seemed to delay progress was that of the rubber band displacement as a

function of pullback. All of the information necessary to make this calculation is provided, but the geometry involved makes it a non-trivial problem. For example, one test subject attempting this required 45 minutes of calculation time. For the main experiment, the administrator will note that this may be roughly estimated graphically using the scale drawing and provided ruler.

3. *Giving an anchor.* It is well known that when making numerical predictions such as these, one of the main strategies is to *anchor and adjust* (Tversky and Kahneman, 1974). In that scenario, a more easily remembered value (the anchor) is first generated and then an estimate is made by adjusting from that value. Subjects in the pilot study seemed to have a particularly difficult time in predicting the value for the first trial. Since the analysis requires only the subject's response at the end of the experiment and in the transition between trials, it makes sense to provide this first value up front. Subjects can then make adjustments to the provided value for the remaining seven trials.

4. *Assisting in PB-L_8 design calculations.* The design calculations for the PB-L_8 algorithm are not difficult, but they are tedious, time consuming and prone to human error when performed by hand. To make this part of the process quicker and more reliable, several changes were made:

   (a) A worksheet for the main effects coefficient calculations is provided. This worksheet is shown on pages 231 and 232 in Appendix B.

   (b) This calculation is now optional, but recommended if the participant does not fully understand how the coefficients are calculated. If the participant declines the manual calculations, or after they are performed, all of the correct values for these coefficients are provided by the administrator using the completed worksheet shown on pages 235 and 236 in Appendix B.

   (c) Instead of having the participant explore the design space manually with this main effects equation, a precalculated table of all 128 configurations is provided. The table is sorted by increasing distance from the target position, and is shown on pages 243 and 244 in Appendix B.

   These changes take the tedium and chance for mistakes out of the process, and allow the participant to focus on understanding and applying the design method.

Conclusion

Making these changes to the experimental protocol should enhance the data quality through a more objective and fair measurement of the dependent variable, a clearer presentation of the device and design task, and speedier completion for each participant.

3.3 Main Study

3.3.1 Experimental Protocol

The format used below follows the American Psychological Association's Journal Article Reporting Standards (JARS).

Participant characteristics.

To be eligible for this study, participants must have the domain-specific knowledge necessary to understand mathematical models of simple mechanical dynamics and aerodynamics. It is assumed that one who has passed a first college-level course in physics for scientists and engineers has this knowledge.

Sampling procedures.

To recruit for this procedure, a person other than the author contacted approximately 750 potential study participants by email, describing the experiment and inviting those interested to contact the author directly. All of those contacted are technical staff members of a nonprofit, engineering design and development company that operates in a broad range of industries. The person who contacted them is responsible for all educational activities at the company and is highly visible but not in a position of authority over any of those solicited. Approval was sought and granted by the Institutional Review Board prior to this action.

56 out of approximately 750 contacted ($\approx$ 7.5%) agreed to participate in the study. Since it was not advertised in any other manner, there was no opportunity for self selection by participants. The experiments were conducted during normal business hours in a small conference room in the main building of the company where the participants are employed. Only the participant and test administrator were in the room during the experiment. This study was conducted on nonconsecutive dates starting on April 21, 2009, and ending on June 9, 2009. Participants in the study were offered an internal account number to cover their time; most accepted but a few declined. Those who completed the experiment (55 out of 56) were offered either a cookie or voucher for a drink in the company's cafeteria.

To meet ethics standards, the protocol for this experiment was reviewed and approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES), 77 Massachusetts Avenue, Room E25-143B, Cambridge, MA 02139. The protocol number assigned by COUHES to identify it is 0709002385. Administration of the experiment was monitored for safety by Dr. Robert Najjar,

Director of Environmental Health and Safety, Charles Stark Draper Laboratory, 555 Technology Square, Cambridge, MA 02139.

Sample size, power, and precision.

The statistical power analysis was detailed in §2.3. It was determined that a sample size of 25 per each of the two treatment groups would be required to resolve a 3-in-5 chance of becoming aware of a problem versus a 1-in-5 chance. The error rates in this analysis were set at 5% for Type I error ($\alpha = 0.05$) and 10% for Type II error ($\beta = 0.10$). If the assumptions are correct, this means that the experiment is powered at 90% to resolve the stated effect size. Another consideration is the minimum event count for a good logistic regression fit, and this was assumed to be 10 per covariate in the analysis in §2.3.

Measures and covariates.

The primary dependent variable is the response to debriefing at the conclusion of the experiment. Each participant is asked a series of questions:

1. Did you think the simulation results were reasonable?

2. Those who expressed doubt were asked to pinpoint the area of concern: "Which of the control factors do you think the problem is tied to?"

3. To those who did not express doubt, the administrator then said, "There is a problem in the simulation. Knowing this now, which of the control factors do you think the problem is tied to?"

If the participant's answer to the second question is *arm material*, *catapult arm*, *bar*, *Magnesium versus Aluminum*, etc. – any wording that unambiguously means the control factor for choice of arm material in the catapult – then the participant is categorized as being aware of the issue ($Y = 1$); otherwise, the participant is categorized as not becoming aware of the issue unprompted ($Y = 0$). These categorizations are made by the author using the video recording of the debriefing portion for each participant. This is essentially a forced choice verbal response.

The exploratory nuisance variable "technical ability" is measured through a written survey administered after the participant grants informed consent, and an exercise to test domain knowledge needed in this experiment.

Research design.

Each participant is randomly assigned to a treatment condition using the following simple method. A sequence of 5-digit identification numbers is created using a random-number generator. The first digit is either 0 or 1 with equal probability and indicates whether the participant uses the aOFAT design (for 0) or the PB-L_8 design (for 1). The second through fifth digits comprise an (possibly zero-padded) integer selected with uniform probability from $[0, 9999]$ without replacement. Participants are assigned an identification number from this sequence in the chronological order in which they participate in the experiment. The author generated the list of random identification numbers and coordinated the testing schedule. Masking is not possible since the treatment variable determines which training is required.

Experimental manipulations or interventions.

All participants are first trained on the operation of the physical device under consideration, by listening to a description of the device including the name of each component of the device, the intended operation of the device, and the factors that may be changed in the design task. The graphical aids used by each participant in this study include:

1. The catapult diagram with components labeled according to nomenclature to be used in the experiment, shown on page 211 in Appendix B.

2. The consolidated model reference sheet shown on pages 212 and 213 in Appendix B.

The participant is then instructed to provide, for each of the seven control factors, a prediction of what will happen to the response of the device if the factor is changed from its nominal value to the alternate setting, the rationale supporting this prediction, and a level of confidence in the prediction on a scale of 1 to 5 in order of increasing confidence. Next the participant is trained on the design method to be used, by being told the assumptions and level of fidelity of the computer simulation, then listening to an explanation of the design task.

Participants with identification numbers starting with 0 use the adaptive one-factor-at-a-time (aOFAT) method:

1. Select a starting configuration either at random or based upon *a priori* knowledge of the system.

2. Evaluate the system response at the starting configuration.

3. For each control factor in the system

    (a) Select a new configuration by using the previously evaluated configuration with the best performance, changing only this factor's setting to its alternate value.

(b) Evaluate the system response at the new configuration.

(c) If the performance improves at the new setting of this factor, keep it at this setting for the remainder of the experiment; otherwise, keep it at the original value for the remainder of the experiment.

4. The configuration obtained after stepping through each control factor exactly once is the optimized result for this design approach.

As an aid in understanding and implementing the aOFAT method, participants were also provided with the sheet shown on page 217 in Appendix B. The top half is a concise summary of the design algorithm, and the bottom half is the design table that is used as a worksheet while stepping through the algorithm.

Participants with identification numbers starting with 1 use the Plackett-Burman L_8 (PB-L_8) method:

1. Evaluate the system response at each of the eight configurations prescribed by the L_8 design matrix.
2. Calculate the coefficients in the linear model used to approximate the relationship between the control factors and system response.
3. Using the linear model, find the configuration that results in the best system response.
4. Optionally, check the system response for this configuration using the simulation results.

In either case, evaluating the system response means getting the computer simulation result from a lookup table of all possible results. For the PB-L_8 design, the estimated responses for the resulting linear model are also provided in tabular form. This approach is taken to reduce the complexity of this experiment and the time required of each participant.

Units of delivery and analysis.

Participants were tested individually and in person by the author.

3.3.2 *Results*

Participant demographics

The participant flow diagram for the main experiment is shown in figure 3-2 on the next page. The format of the diagram is a hybrid of the participant flow diagrams proposed in the JARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) and

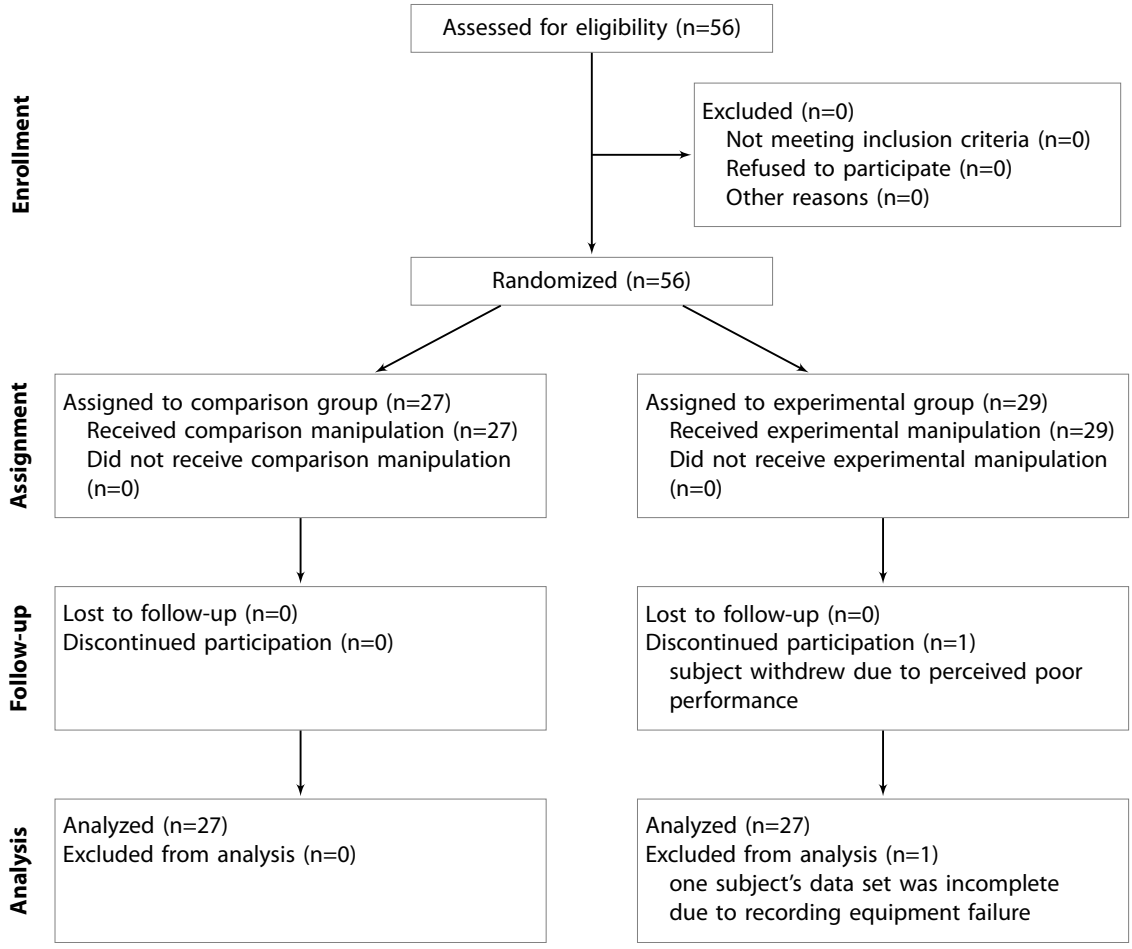


Figure 3-2: Participant flow diagram for the main study

CONSORT (Moher et al., 2001) guidelines. Fifty-six (56) engineers volunteered to participate in this study. Of these, two data sets were discarded – one subject withdrew from the experiment and there was a recording equipment malfunction for another subject. Demographic characteristics for the 54 participants who successfully completed the study are shown in figure 3-3 on the following page.

The distribution of participants was fairly even in most categories except for gender, for which nearly all participants (51 of 54) were male.

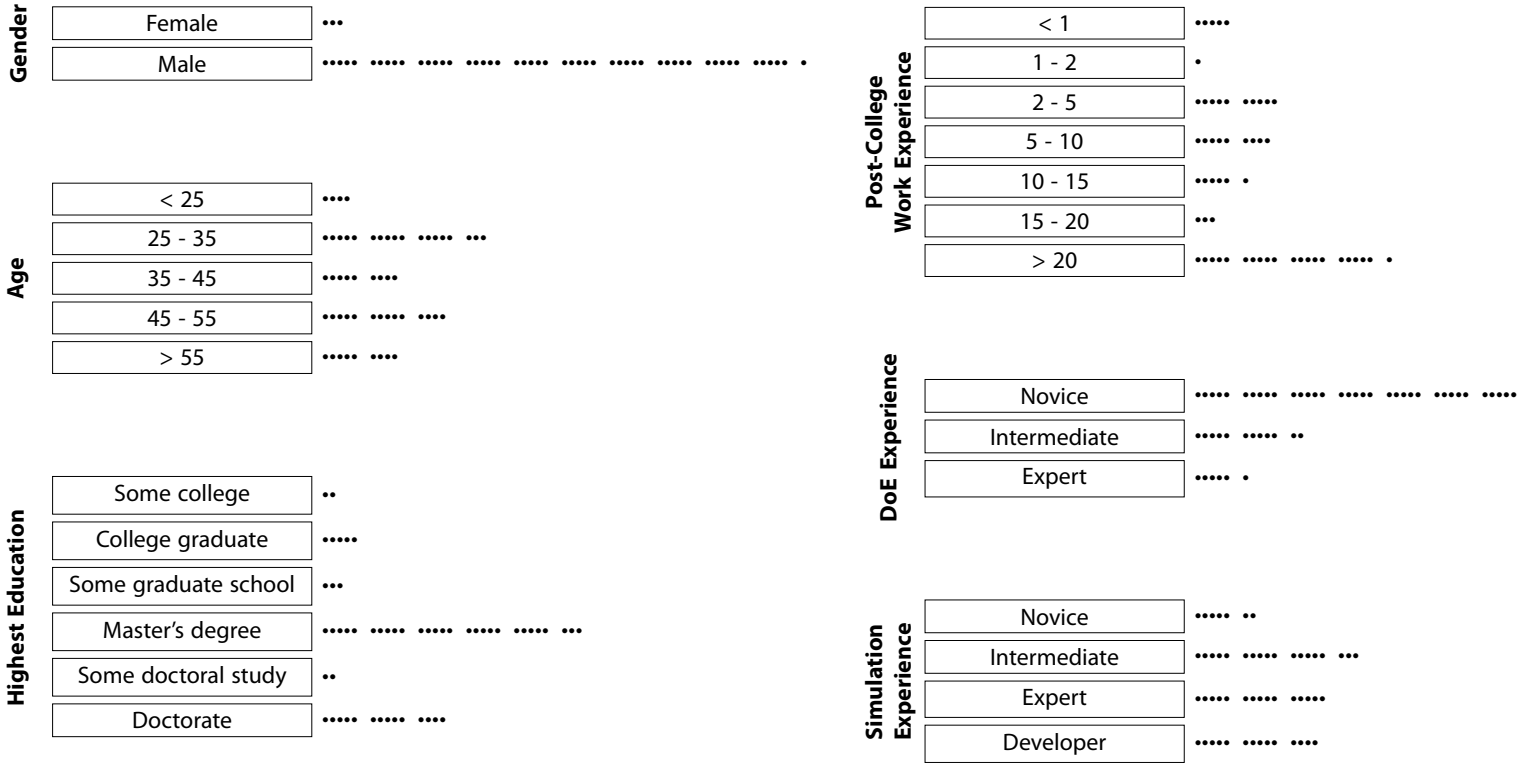


Figure 3-3: Demographics for the main study. Each dot represents one participant.

Chapter 4

Analysis

4.1 Measure of Success in Identifying Anomaly

The primary dependent variable to be analyzed is the debriefing result, one value for each of the 54 valid participants. Disregarding the confounding factor *domain knowledge* for the moment, the results are shown in the 2×2 contingency table given as table 4.1.

Table 4.1: Debriefing Results (Raw Counts)

		Identifies Anomaly?	
		No	Yes
Condition	aOFAT	13	14
	PB-L_8	26	1

4.1.1 *Using Simple Proportions without Domain Knowledge*

The first and most important effect to be considered is the proportion of each group that identifies the anomaly without being told of its existence. For each of the values in the contingency table, the binomial proportion and its 95% confidence interval are calculated and shown in table 4.2 on the following page. The confidence intervals in this table were calculated with the so-called "exact" method introduced by Clopper and Pearson (1934). This method is called exact because it is based directly on the binomial probability distribution rather than an approximation. The lower bound is $p_0 = 0$ when $x = 0$, or the solution for p_0 in the equation

$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1 - p_0)^{n-k} = \alpha/2 \tag{4.1}$$

Table 4.2: Debriefing Results (Proportions with 95% Confidence Intervals)

		Identifies Anomaly?	
		No	Yes
Condition	aOFAT	0.48 (0.29,0.68)	0.52 (0.32,0.71)
	PB-L_8	0.96 (0.81,1.00)	0.04 (0.00,0.19)

when $x > 0$. Similarly, the upper bound is $p_0 = 1$ when $x = n$, or the solution for p_0 in the equation

$$\sum_{k=x}^{n} \binom{n}{k} p_0^k (1 - p_0)^{n-k} = \alpha/2 \tag{4.2}$$

when $x < n$. So for the Clopper-Pearson method, the $1 - \alpha$ confidence bounds are given by

$$p_{1-\alpha,low} = \begin{cases} 0 & \text{if } x = 0, \\ \left[1 - \frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} & \text{if } x > 0. \end{cases} \tag{4.3}$$

$$p_{1-\alpha,high} = \begin{cases} \left[1 - \frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1} & \text{if } x < n, \\ 1 & \text{if } x = n. \end{cases} \tag{4.4}$$

with nomenclature adopted from Agresti and Coull (1998) such that $F_{a,b,c}$ is the $1 - c$ quantile from the F distribution using a and b for degrees of freedom.

4.1.2 Using Logistic Regression with Domain Knowledge

In the literature survey, domain knowledge was identified as a probable confounding factor. To control for this, each subject was quizzed on the relative effect magnitude and direction for each of the seven control factors in the catapult, when changed from its nominal setting to its alternate setting. A simple index representing domain-specific knowledge level can be constructed by assigning one point for each control factor prediction in the correct direction. The magnitude predictions are subjective and are not used for this index. For a recent example of this approach to assessing domain-specific knowledge in a human-subjects experiment involving an engineering design task, see Klahr et al. (2007).

Distribution of the Domain Knowledge Score

Before usage in the regression for predicting debriefing result, the domain knowledge score is considered alone to determine whether a bias may be introduced by an inequitable distribution of expertise between the two groups of subjects. The approach taken here is to assume a probability distribution for each frequency response, formally test the assumption for each, then perform the appropriate comparison between the two distributions.

There are 27 test points from each of the two participant groups – one for each valid participant data set. The central limit theorem states that for a sample size greater than about 30, the sampling distribution of the mean is approximately normal (Spiegel and Stephens, 1998, p. 182). However, an interest in the distribution of the scores themselves rather than the distribution of the mean of a sample of the scores precludes use of the central limit theorem in this situation. If a normal distribution is assumed, then a test of normality is necessary to verify the assumption.

Many tests of normality are available, each with various strengths and weaknesses. The test used here is the D'Agostino-Pearson K^2 omnibus test, which has the advantages of addressing skewness and kurtosis and performing well for sample sizes in the range of $20 \leq n \leq 200$ (D'Agostino and Pearson, 1973). The K^2 statistic in this test is approximately χ^2-distributed with two degrees of freedom. For a Type I error rate $\alpha < 0.05$, the critical value of K^2 is 5.99. For the domain knowledge scores in the aOFAT group, K^2 is 1.267. For the PB-L_8 group, K^2 is 4.899. In both cases, the null hypothesis cannot be rejected at the 0.05 level, and the null hypothesis in this case is that the distribution is normal.

Figure 4-1 shows the domain knowledge scores as separate histograms for each test group, along with the distributions that fit these data. Now that it is known that the normality assumption likely holds, a two-sample t test may be used to test for equal means. The test statistic is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.5}$$

where $\bar{x}_i$ is the sample mean, s_i is the sample standard deviation, and n_i is the number of samples. For the domain knowledge score data, the test statistic was computed as $t = 1.553$. Using a two-tailed test and the t distribution with $n_i - 1 = 26$ degrees of freedom, the critical value for a significance level of 0.05 is 2.056. For t values greater than the critical value, the data from the two samples are unlikely to be observed if the distributions are the same. Since the test statistic here is less than the critical value, the null hypothesis that the means are equal cannot be rejected.

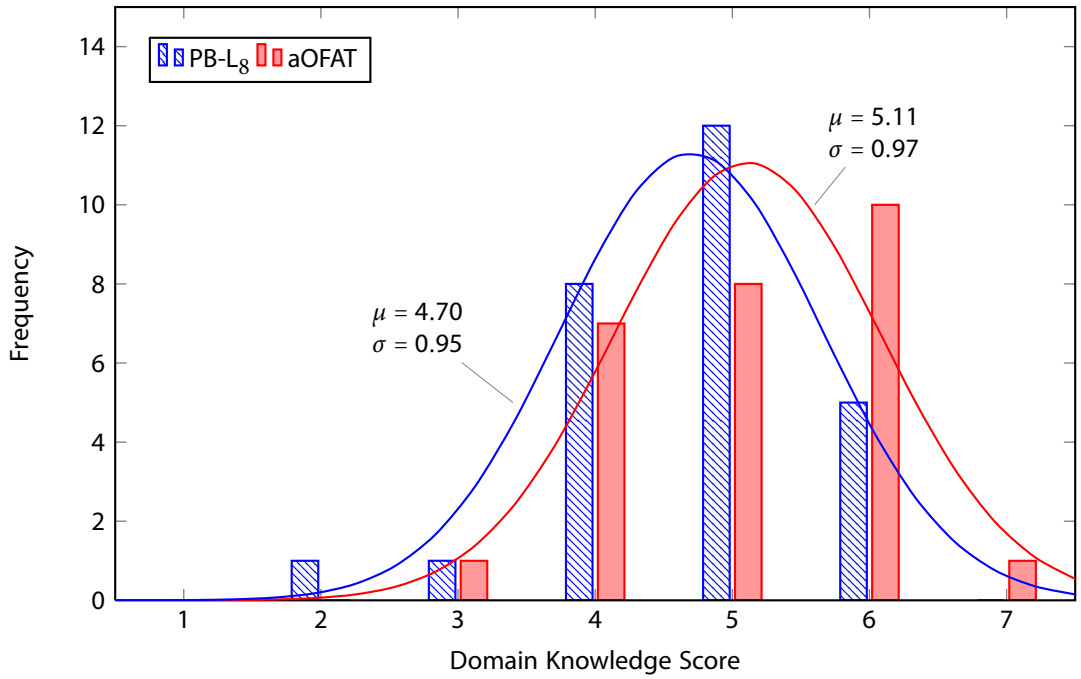


Figure 4-1: Histograms of domain knowledge score by treatment group

Performing the Logistic Regression

Table 4.3 shows the result of this regression. The two main effects were found to be significant, but the interaction between the effects was not. Interpretation of these results is in the next chapter.

Table 4.3: Logistic Regression Coefficients for Subject Debriefing

Explanatory Variable	Estimate	Standard Error	P-value	95% CI	Odds Ratio
Intercept	−0.105	0.438	0.810	(−0.965, 0.754)	–
X_{DM}	−3.518	1.155	0.002	(−5.782, −1.255)	(0.003, 0.285)
X'_{DKS}	1.100	0.471	0.020	(0.176, 2.024)	(1.193, 7.564)

As a measure of model fit, the coefficient of determination recommended by Menard (2000) was calculated by

$$R_L^2 = \frac{G_M}{D_0}. \tag{4.6}$$

In this equation, D_0 is the *null deviance* of the model calculated by −2 times the log likelihood of the intercept-only model. G_M is the difference between D_0 and −2 times the log likelihood of the model with explanatory variables in place. R_L^2 ranges in value from 0 for a completely ineffective model to 1 for a model that perfectly predicts the observed values. For this regression, D_0 was 64.4 and G_M was 39.3, resulting in $R_L^2 = 0.390$.

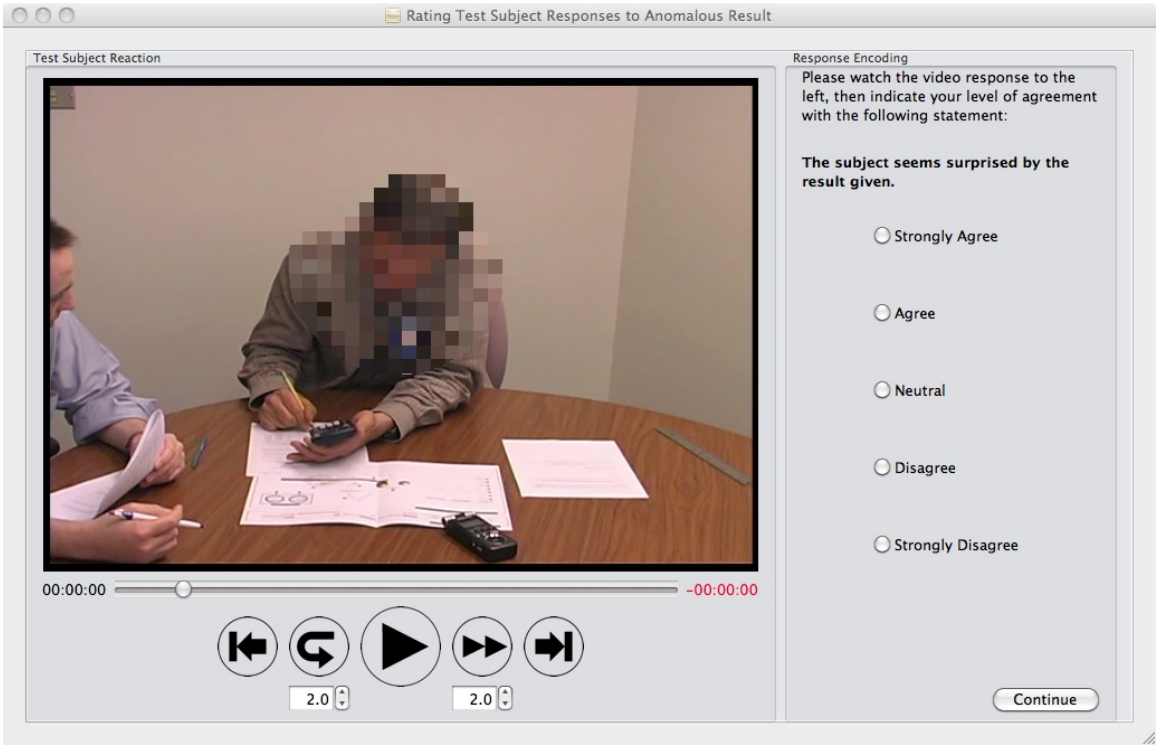


Figure 4-2: Graphical user interface for rating participant reactions. The image has been digitally altered to protect the identity of the participant, but the video viewed by the independent raters was not obscured in this manner.

4.2 Measure of Expectation Violation

One measure of participant behavior in reaction to the numerical simulation result is whether the participant expresses surprise, a well-known manifestation of expectation violation. As this is a subjective measure, it was obtained through ratings by independent judges. Two analysts, with certification in "MIT Research Involving Human Subjects" and prior experience in a similar study, were hired for this task. The analysts worked alone, viewing the video recordings on a notebook computer using a custom graphical user interface that randomized the order in which responses were shown, enforced viewing the entire response before entering a rating, and collected each rating on a five-point Likert scale. An example screen from this custom interface is shown in figure 4-2.

To use the rating data in any analysis, disagreements between raters must first be resolved. The rule of agreement used here is shown in figure 4-3 on the next page, where the outcome variable is collapsed into a binomial in which $Y_s = 1$ indicates that the raters agreed that the subject seemed surprised and $Y_s = 0$ indicates that the raters agreed that the subject did not seem surprised.

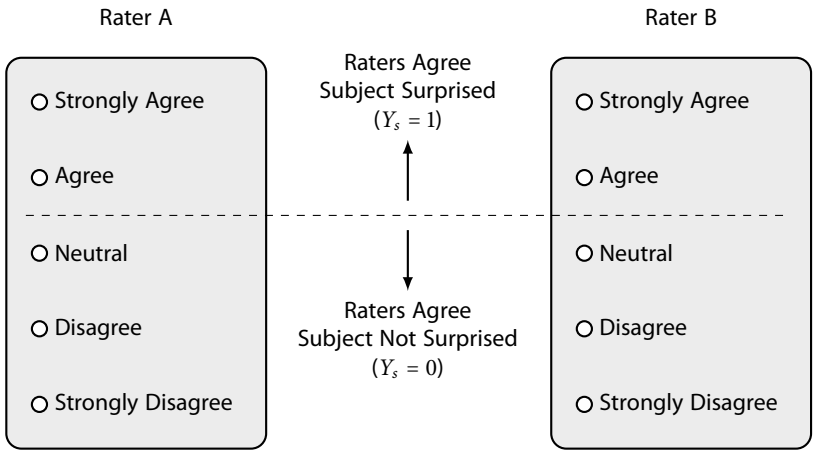


Figure 4-3: Decision rule for resolving disagreements between independent video raters. Combinations that cross the dashed line are discarded.

4.2.1 Interrater Reliability

There are numerous measures of interrater reliability in the literature (Lombard et al., 2002; Sim and Wright, 2005). Two are reported here: percent agreement and Cohen's kappa. Percent agreement is the most widely reported measure in many fields (Lombard et al., 2002), with the advantages of being intuitively appealing and trivial to calculate. Critics of percent agreement assert that it is not conservative enough since it does not account for the chance of random agreement. Cohen's kappa is the most widely reported measure in behavioral research and does take chance agreement into account, but it has been criticized as being too conservative (Perreault, Jr and Leigh, 1989).

Two analysts each viewed and rated 385 video recorded reactions. The analysts agreed, according to the decision rule in figure 4-3, on 319 of these. The agreement was therefore approximately 82.9%. It should be noted that the raters agreed on 157 of 189 (83.1%) reactions for the aOFAT treatment group, and they agreed on 162 of 196 (82.7%) reactions for the PB-L_8 treatment group.

The second measure of interrater reliability is Cohen's kappa (Cohen, 1960). The equation for it is

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{4.7}$$

where P_o is the proportion of agreement in the ratings and P_c is the probability of chance agreement between the raters. P_c is calculated by assuming that each rater has an intrinsic tendency to answer one way or the other, estimable from the observed marginal proportions. For two raters, the equation for calculating this is

$$P_c = P_A(Y_s' = 1) \cdot P_B(Y_s' = 1) + P_A(Y_s' = 0) \cdot P_B(Y_s' = 0) \tag{4.8}$$

Table 4.4: Contingency Table for Independent Analysts' Ratings of Subject Reactions

		Rater A		
		$Y'_s = 1$	$Y'_s = 0$	
Rater B	$Y'_s = 1$	148/385 (0.3844)	26/385 (0.0675)	174/385 (0.4519)
	$Y'_s = 0$	40/385 (0.1039)	171/385 (0.4442)	211/385 (0.5481)
		188/385 (0.4883)	197/385 (0.5117)	385/385 (1.0000)

where $P_A(Y'_s = 1)$ is the proportion that rater A chooses "strongly agree" or "agree"; $P_B(Y'_s = 1)$ is the proportion that rater B chooses "strongly agree" or "agree"; $P_A(Y'_s = 0)$ is the proportion that rater A chooses "neutral", "disagree" or "strongly disagree"; and $P_B(Y'_s = 0)$ is the proportion that rater B chooses "neutral", "disagree" or "strongly disagree". This notation uses Y'_s to indicate the response by a single rater that would result in a resolved value of Y_s if the decision rule for agreement is satisfied. For this data set, these values are shown in table 4.4. Substituting into (4.8), P_c is calculated to be 0.5011. Finally, (4.7) may be used to calculate $\kappa = 0.656$.

Cohen (1960) gives the standard error for κ as

$$\sigma_\kappa = \sqrt{\frac{P_o\,(1 - P_o)}{N\,(1 - P_c)^2}} \tag{4.9}$$

where N is the number of samples per rater. This is calculated as $\sigma_\kappa = 0.0385$, resulting in a 95% confidence interval for κ of $(0.5805, 0.7315)$.

To place the value of κ into context, one must also calculate the maximum achievable value κ_m, which depends on the relative tendencies of the raters according to

$$\kappa_m = \frac{P_{oM} - P_c}{1 - P_c} \tag{4.10}$$

where P_{oM} is the sum of the minimum marginal proportions by category. Using the data in table 4.4,

$$P_{oM} = \min(0.4883, 0.4519) + \min(0.5117, 0.5481) = 0.9636 \tag{4.11}$$

which results in $\kappa_m = 0.927$. The marginally-permitted agreement indicated by Cohen's kappa is the ratio of κ to κ_m, which in this case is 0.708 (95% confidence interval: 0.626, 0.789).

There is no standard for what constitutes an acceptable level of agreement, but in a recent article Lombard et al. (2002) advise

> Coefficients of .90 or greater are nearly always acceptable, .80 or greater is acceptable in most situations, and .70 may be appropriate in some exploratory studies for some indices.

> Higher criteria should be used for indices known to be liberal (i.e., percent agreement) and lower criteria can be used for indices known to be more conservative (Cohen's kappa, Scott's pi, and Krippendorff's alpha).

According to this advice, values of 0.829 for percent agreement and 0.708 for Cohen's kappa (marginally adjusted) suggest that the interrater reliability is sufficient for this exploratory portion of the analysis.

4.2.2 Explanatory Variable: Comparison Elicits Anomaly

The surprise ratings discussed above may be used to study the difference in performance between the two treatment groups. The key to this comparison is to determine whether the subjects were surprised when they should have been or whether they were surprised when they should not have been. The subjects should presumably be surprised, assuming that prediction ability is sound, when the anomaly is elicited. In most cases in the experiment, when a subject makes a prediction it is based on an anchor-and-adjust strategy (Tversky and Kahneman, 1974), where a previously revealed result is used as the anchor. The anomaly in the simulation is tied to one factor, and when the subject bases a prediction on a prior result with this factor at the other level, the anomaly is elicited. For the aOFAT subjects, the trial in which this happens is clear: the catapult arm material is only changed in one trial. For the PB-L_8 subjects, the nature of the design matrix is that there are four trials in which a comparison may be made between configurations that have opposite settings for the arm material. For this reason, the experiment's administrator asked the subject after each prediction to identify any previous trial result(s) used to make the prediction. The answers provided were unambiguous for 148 of the 162 PB-L_8 data points that were agreed upon during independent rating. Added to the 157 aOFAT data points, these 305 points form the variable *comparison elicits anomaly* indicated by $X_{CEA} = 0$ when the anomaly is not elicited or $X_{CEA} = 1$ when the anomaly is elicited.

4.2.3 Contingency Table

This section presents an analysis the contingency table with frequencies for Y_s, X_{CEA} and X_{DM}, shown as table 4.5a on the next page. In contrast to the regression for debriefing results, which included a continuous predictor variable, the data for this analysis include only dichotomous predictor variables. Although there is nothing precluding the use of such data in a loglinear or logistic regression, the risk or odds ratios of interest may be estimated directly from the contingency table, resulting in the identical values as in the corresponding regressions.

Analysis of this table allows me to address the propriety of surprise response as a function of anomaly elicitation. Specifically, I would like to know the relative risk of surprise when the anomaly is elicited, as a function of treatment group. Since surprise is a manifestation of expectation violation, participants

Table 4.5: Full and Reduced Contingency Tables for Surprise Ratings

(a) All responses

		Anomaly Elicited?	Surprised? No	Yes
Condition	aOFAT	No	79	52
		Yes	4	22
	PB-L_8	No	59	43
		Yes	24	22

(b) Anomaly elicted

Condition	Surprised? Yes	No
PB-L_8	22	24
aOFAT	22	4

(c) No anomaly elicited

Condition	Surprised? Yes	No
PB-L_8	43	59
aOFAT	52	79

showing surprise when the anomaly is present is somewhat analogous to a "hit" in a *signal detection* experiment. Table 4.5b is generated from the data in table 4.5a, to match the format shown in table 2.1 on page 32 required for the risk ratio calculation with (2.14). Parameters calculated for the risk ratio in this case are shown in the middle row of table 4.6, where the additional parameters come from (2.15) and (2.16).

Extending the signal detection analogy, the *hit rate* is only half of the picture in characterizing detection performance. The other essential bit of information is the *false alarm rate*, wherein the subject judges the signal to be present when it is not. In this study, a false alarm would roughly correspond to showing surprise when the anomaly is not elicited. One may calculate the risk ratio for this case using the reduced contingency data shown in table 4.5c. Again using (2.14)–(2.16), the parameters of interest are calculated as the values shown in the bottom row of table 4.6.

Table 4.6: Contingency Analysis of Surprise Rating Results

	RR	$\log RR$	$SE(\log RR)$	$\log RR \pm 1.96 \cdot SE(\log RR)$	$RR_{95\%CI}$
$\frac{\text{PB-L}_8\vert\text{anomaly}}{\text{aOFAT}\vert\text{anomaly}}$	0.57	−0.57	0.18	$[-0.91, -0.23]$	$[0.40, 0.80]$
$\frac{\text{PB-L}_8\vert\text{no anomaly}}{\text{aOFAT}\vert\text{no anomaly}}$	1.06	0.06	0.16	$[-0.25, \quad 0.37]$	$[0.91, 1.24]$

This page intentionally left blank.

Chapter 5

Discussion

5.1 Main Findings

There were two dependent or response variables of interest in the main experiment: the debriefing result, one per subject, recorded after the completion of the design task, and the independent judges' ratings of whether the subject seemed surprised, based on the video recorded reaction after each of the 7 trials in which simulation results were predicted by the subject before being provided by the administrator.

5.1.1 *Confirmatory Analysis of Debriefing Results*

The contingency data shown as table 4.1 on page 75 gives the unprocessed debriefing results. In the analysis of these data, only the results taken before disclosure of the simulation flaw were used, as discussed in §3.3. This represents the natural response of the subjects. Including the results obtained after disclosure would artificially inflate the estimated rate of detection for all subjects.

Before getting into the analysis, it is remarkable to consider only the simple 2×2 contingency data in table 4.1. Regardless of any analysis that is to follow, there is no more powerful support for the hypothesis than this: about half (14 of 27 ≈ 52%) of the aOFAT group recognized the simulation problem, while nearly none (1 of 27 ≈ 4%) of the PB-L_8 group did so. The data in table 4.1 are also presented as proportions with confidence intervals in table 4.2 on page 76 for the complete picture of this contingency data.

While a proportional comparison could be used for a hypothesis test, the effect of domain knowledge was included for significance testing here. In the logistic regression of the debriefing outcome on both design method X_{DM} and normalized domain knowledge score X'_{DKS}, calculation of the regression

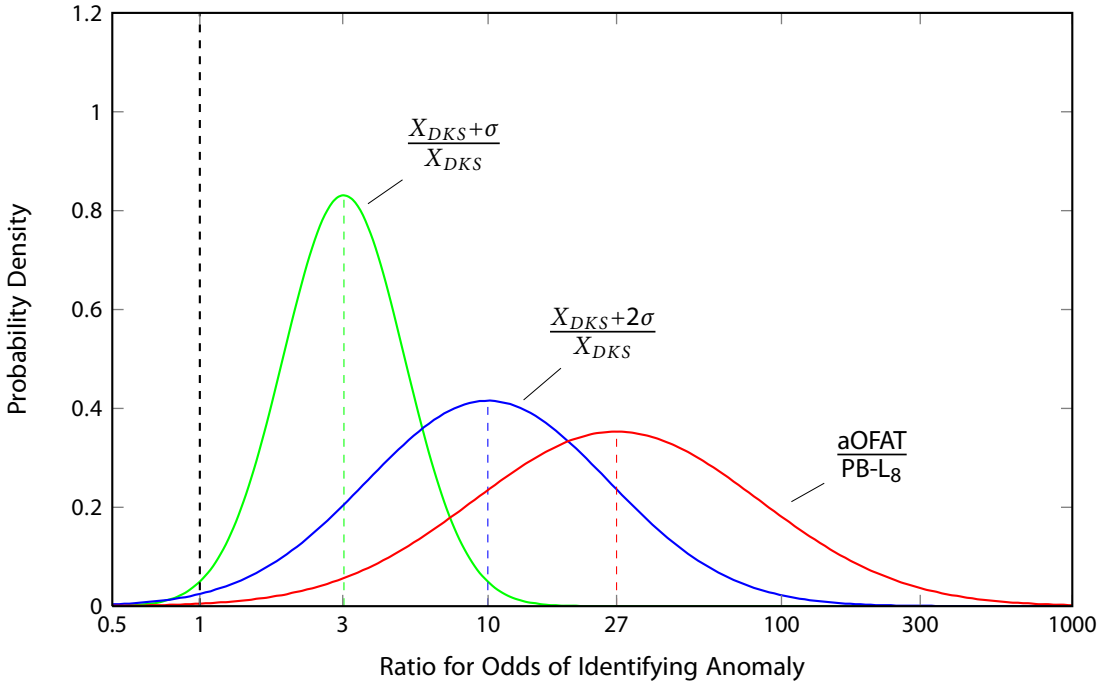


Figure 5-1: Odds ratios for participant detection of the anomaly

coefficients serves as the hypothesis test. From the results shown in table 4.3 on page 78, both design method and domain knowledge were statistically significant predictors of the debriefing outcome.

In the logistic equation, the coefficients are assumed to have a normal probability distribution. When converting from the logit to the odds ratio (by exponentiation), the probability distribution becomes asymmetric. However, if plotted on a semilog axis as shown in figure 5-1, the odds ratios appear symmetric and normally distributed. The abscissa indicates the odds ratio, and each plotted curve is labeled with a fraction that reveals the two characteristics that comprise the odds ratio for that curve: odds of a participant with the characteristic in the numerator detecting the simulation mistake, divided by the odds of a participant with the characteristic in the denominator detecting the simulation mistake. An odds ratio of 1, shown in figure 5-1 as the leftmost dashed line, indicates that there is no difference in performance with respect to the numerator and denominator characteristics – it is the null hypothesis. Odds ratios to the left of 1 indicate that participants in the denominator perform better, while odds ratios to the right of 1 indicate that participants in the numerator perform better.

Dichotomous predictor variables have only one curve when plotted in this manner; however, continuous predictor variables have infinitely many such curves. There are three probability distributions shown in this figure: one relates the aOFAT group to the PB-L_8 group, and the other two show the effect of domain knowledge. The regression coefficient for X'_{DKS} is given in units of standard deviations in table 4.3, because X'_{DKS} is the normalized version of X_{DKS}. The two curves selected to represent domain knowledge show the effect of one-standard-deviation and two-standard-deviation

increases in X_{DKS} on anomaly detection ability.

As one would expect, in this experiment a higher domain knowledge is shown to be an advantage in locating the simulation mistake. However, the largest advantage comes from choice of design method, with the aOFAT group bettering the PB-L_8 group by a factor of 27 to 1. The trend shown in the figure suggests that one would need an advantage of more than two standard deviations in domain knowledge to compensate for use of the more complex PB-L_8 design.

5.1.2 *Exploratory Analysis of Surprise Rating Results*

Interpretation of Interrater Reliability Levels Attained

One thorny issue worth further exposition before getting into this exploratory analysis is whether the surprise rating results are sufficiently reliable. Both measures of inter rater reliability – percent agreement and Cohen's kappa – suggested that these results were slightly better than the minimum threshold for reliability. Placing the precise task for the raters into context shows that the reliability is actually quite good.

For this task, the most similar work described in the literature is the research in which photographs of facial expressions are viewed by subjects who judge the emotion depicted in the image. For a group of studies using the same methodology, a review article by Russell (1994) found the median percentage agreement for *surprise* was 87.5 for Western cultures (p. 130). It appears that our independent raters performed similarly, with a simple percentage agreement level of 83. However, the following details regarding the "standard method" (p. 109) used in that group of studies suggest that our raters in fact did very well in comparison:

1. *Preselection.* Candidate photographs to be rated were selected from a large pool of photographs, and grounds for exclusion could be the investigator's own intuition, or in some cases, "if they failed to achieve 70% agreement with prediction from a panel of 25 to 30 judges." (p. 113)

2. *Posed expressions.* Most of the photographs were of actors hired to pose expressions on command. Motley and Camden (1988) performed a study to investigate the difference between posed and spontaneous expressions and found judgment accuracy for the former to be 81.4%, while the latter was 26.0%.

3. *Forced choice.* The task for raters was to view a photograph, then judge the emotion portrayed by selecting from a multiple-choice list of six different emotions possible. If there were an equal number of each emotion in the pool of photographs, this six-item forced-choice response artificially inflates the agreement proportion by about $1/6 \approx 17\%$ for choosing at random.

In contrast to those idyllic conditions, our raters were judging the natural reactions of engineers during

parts of an engineering task. It should be noted that engineers are not known for being particularly emotionally demonstrative when grinding through technical work, and these engineers were not told that any measurement of their emotions would be attempted later.

Interpretation of Risk Ratio for Expressing Surprise

The debriefing results provide strong support for the hypothesis that a more complex experimental design reduces the engineer's ability to detect a simulation mistake, and now the surprise rating results are used to investigate this phenomenon more deeply. A rigorous approach to significance testing was used for the debriefing results – one in which the odds ratio was clearly the appropriate measure. However, odds is not the same as probability, and for the exploratory component considering the surprise rating results, the more intuitive risk ratio is used. A simple thought experiment illustrates the difference between the odds ratio and the risk ratio:

> Consider two subjects performing the same task independently: A, who is successful 50% of the time, and B, who is successful 10% of the time. By frequentist logic the probability of success for A is 0.50 and for B is 0.10. If given ten opportunities to perform the task, A is expected to be successful five times and to fail five times, while B is expected to be successful once and to fail nine times. The odds of success for A is then 5:5, and for B is 1:9. The odds ratio of success for A compared to B is $\frac{5/5}{1/9} = 9$, and the risk ratio of success for A compared to B is $\frac{0.50}{0.10} = 5$. The issue to consider is this: when describing the relative performance, does one say A is nine times as capable as B or that A is five times as capable as B? For one who observed A's five successes versus B's one success in the same number of tries, it would be unnatural to consider A to be nine times as capable as B.

Using the risk ratio as an intuitive measure of relative performance, the two most important questions that can be answered with the surprise rating results are: (1) Was the participant surprised when it was appropriate? and (2) Was the participant surprised when it was not appropriate? Justification for asking these particular questions is given by noting parallels between this experiment and a signal detection experiment.

A signal detection experiment is one in which a human subject is either presented with a stimulus (visual, aural, etc.) or not (background noise only), then asked to judge whether the stimulus is present. If the stimulus is present and the participant judges this correctly, this is called a *hit*. If the stimulus is not present but the participant says it is, this is called a *false alarm*. The *hit rate* is the proportion of stimulus-present trials that result in hits, and the *false alarm rate* is the proportion of stimulus-absent trials that result in hits. There are two other possible scenarios: a *miss* is when the stimulus is present but the participant says it is not, and a *correct rejection* is when the stimulus is not present and the participant judges this correctly. The *miss rate* is one minus the hit rate, and the *correct rejection rate*

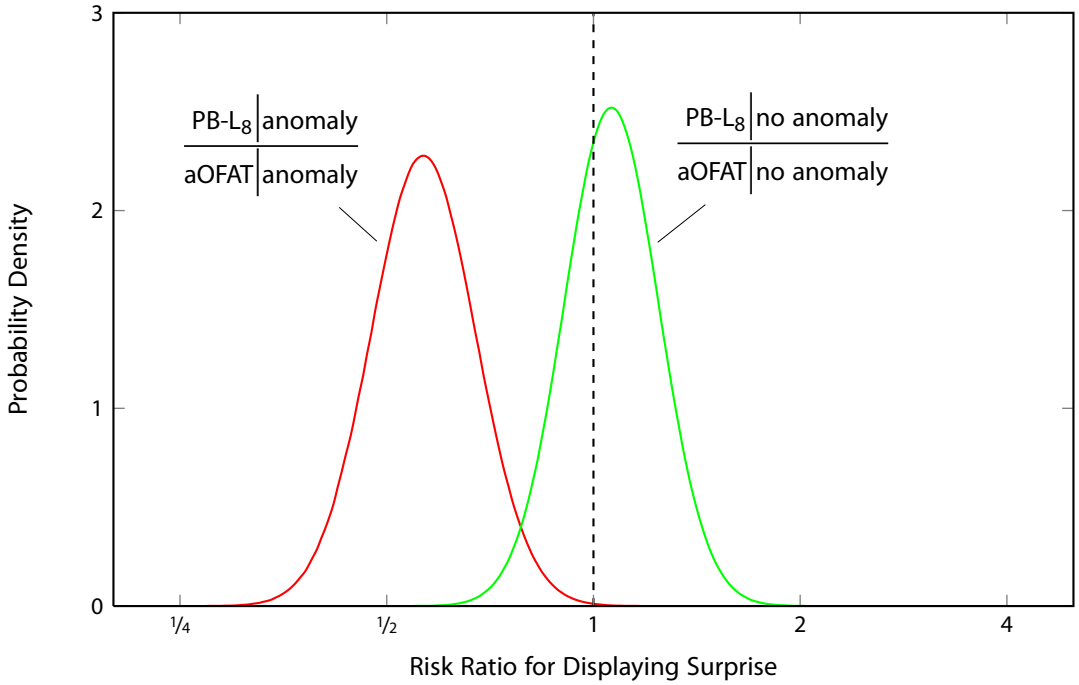


Figure 5-2: Risk ratios for participant expressing surprise at the simulation result

is one minus the false alarm rate, so only the first two rates are necessary to characterize detection performance fully. Good detection ability requires the hit rate to be higher than the false alarm rate.

In this experiment, the stimulus (the simulation mistake, if the anomaly is elicited) is presented or not, and the reaction is judged for surprise. If the participant shows surprise when the anomaly is elicited, this suggests that the participant had a correct expectation that was violated – a hit. If the participant shows surprise when the anomaly is not elicited, this suggests that the participant did not have the correct expectation – a false alarm.

Although they were calculated using the short cut equation given in (2.14), risk ratios may also be calculated by exponentiating the coefficients in a loglinear regression, the same way that odds ratios are calculated from the coefficients of the logistic regression. This helps to explain why, as with the odds ratio, the probability distribution for the risk ratio is asymmetric. However, if plotted on a semilog axis as shown in figure 5-2, the risk ratios appear symmetric and normally distributed. The abscissa indicates the risk ratio, and each plotted curve is labeled with a fraction that reveals the two characteristics that comprise the risk ratio for that curve: probability of a participant with the characteristic in the numerator detecting the simulation mistake, divided by the probability of a participant with the characteristic in the denominator detecting the simulation mistake. A risk ratio of 1, shown in the figure as the dashed line in the center, indicates that there is no difference in performance with respect to the numerator and denominator characteristics – it is the null hypothesis. Risk ratios to the left of 1 indicate that participants in the denominator perform better, while risk

ratios to the right of 1 indicate that participants in the numerator perform better.

From the curves shown in figure 5-2 on the preceding page, it appears that the aOFAT group is better at correctly reacting to anomalous behavior, since the density curve for elicited anomalies is entirely to the left of the dashed line. From the parallel with signal detection experiments, this is analogous to the aOFAT group having a higher hit rate than the PB-L_8 group. The other curve shown in this figure is the probability density of the risk ratio when no anomaly is elicited, and its position – nearly centered on the dashed line representing the null hypothesis – indicates that neither experimental group has an advantage in the analog to false-alarm performance in detection theory.

A partial explanation for this behavior was observed informally by the author over the course of the experiment: some subjects did not put much effort into generating high-quality predictions, preferring instead to guess "in the ballpark". As a result, many of those predictions had significant error. By the very definition of expectation violation, a large discrepancy between the predicted value and the reported value should result in an expression of surprise, without regard to the correctness of the prediction. This is precisely what happened in a substantial portion of the responses where the anomaly was not elicited yet the participant reacted with surprise.

5.2 Limitations

This study required a compromise to include mainly participants from the population to which the research directly applies: full-time working professionals with degrees in engineering and science. To accommodate the busy schedules of the participants, and to limit expenses incurred by reimbursement of high hourly wages, the design task was limited to the minimum required for a hypothesis test.

The motivation of subjects for the design task in this study is likely not the same as motivation in the subjects' daily work. There was nothing at stake for the subjects here except perhaps pride. If this introduced a bias, however, it would likely apply to all subjects, and since the analysis here focuses on the relative performance such an effect may not be important in this context. This is a known limitation in choosing an *in vitro* methodology over an *in vivo* approach.

In resolving independent analysts' ratings of video recorded reactions, points of disagreement were discarded, and the analysis was not changed to address the omitted data. There is no clear agreement in the statistics research community on the best course of action for handling missing data. Some claim that techniques for including this information in any analyses (e.g. multiple imputation) amount to fabricating data, while others claim that it requires a stronger assumption to discard the data than it does to account for missing values. If the recent literature on this is an accurate indicator, missing data analysis as a valid approach does seem to be gaining momentum (Graham, 2009); however, commercial software capability in this area appears to be scant.

Chapter 6

Conclusions and Suggested Future Work

At the outset of this study, a literature review of the concepts involved suggested that a properly designed human subjects experiment would produce results consistent with the hypothesis that complexity of experimental design hampers ability to recognize anomalous behavior, and the final results did not disappoint in this respect. The main uncertainty seemed to be in the exact magnitude of the difference the design method would make. The answer to this question, while in line with *a priori* estimates, was nevertheless stunning: *about half* of the one-factor-at-a-time designers successfully recognized the simulation problem, and *nearly none* of the fractional-factorial designers did so. This study was intended to expose this phenomenon, but not necessarily to investigate the specific cognitive mechanisms by which it occurs. However, I begin the discussion on those mechanisms with the first conjecture below. There were also some interesting findings on the periphery of these expected results, and much of the focus in this final chapter is on those findings and the implications of the phenomenon observed in this study.

6.1 Surprises and Conjectures

From the literature review, I deduced that an engineer using a more complex method would be less likely to discover a problem because of the increased strain on working memory. In an equivalent non-experimental task, I would expect that upon noticing anomalous behavior, the engineer would cease working toward the design goal until the problem is resolved. In this controlled experiment, subjects were not permitted to do this. Because the confirmatory dependent variable was measured during the debriefing, the limitations of working memory do not fully explain the result.

Both groups encountered the anomaly during a sequence of trials, then were questioned at the end. The *von Restorff effect* states that when considering items in a sequence, those that are distinctive are more easily remembered than those that are not (von Restorff, 1933; Hunt, 1995). This effect is

also called the *isolation effect*, as it was thought that an item being isolated from the others would produce the effect. However, Green (1956) showed that the likely mechanism for the effect is an unexpected change – a surprise – encountered in the recalled item. The first conjecture is that the one-factor-at-a-time designers were better at recalling the anomalous behavior during the debriefing because they were more likely to be surprised when the anomaly was elicited, as seen in the surprise rating results.

For the overall results, I did not expect the rates of detection to be so low. Compared to the predictions used for sample-size planning, the resulting detection rates do not appear dramatically lower. However, the predictions in those statistical power calculations were thought to be conservative estimates. It would not have been surprising for the detection rates to be higher, particularly considering the demographics of the main study, in which 44 of 54 subjects had a master's degree or higher in an engineering or science field and 21 of 54 subjects had over 20 years of full-time working experience. The conjecture here is that the low overall performance was primarily due to weakness in domain knowledge and poor effort in making predictions.

Weakness in domain knowledge appeared to affect subjects in several ways. Some subjects recalled the correct principles related to the counterintuitive behavior but missed key concepts in applying them. For example, when observing that increasing the mass of the catapult arm resulted in the ball traveling farther, several participants concluded that this was due to momentum transfer. The issue is that if momentum is conserved when the ball transitions from the catapult arm into ballistic flight, then a heavier arm implies more momentum on the catapult side of this transition. Since the mass of the ball does not change in this case, the velocity of the ball would increase to attain the higher momentum transferred by the heavier arm. The glaring weakness in this argument is that, in fact, momentum is not conserved since much of it is lost when the catapult arm crashes into the stopper to launch the ball. The ball's momentum is no doubt the same at the instant before and after launch, but all of the catapult arm's momentum is lost when it is abruptly halted.

Another surprising facet to this domain-knowledge-weakness effect is that a deficiency in theory that only mildly influenced the outcome appeared to lower confidence in areas that should have been strengths for some participants. For example, many participants admitted to a lack of confidence in aerodynamics theory, but they generally concluded early on that changes to the ambient temperature or relative humidity of the operating environment would have little impact on the outcome, in comparison to the effects of the other control factors. Some subjects using the more complex algorithm, in which environmental conditions were changing at the same time as the catapult arm, concluded that the counterintuitive behavior must be caused by something they did not understand in the aerodynamics.

Physically plausible reasoning for the counterintuitive behavior was rare. For subjects using the method with multiple factors changing at the same time, a convenient explanation was that there were either interactions or misunderstood principles for one or more other factors involved. For

subjects using the one-factor-at-a-time approach, the isolated nature of the counterintuitive behavior resulted in more explanations attempted. Of these, only one participant offered a physically plausible explanation, by reasoning through the effect of Magnus lift on the ball. The observed behavior was that a heavier arm caused the ball to travel farther, and this was explained by correctly noting that the heavier arm would cause a lower ball speed at launch, but would also cause a lower ball spin at launch, which would in turn cause a lower Magnus force. Since the direction of the Magnus force is usually[1] toward the direction of spin (thus, toward the ground for the top-spinning ball), a lower force meant the ball would not be pushed down as quickly, and the subject's reasoning was that this effect must have been enough to compensate for the lower speed of the ball at launch.

In some cases, the effort put forth in making predictions seemed to lack rigor. Evidence in support of this claim includes the observed phenomenon of subjects predicting close to the provided answer but then expressing surprise. Surprise is a manifestation of expectation violation, so a subject's display of surprise in these cases implies that an expectation of poor prediction was violated.

6.2 Implications of Observed Phenomenon

It is clear that anomaly detection ability is severely compromised in a complex experimental plan. Such approaches offer great benefits like efficiency, robustness to noise and ability to resolve more interactions than with the simpler one factor at a time method. However, if this path is taken, one should be vigilant in rooting out flaws before using the method, as these results suggest that an undiscovered simulation problem is unlikely to be detected during the experiment or its analysis.

Advice given by design methodologists usually excludes approaches specifically geared toward discovering flaws in the data. For example, in the single case study applying a fractional factorial design to a computer simulation in Box et al. (1978, pp. 429-432), the authors advised skipping a one-factor-at-a-time sensitivity analysis, reasoning that since the complex simulation included statistical sampling to model variance in the system, a fractional factorial design would be necessary to evaluate the control factor effects. Although the model had been verified using actual data, a counterintuitive main effect discovered in the post-experiment analysis caused the subject-matter experts to question the simulation. Our results, however, suggest caution in placing trust in subject-matter experts to discover simulation problems by only considering the control factor effects. In our study, 27 subject-matter experts reviewed the control factor effects and could have noticed the problem by observing the wrong sign for the flawed factor, yet only 1 of 27 was able to do so.

As is more typically the case, Johnson et al. (2008) issue the following caveat before getting into design

[1] In this case, the ball speed was low enough to encounter a bizarre effect seen only in smooth spheres in the subcritical regime: a reversal of the normal direction of the Magnus force. This negates the creative argument of the participant, but in fairness it is not a widely known phenomenon.

theory:

> This research assumes that the computer simulation has been calibrated, verified, and validated, and that it can be used to make accurate predictions about the behavior of the physical system it models. Once an adequate computer simulation model is obtained, the next major consideration is to decide how experiments will be designed and carried out.

One notable exception to counter these two examples is found in Kleijnen et al. (2005), where several techniques for integrating validation exercises into early experimentation are provided. In particular, Kleijnen et al. advocate checking the signs of main effects in any experimental design using simulation models, and they also propose using preliminary sensitivity analyses (SA) and sequential bifurcation (SB) for combination screening and validation studies. This is sound advice, in the opinion of this author. In fact, if the participants in this study were specifically instructed to check effect signs for errors, the outcome of the experiment may have been quite different.

The observed phenomenon has implications that go beyond computational mistakes, into the realm of scientific learning. It is well known in that research community that scientific insight is built through encountering anomalies and then focusing on the anomalous behavior until it is understood. In a research study with similarities to this one, Dunbar (1995) performed a series of *in vitro* experiments in which scientists were given a set of real data including an anomaly that led to an important discovery in the original work. The scientist subjects in this experiment were instructed to perform the same task as in the original work, then observed to determine whether they would pursue the anomaly or the stated goal, and how this choice affected the ability to make the discovery. Dunbar found that:

> When subjects maintained their initial goal they did not make a discovery. When subjects changed their goal to one of exploring the cause of unexpected and/or inconsistent findings, they then made the discovery.

In the context of engineering design practice, these results suggest that ability to recognize and pursue explanations for inconsistent behavior is the key to the conceptual change necessary to build expertise in the long run. Choosing a design method that hampers this ability means the engineer might miss an opportunity to learn from the experience.

6.3 Suggested Future Work

The point of this study was to reveal another property of formal engineering design methods that should be considered when choosing a method to use with computer simulation. The problem of computational mistakes in engineering design is significant and growing, and it is important for the design research community to devote resources to better understanding and ultimately managing the problem effectively.

The obvious next step is to replicate the experiment with a different population. Based on direct experience with recruiting and scheduling busy full-time engineers, this author's recommendation is to repeat the experiment using engineering students as subjects. This is contrary to the numerous articles in the experimental psychology literature lamenting the overuse of students as experimental subjects (Bodner, 2006; Wintre et al., 2001). However, if the results were found to be similar to those obtained in this study, then future studies may choose to avoid the high costs and logistical challenges involved in using experienced engineers.

The next recommendation, also aimed at improving the experimenter's productivity, would be to test the hypothesis that reaction time is a suitable surrogate variable to replace the independent analysts' ratings of subjects' surprise response. The work by Meyer et al. (1997) suggests that this may be the case.

There are many avenues for expanding this experiment to investigate different aspects of the problem. The number of factors in the design space, strength of the effects, presence of strong interactions or noise, and alternate design methods are all worthy candidates for further investigation. The flaw chosen for this experiment was an effect *sign* error, and a similar experiment using an effect *scale* error would be a valuable contribution. A simple substitution of the device under design to other popular devices used to demonstrate DOE, for instance the paper helicopter (Box, 1992) or ball and funnel (Gunter, 1993), might lead to further insight.

Another recommended study is to include in the complex approach one or more of the techniques suggested by Kleijnen et al. (2005) for countering validity problems: checking the signs of the main effects or starting with a sensitivity analysis or sequential bifurcation.

Studying the effect of pairs of engineers working on the design task in tandem could be enlightening in a similar experiment. There are competing views in the literature regarding what effect this might have. Advocates of *pair programming* claim improved productivity (Cockburn and Williams, 2000), but the theory of *social loafing* states that, at least in certain cases, the opposite effect occurs (Karau and Williams, 1993).

Finally, an intriguing aspect of this problem might be the effect of specific *personality traits* on the ability or willingness to detect anomalies in engineering design. This could be measured by administration of a survey "instrument" during the experiment. Two potential candidates for this were identified during the literature search: the Curiosity and Exploratory Inventory (CEI) (Kashdan et al., 2004) and the Hurtt skepticism scale (Hurtt, 1999).

This page intentionally left blank.

Bibliography

Agresti, A., and B. A. Coull, "Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions," *The American Statistician*, Vol. 52, No. 2, 1998, pp. 119–126.

AIAA Computational Fluid Dynamics Committee, "Guide for the Verification and Validation of Computational Fluid Dynamics Simulations," Standard AIAA-G-077-1998, American Institute of Aeronautics and Astronautics, Reston, VA, 1998.

Alberdi, E., D. H. Sleeman, and M. Korpi, "Accommodating Surprise in Taxonomic Tasks: The Role of Expertise," *Cognitive Science*, Vol. 24, No. 1, 2000, pp. 53–91.

Anderson, M. J., "Tabletop Hockey Meets Goals for Teaching Experimental Design," URL: http://www.statease.com/pubs/hockey.pdf [accessed 02 December 2009].

Antony, J., "Training for Design of Experiments Using a Catapult," *Quality and Reliability Engineering International*, Vol. 18, 2002, pp. 29–35.

Antony, J., and F. J. Antony, "Teaching the Taguchi method to industrial engineers," *Work Study*, Vol. 50, No. 4/5, 2001, pp. 141–149.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards, "Reporting Standards for Research in Psychology: Why Do We Need Them? What Might They Be?" *American Psychologist*, Vol. 63, No. 9, 2008, pp. 839–851.

Bacchetti, P., L. E. Wolf, M. R. Segal, and C. E. McCulloch, "Ethics and Sample Size," *American Journal of Epidemiology*, Vol. 161, No. 2, 2005, pp. 105–110.

Baker, L., "Differences in the Standards Used by College Students to Evaluate Their Comprehension of Expository Prose," *Reading Research Quarterly*, Vol. 20, No. 3, 1985, pp. 297–313.

Besnard, D., and L. Cacitti, "Trouble-Shooting in Mechanics: A Heuristic Matching Process," *Cognition, Technology & Work*, Vol. 3, 2001, pp. 150–160.

Bodner, T. E., "Designs, Participants, and Measurement Methods in Psychological Research," *Canadian Psychology*, Vol. 47, No. 4, 2006, pp. 263–272.

Booker, A. J., "Design and Analysis of Computer Experiments," in *7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis & Optimization*, AIAA-98-4757, St. Louis, MO, September 2-4, 1998, pp. 118–128.

Box, G. E. P., "George's Column: Teaching Engineers Experimental Design with a Paper Helicopter," *Quality Engineering*, Vol. 4, No. 3, 1992, pp. 453–459.

Box, G. E. P., W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*, John Wiley & Sons, 1978.

Box, G. E. P., and K. B. Wilson, "On the Experimental Attainment of Optimal Conditions," *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 13, No. 1, 1951, pp. 1–45.

Briggs, C., "A Process for Design Engineering," in *AIAA Space 2004 Conference and Exhibit*, San Diego, California, September 28-30, 2004, pp. 1–5.

Chinn, C. A., and W. F. Brewer, "The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction," *Review of Educational Research*, Vol. 63, No. 1, 1993, pp. 1–49.

———, "Mental Models in Data Interpretation," in *Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers*, The University of Chicago Press, Vol. 63, September 1996, pp. S211–S219.

———, "An Empirical Test of a Taxonomy of Responses to Anomalous Data in Science," *Journal of Research in Science Teaching*, Vol. 35, No. 6, 1998, pp. 623–654.

Clopper, C. J., and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, Vol. 26, No. 4, 1934, pp. 404–413.

Cockburn, A., and L. Williams, "The Costs and Benefits of Pair Programming," in *Proceedings of the First International Conference on Extreme Programming and Flexible Processes in Software Engineering (XP 2000)*, Addison-Wesley, 2000, pp. 223–247.

Cohen, J., "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, Vol. 20, No. 1, 1960, pp. 37–46.

———, *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, 2nd ed., 1988.

Cohen, J., P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, 3rd ed., 2003.

Craik, K., *The Nature of Explanation*, Cambridge University Press, 1943.

D'Agostino, R., and E. S. Pearson, "Tests for Departure from Normality. Empirical Results for the Distributions of $b2$ and $\sqrt{b1}$," *Biometrika*, Vol. 60, No. 3, 1973, pp. 613–622.

Daniel, C., "One-at-a-Time Plans," *Journal of the American Statistical Association*, Vol. 68, No. 342, 1973, pp. 353–360.

Davis, R., "Diagnostic Reasoning Based on Structure and Behavior," *Artificial Intelligence*, Vol. 24, 1984, pp. 347–410.

Dubois, P. F., "Maintaining Correctness in Scientific Programs," *IEEE Computing in Science and Engineering*, Vol. 7, No. 3, 2005, pp. 80–85.

Dunbar, K., "How scientists really reason: Scientific reasoning in real-world laboratories," in *Mechanisms of Insight*, edited by R. J. Sternberg and J. Davidson, MIT Press, 1995.

Elmer-Dewitt, P., "Ghost in the Machine," *Time*, Vol. 135, No. 5, 1990, pp. 58–59.

Ephrath, A. R., and R. E. Curry, "Detection by Pilots of System Failures During Instrument Landings," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-7, No. 12, 1977, pp. 841–848.

Ericsson, K. A., and H. A. Simon, *Protocol analysis: verbal reports as data*, MIT Press, Cambridge, MA, 1993.

Fisher, R. A., "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, Vol. 33, 1926, pp. 503–513.

———, *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935.

Frey, D. D., F. Engelhardt, and E. M. Greitzer, "A role for "one-factor-at-a-time" experimentation in parameter design," *Research in Engineering Design*, Vol. 14, 2003, pp. 65–74.

Frey, D. D., and R. Jugulum, "The Mechanisms by Which Adaptive One-factor-at-a-time Experimentation Leads to Improvement," *Journal of Mechanical Design*, Vol. 128, 2006, pp. 1050–1060.

Frey, D. D., and H. Wang, "Adaptive One-Factor-at-a-Time Experimentation and Expected Value of Improvement," *Technometrics*, Vol. 48, No. 3, 2006, pp. 418–431.

Friedman, M., and L. J. Savage, "Planning Experiments Seeking Maxima," in *Techniques of Statistical Analysis*, edited by C. Eisenhart and M. W. Hastay, McGraw-Hill, New York, pp. 365–372, 1947.

Gai, E. G., and R. E. Curry, "Failure Detection by Pilots during Automatic Landing: Models and Experiments," *Journal of Aircraft*, Vol. 14, No. 2, 1977, pp. 135–141.

Gentner, D., *Mental Models*, Lawrence Erlbaum Associates, 1983.

Giunta, A. A., S. F. W. Jr., and M. S. Eldred, "Overview of Modern Design of Experiments Methods for Computational Simulations," in *41st AIAA Aerospace Sciences Meeting and Exhibit*, American Institute of Aeronautics and Astronautics, Reno, Nevada, January 6-9, 2003, pp. 1–17.

Gorsky, P., and M. Finegold, "The role of anomaly and of cognitive dissonance in restructuring students' concepts of force," *Instructional Science*, Vol. 22, 1994, pp. 75–90.

Graham, J. W., "Missing Data Analysis: Making It Work in the Real World," *Annual Review of Psychology*, Vol. 60, No. 1, 2009, pp. 549–576.

Green, R. T., "Surprise as a Factor in the von Restorff Effect," *Journal of Experimental Psychology*, Vol. 52, No. 5, 1956, pp. 340–344.

Gunter, B., "Through a Funnel Slowly With Ball Bearing and Insight to Teach Experimental Design," *The American Statistician*, Vol. 47, No. 4, 1993, pp. 265–269.

Halpern, S. D., J. H. T. Karlawish, and J. A. Berlin, "The Continuing Unethical Conduct of Underpowered Clinical Trials," *Journal of the American Medical Association*, Vol. 288, No. 3, 2002, pp. 358–362.

Hatton, L., "The T Experiments: Errors in Scientific Software," *IEEE Computational Science and Engineering*, Vol. 4, No. 2, 1996, pp. 27–38.

Hazelrigg, G. A., "On the Role and Use of Mathematical Models in Engineering Design," *Transactions of the ASME*, Vol. 121, 1999, pp. 336–341.

———, "Thoughts on Model Validation for Engineering Design," in *Proceedings of the ASME Design Engineering Technical Conference*, Chicago, Illinois, Vol. 3, September 2-6, 2003, pp. 373–380.

Hirschi, N. W., and D. D. Frey, "Cognition and complexity: An experiment on the effect of coupling in parameter design," *Research in Engineering Design*, Vol. 13, No. 3, 2002, pp. 123–131.

Hosmer, Jr., D. W., and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, 2nd ed., 2000.

Hsieh, F. Y., D. A. Bloch, and M. D. Larsen, "A Simple Method of Sample Size Calculation for Linear and Logistic Regression," *Statistics in Medicine*, Vol. 17, 1998, pp. 1623–1634.

Hunt, R. R., "The subtlety of distinctiveness: What von Restorff really did," *Psychonomic Bulletin & Review*, Vol. 2, No. 1, 1995, pp. 105–112.

Hurtt, R. K., *Skeptical About Skepticism: Instrument Development and Experimental Validation*, Ph.D. thesis, The University of Utah, August 1999.

Johnson, R. T., D. C. Montgomery, B. Jones, and J. W. Fowler, "Comparing Designs for Computer Simulation Experiments," in *Proceedings of the 2008 Winter Simulation Conference*, 2008, pp. 463–470.

Johnson-Laird, P. N., *Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness*, Harvard University Press, 1983.

JPL Special Review Board, "Report on the Loss of the Mars Climate Orbiter Mission," Tech. Report JPL D-18441, NASA Jet Propulsion Laboratory, Pasadena, CA, November 1999.

———, "Report on the Loss of the Mars Polar Lander and Deep Space 2 Missions," Tech. Report JPL D-18709, NASA Jet Propulsion Laboratory, Pasadena, CA, March 2000.

Kalagnanam, J. R., and U. M. Diwekar, "An efficient sampling technique for off-line quality control," *Technometrics*, Vol. 39, No. 3, 1997, pp. 308–319.

Karau, S. J., and K. D. Williams, "Social Loafing: A meta-analytic review and theoretical integration," *Journal of Personality and Social Psychology*, Vol. 65, 1993, pp. 681–706.

Kashdan, T. B., P. Rose, and F. D. Fincham, "Curiosity and Exploration: Facilitating Positive Subjective Experiences and Personal Growth Opportunities," *Journal of Personality Assessment*, Vol. 82, No. 3, 2004, pp. 291–305.

Klahr, D., and K. Dunbar, "Dual Space Search During Scientific Reasoning," *Cognitive Science*, Vol. 12, 1988, pp. 1–48.

Klahr, D., L. M. Triona, and C. Williams, "Hands on What? The Relative Effectiveness of Physical Versus Virtual Materials in an Engineering Design Project by Middle School Children," *Research in Science Teaching*, Vol. 44, No. 1, 2007, pp. 183–203.

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa, "State-of-the-Art Review: A User's Guide to the Brave New World of Designing Simulation Experiments," *INFORMS Journal on Computing*, Vol. 17, No. 3, 2005, pp. 263–289.

Koehler, J. R., and A. B. Owen, "Computer Experiments," in *Handbook of Statistics 13: Design and Analysis of Experiments*, edited by S. Ghosh and C. R. Rao, Elsevier Science, pp. 261–308, 1996.

Ligetti, C. B., and T. W. Simpson, "Metamodel-Driven Design Optimization Using Integrative Graphical Design Interfaces: Results From a Job-Shop Manufacturing Simulation Experiment," *Journal of Computing and Information Science in Engineering*, Vol. 5, No. 1, 2005, pp. 8–17.

Lin, H., "The Development of Software for Ballistic-Missile Defense," *Scientific American*, Vol. 253, No. 6, 1985, pp. 46–53.

Lloyd, K. E., L. S. Reid, and J. B. Feallock, "Short-Term Retention as a Function of the Average Number of Items Presented," *Journal of Experimental Psychology*, Vol. 60, No. 4, 1960, pp. 201–207.

Lombard, M., J. Snyder-Duch, and C. C. Bracken, "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability," *Human Communication Research*, Vol. 28, No. 4, 2002, pp. 587–604.

McConnell, S., "Gauging Software Readiness With Defect Tracking," *IEEE Software*, Vol. 14, No. 3, 1997, pp. 135–136.

Mckay, M. D., R. J. Beckman, and W. J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, Vol. 21, No. 2, 1979, pp. 239–245.

Mellor, P., "A Model of the Problem or a Problem with the Model?" *Computing and Control Engineering Journal*, Vol. 9, No. 1, 1998, pp. 8–18.

Menard, S., "Coefficients of Determination for Multiple Logistic Regression Analysis," *The American Statistician*, Vol. 54, No. 1, 2000, pp. 17–24.

Metropolis, N., and S. Ulam, "The Monte Carlo Method," *Journal of the American Statistical Association*, Vol. 44, No. 247, 1949, pp. 335–341.

Meyer, W.-U., R. Reisenzein, and A. Schützwohl, "Toward a Process Analysis of Emotions: The Case of Surprise," *Motivation and Emotion*, Vol. 21, No. 3, 1997, pp. 251–274.

Miller, G. A., "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *The Psychological Review*, Vol. 63, No. 2, 1956, pp. 81–97.

Moher, D., K. F. Schulz, and D. Altman, "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials," *JAMA*, Vol. 285, No. 15, 2001, pp. 1987–1991.

Morris, J. A., and M. J. Gardner, "Calculating Confidence Intervals For Relative Risks (Odds Ratios) And Standardised Ratios And Rates," *British Medical Journal (Clinical Research Edition)*, Vol. 296, No. 6632, 1988, pp. 1313–1316.

Motley, M. T., and C. T. Camden, "Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting." *Western Journal of Speech Communication*, Vol. 52, 1988, pp. 1–22.

Muller, K. E., and V. A. Benignus, "Increasing Scientific Power With Statistical Power," *Neurotoxicology and Teratology*, Vol. 14, No. 3, 1992, pp. 211–219.

Neumann, P. G., *Computer Related Risks*, Addison-Wesley, January 1995.

Oberkampf, W. L., T. G. Trucano, and C. Hirsch, "Verification, validation, and predictive capability in computational engineering and physics," *Applied Mechanics Reviews*, Vol. 57, No. 5, 2004, pp. 345–384.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, Vol. 49, No. 12, 1996, pp. 1373–1379.

Peloton Systems LLC., "Xpult Experimental Catapult for design of experiments and Taguchi methods." URL: http://www.xpult.com [accessed September 22, 2009].

Perreault, Jr, W. D., and L. E. Leigh, "Reliability Of Nominal Data Based On Qualitative Judgements," *Journal of Marketing Research*, Vol. 26, No. 2, 1989, pp. 135–148.

Phadke, M. S., *Quality Engineering Using Robust Design*, Prentice Hall, Englewood Cliffs, New Jersey, 1989.

Plackett, R. L., and J. P. Burman, "The Design of Optimal Multifactorial Experiments," *Biometrika*, Vol. 33, No. 4, 1946, pp. 305–325.

Pollack, I., L. B. Johnson, and P. R. Knaff, "Running Memory Span," *Journal of Experimental Psychology*, Vol. 57, No. 3, 1959, pp. 137–146.

Rapp, D. N., "How do readers handle incorrect information during reading?" *Memory & Cognition*, Vol. 36, No. 3, 2008, pp. 688–701.

Rasmussen, J., and A. Jensen, "Mental Procedures in Real-Life Tasks: A Case Study of Electronic Trouble Shooting," *Ergonomics*, Vol. 17, No. 3, 1974, pp. 293–307.

Robertson, D., K. Ulrich, and M. Filerman, "CAD and Cognitive Complexity: Beyond the Drafting Board Metaphor," *Manufacturing Review*, Vol. 4, No. 3, 1991, pp. 194–204.

Robinson, T. J., C. M. Borror, and R. H. Myers, "Robust Parameter Design: A Review," *Quality and Reliability Engineering International*, Vol. 20, No. 1, 2004, pp. 81–101.

Ruffle-Smith, H. P., "A Simulator Study of the Interaction of Pilot Workload With Errors, Vigilance, and Decisions," Tech. Report TM-78482, National Aeronautics and Space Administration, January 1979.

Russell, J. A., "Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies." *Psychological Bulletin*, Vol. 115, No. 1, 1994, pp. 102–141.

Sadoff, M., "A Study of a Pilot's Ability to Control During Simulated Stability Augmentation System Failures," Tech. Report TN D-1552, National Aeronatics and Space Administration, November 1962.

Sarin, S., "Teaching Taguchi's Approach to Parameter Designs," *Quality Progress*, Vol. 30, No. 5, 1997, pp. 102–106.

Sim, J., and C. Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy*, Vol. 85, No. 3, 2005, pp. 257–268.

Simon, H. A., "The Information Processing Explanation of Gestalt Phenomena," *Computers in Human Behavior*, Vol. 2, 1986, pp. 241–255.

Simpson, T. W., D. K. J. Lin, and W. Chen, "Sampling Strategies for Computer Experiments: Design and Analysis," *International Journal of Reliability and Applications*, Vol. 2, No. 3, 2001, pp. 209–240.

Spiegel, M. R., and L. J. Stephens, *Schaum's Outline of Theory and Problems of Statistics*, McGraw-Hill, 3rd ed., 1998.

Stevenson, D. E., "A critical look at quality in large-scale simulations," *Computing in Science and Engineering*, Vol. 1, No. 3, 1999, pp. 53–63.

Stiensmeier-Pelster, J., A. Martini, and R. Reisenzein, "The Role of Surprise in the Attribution Process," *Cognition and Emotion*, Vol. 9, No. 1, 1995, pp. 5–31.

Taguchi, G., *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs*, Quality Resources: A division of the Kraus Organization Limited, White Plains, NY; and American Supplier Institute, Inc., Dearborn, MI, 1987.

Trickett, S. B., J. G. Trafton, C. D. Schunn, and A. Harrison, "That's Odd! How Scientists Respond to Anomalous Data," in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, edited by J. D. Moore and K. Stenning, Edinburgh, Scotland, August 1-4, 2001.

Turkle, S., "How Computers Change the Way We Think," *The Chronicle of Higher Education*, Vol. 50, No. 21, 2004, pp. B26–B28.

Tversky, A., and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science*, Vol. 185, No. 4157, 1974, pp. 1124–1131.

van Eekhout, J. M., and W. B. Rouse, "Human Errors in Detection, Diagnosis, and Compensation for Failures in the Engine Control Room of a Supertanker," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-11, No. 12, 1981, pp. 813–816.

Venturino, M., "Interference and Information Organization in Keeping Track of Continually Changing Information," *Human Factors*, Vol. 39, No. 4, 1997, pp. 532–539.

Vittinghoff, E., and C. E. McCulloch, "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression," *American Journal of Epidemiology*, Vol. 165, No. 6, 2007, pp. 710–718.

von Restorff, H., "Analyse von Vorgängen im Spurenfeld: I. Über die Wirkung von Bereichsbildungen im Spurenfeld (Analysis of Processes in the Memory Trace: I. On the Effect of Region-Formation on the Memory Trace)," *Psychologie Forschung*, Vol. 18, 1933, pp. 299–342.

White, R., and R. Gunstone, *Probing Understanding*, Falmer Press, London, 1992.

Whittemore, A. S., "Sample Size for Logistic Regression with Small Response Probability," *Journal of the American Statistical Association*, Vol. 76, No. 373, 1981, pp. 27–32.

Wilson, G. V., "What Should Computer Scientists Teach to Physical Scientists and Engineers?" *IEEE Computational Science and Engineering*, Vol. 3, No. 2, 1996, pp. 46–55.

Wintre, M. G., C. North, and L. A. Sugar, "Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective." *Canadian Psychology*, Vol. 42, No. 3, 2001, pp. 216–225.

Wu, C. F. J., and M. Hamada, *Experiments: Planning, Analysis, and Parameter Design Optimization*, John Wiley & Sons, Inc., 2nd ed., 2009.

Yntema, D. B., "Keeping Track of Several Things at Once," *Human Factors*, Vol. 5, 1963, pp. 7–17.

Appendix A

Mathematical Model of the Catapult Device

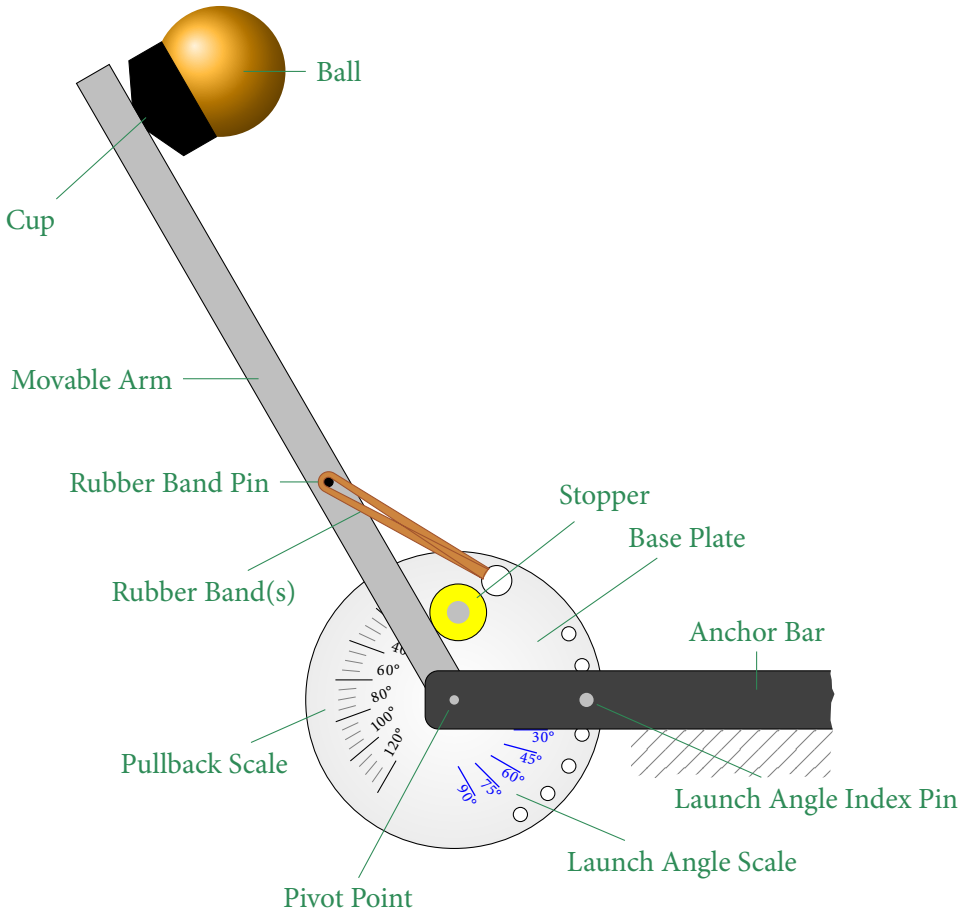
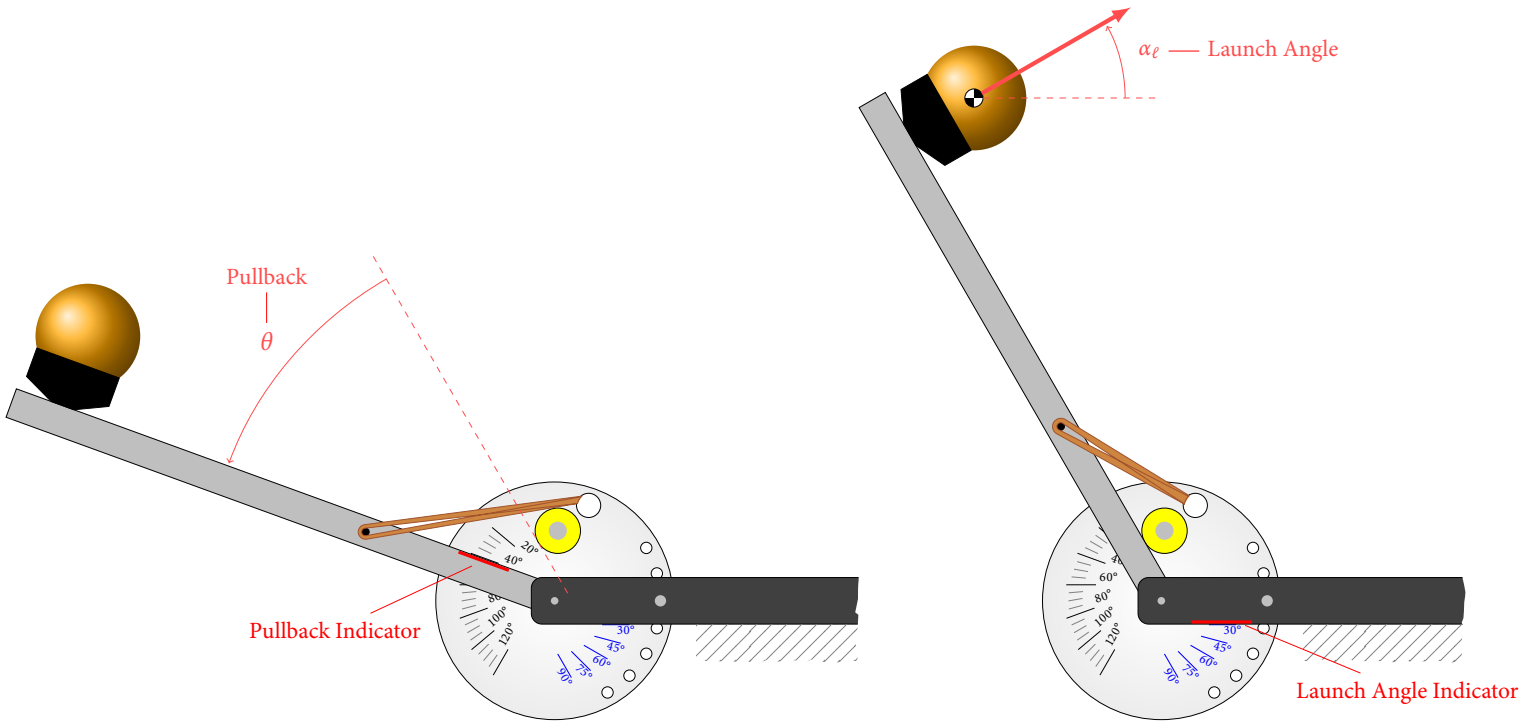


Figure A-1: The Xpult catapult, showing component nomenclature



(a) Catapult device in the 30°-launch, 40°-pullback position

(b) Catapult device in the 30°-launch, 0°-pullback position

Figure A-2: Catapult angle definitions. Note that pullback is relative to the stopper and launch angle is relative to the horizon.

A.1 Overview

In the operation of the catapult, the dynamics of interest are different before and after launch. To launch, the catapult arm is pulled back to the desired position θ_o and then released. Energy stored in the rubber band(s) accelerates the catapult arm with the ball held in place by the cup. When the arm hits the stop, at the selected launch angle α_ℓ, the ball begins a ballistic trajectory with an initial direction, speed and rotation.

Table A.1: Salient Features of the Simulation Model

	from release to launch	from launch to (first) landing
Initial State	released from rest at initial pullback	launch speed, position & direction
Kinetic Energy	arm (rigid body); ball (point mass)	ball (point mass)
Potential Energy	rubber band(s) (ideal spring); gravity	gravity
Losses	none	aerodynamic drag; Magnus lift due to ball spin
Final State	launch speed, position & direction	landing position

A.2 Prelaunch Dynamics

For the prelaunch dynamics, there are several simplifying assumptions that we can make. First, since the ball and cup are very close together, we may lump them together into a single point mass at the location of their shared center of gravity. Second, the pivot point may be assumed to be frictionless. In reality, there is a small amount of friction present but it is much less than the other modeled effects. Third, we are ignoring any damping due to aerodynamic drag in prelaunch. Finally, we may assume that the cup and the arm are both perfectly rigid (i.e., their elasticity is not modeled).

A free-body diagram of the arm with point mass is shown in figure A-3. Indicated here are the reaction forces at the pivot ($\vec{F}_{px}, \vec{F}_{py}$), the reaction torque at the pivot ($\vec{\tau}_p$), the gravitational forces on the ball/cup point mass ($\vec{F}_{gbc}$) and the arm ($\vec{F}_{ga}$), and the input force from the rubber band ($\vec{F}_{rb}$). If we were concerned with including the effects of pivot friction, we would need to know the reaction forces at the pivot point to calculate the reaction torque ($\vec{\tau}_p$). Since we are neglecting friction, $\vec{\tau}_p \equiv 0$ and these forces do not matter to the dynamics calculations at all.

We may calculate the rotational acceleration of the arm by summing the moments about the pivot:

$$\sum \vec{M}_\theta = \left(\vec{R}_a \times \vec{F}_{ga}\right) + \left(\vec{R}_{bc} \times \vec{F}_{gbc}\right) + \left(\vec{R}_{rbp} \times \vec{F}_{rb}\right) \tag{A.1}$$

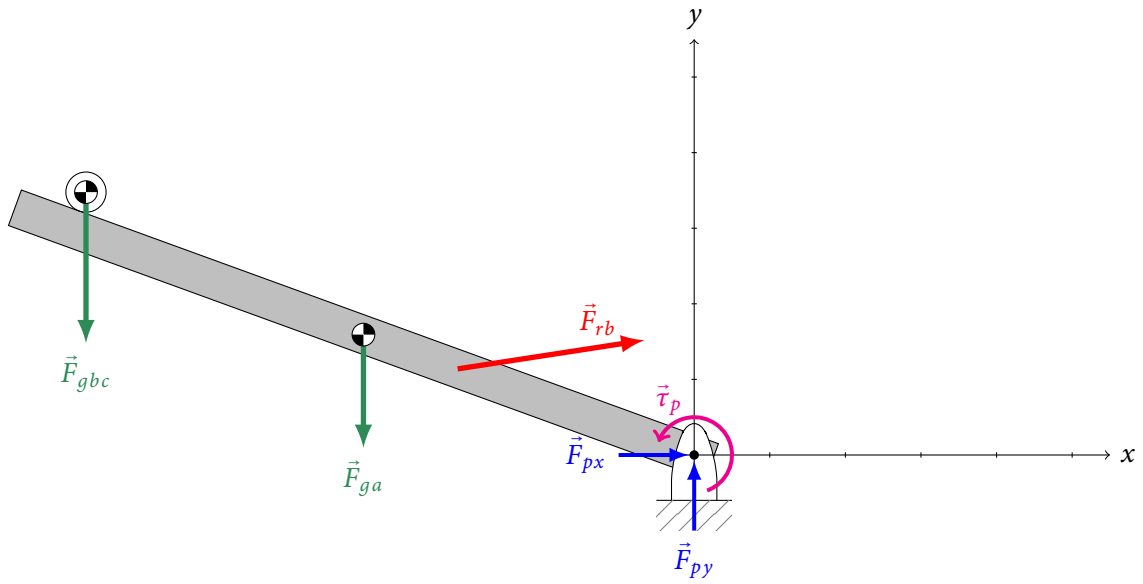


Figure A-3: Free body diagram of the catapult arm before the ball is launched

Using lumped parameter modeling, we know that this moment sum is equal to

$$\sum \vec{M}_\theta = I_{abc} \frac{d^2\theta}{dt^2} \tag{A.2}$$

where I_{abc} is the second moment of inertia of the arm and point mass about the pivot point, and $\frac{d^2\theta}{dt^2}$ is the second derivative of the pullback θ with respect to time t. We have formulated these equations in terms of the pullback, so here counterclockwise motion of the arm is considered to be in the positive direction. We can combine the above equations into the state-space formulation

$$\begin{bmatrix} \frac{d\theta}{dt} \\ \frac{d\omega}{dt} \end{bmatrix} = \begin{bmatrix} \omega \\ \frac{1}{I_{abc}} \left[\left(\vec{R}_a \times \vec{F}_{ga} \right) + \left(\vec{R}_{bc} \times \vec{F}_{gbc} \right) + \left(\vec{R}_{rbp} \times \vec{F}_{rb} \right) \right] \end{bmatrix} \tag{A.3}$$

where ω is the rotational speed of the arm. Note that with the chosen sign convention, ω will always be less than or equal to zero after release.

Next, we express all values in (A.3) in terms of the engineering parameters given in table A.3 on page 119. Starting with the moment of inertia about the pivot,

$$I_{abc} = I_a + I_{bc} = \frac{m_a L_a^2}{3} + \left(m_b + m_c \right) \left(R_{bc}^2 + L_{bc}^2 \right) \tag{A.4}$$

where the offset for the equivalent point mass center of the ball plus cup may be calculated as

$$L_{bc} = \frac{L_b m_b + L_c m_c}{m_b + m_c} \tag{A.5}$$

For the gravitational forces, we have

$$\vec{R}_a = -R_a \sin(\theta + \alpha_\ell)\,\hat{\imath} + R_a \cos(\theta + \alpha_\ell)\,\hat{\jmath} \tag{A.6}$$

$$\vec{F}_{ga} = -m_a g\,\hat{\jmath} \tag{A.7}$$

$$\vec{R}_{bc} = \left(-R_{bc} \sin(\theta + \alpha_\ell) + L_{bc} \cos(\theta - \alpha_\ell)\right)\,\hat{\imath} + \left(R_{bc} \cos(\theta + \alpha_\ell) + L_{bc} \sin(\theta - \alpha_\ell)\right)\,\hat{\jmath} \tag{A.8}$$

$$\vec{F}_{gbc} = -m_{bc} g\,\hat{\jmath} \tag{A.9}$$

where $\hat{\imath}$ and $\hat{\jmath}$ are the unit vectors parallel to the positive x and positive y axes, respectively, and $g = 32.174\ \ ft/s^2$ is the specific force (force per unit mass) due to gravity.

For the rubber band force,

$$\vec{R}_{rbp} = -R_{rbp} \sin(\theta + \alpha_\ell)\,\hat{\imath} + R_{rbp} \cos(\theta + \alpha_\ell)\,\hat{\jmath} \tag{A.10}$$

$$\left\|\vec{F}_{rb}\right\| = n_{rb} k_{rb} \left(\left\|\vec{L}_{rb}(\theta)\right\| - L_{rbf}\right) \tag{A.11}$$

We can get the vector form of the latter equation by using the vector form of the rubber band length

$$\begin{aligned} \vec{L}_{rb}(\theta) &= L_{rbx}\,\hat{\imath} + L_{rby}\,\hat{\jmath} \\ &= \underbrace{\sqrt{L_{rbx}^2 + L_{rby}^2}}_{\text{magnitude}} \cdot \underbrace{\left[\frac{1}{\sqrt{L_{rbx}^2 + L_{rby}^2}} \left(L_{rbx}\,\hat{\imath} + L_{rby}\,\hat{\jmath}\right)\right]}_{\text{unit vector}} \end{aligned} \tag{A.12}$$

where

$$L_{rbx} = R_{rbh} \sin(\phi - \alpha_\ell) + R_{rbp} \sin(\theta + \alpha_\ell) \tag{A.13a}$$

$$L_{rby} = R_{rbh} \cos(\phi - \alpha_\ell) - R_{rbp} \cos(\theta + \alpha_\ell) \tag{A.13b}$$

Substituting,

$$\vec{F}_{rb} = n_{rb} k_{rb} \left[1 - \left(\frac{L_{rbf}}{\sqrt{L_{rbx}^2 + L_{rby}^2}}\right)\right] \left(L_{rbx}\,\hat{\imath} + L_{rby}\,\hat{\jmath}\right) \tag{A.14}$$

Evaluating the three cross products in Equation A.3 results in

$$\vec{R}_a \times \vec{F}_{ga} = m_a g R_a \sin(\theta + \alpha_\ell)\,\hat{k} \tag{A.15}$$

$$\vec{R}_{bc} \times \vec{F}_{gbc} = m_{bc} g \left(R_{bc} \sin(\theta + \alpha_\ell) - L_{bc} \cos(\theta - \alpha_\ell)\right)\,\hat{k} \tag{A.16}$$

$$\vec{R}_{rbp} \times \vec{F}_{rb} = -n_{rb} k_{rb} R_{rbp} \left[1 - \left(\frac{L_{rbf}}{\sqrt{L_{rbx}^2 + L_{rby}^2}}\right)\right] \left(L_{rbx} \cos(\theta + \alpha_\ell) + L_{rby} \sin(\theta + \alpha_\ell)\right)\,\hat{k} \tag{A.17}$$

where $\hat{k}$ is the unit vector parallel to the positive θ axis (i.e., out of the page for this right hand system). Substituting everything back into Equation A.3 and collecting terms, we get

$$\frac{d\omega}{dt} = \frac{1}{\frac{m_a L_a^2}{3} + (m_b + m_c)\left(R_{bc}^2 + L_{bc}^2\right)} \Bigg\{ m_a g R_a \sin(\theta + \alpha_\ell) + m_{bc} g \left(R_{bc} \sin(\theta + \alpha_\ell) - L_{bc} \cos(\theta - \alpha_\ell)\right) - n_{rb} k_{rb} R_{rbp} \left[1 - \left(\frac{L_{rbf}}{\sqrt{L_{rbx}^2 + L_{rby}^2}}\right)\right] \left(L_{rbx} \cos(\theta + \alpha_\ell) + L_{rby} \sin(\theta + \alpha_\ell)\right) \Bigg\} \quad \text{(A.18)}$$

A.3 Ballistics Dynamics

At the point of launching, the dynamics of interest switches to the ballistics of the projectile (ball). The initial conditions for this simulation come from the last state of the catapult dynamics simulation: linear and rotational speeds at launch, direction of travel (launch angle) and position of the ball at launch.

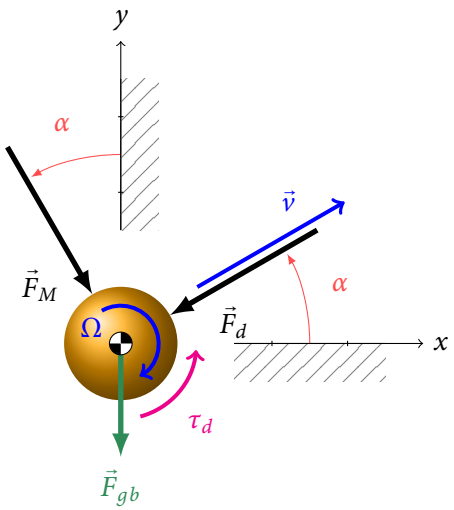


Figure A-4: Free body diagram of the ball in ballistic flight

A free-body diagram of the ball in free flight after launch is shown in figure A-4. Indicated here are the velocity vector ($\vec{v}$) and rotational speed (Ω), gravitational force ($\vec{F}_{gb}$), aerodynamic drag force ($\vec{F}_d$), aerodynamic drag torque ($\vec{\tau}_d$) and the Magnus force ($\vec{F}_M$) generated by the topspin. From the

free-body diagram, we see that simple Newtonian physics gives us

$$m_b \frac{d\vec{v}}{dt} = \left\|\vec{F}_M\right\|\left\{\sin\alpha\,\hat{\imath} - \cos\alpha\,\hat{\jmath}\right\} - \left\|\vec{F}_d\right\|\left\{\cos\alpha\,\hat{\imath} + \sin\alpha\,\hat{\jmath}\right\} - \left\|\vec{F}_{gb}\right\|\hat{\jmath} \tag{A.19}$$

This is largely an aerodynamics problem, and as is usually the case with such problems, the predicted behavior is based on empirically generated formulas that apply under specific conditions. Of particular importance here are the shape and surface roughness of the projectile, and the rotational and translational speed ranges of interest. The high-quality table tennis balls used here may be assumed to be smooth spheres, and the operational regime of interest is about 175 to 300 RPM for rotation and 10 to 25 $^{\text{ft}}/_{\text{sec}}$ for translation.

The magnitude of the aerodynamic drag force is calculated as

$$\left\|\vec{F}_d\right\| = \frac{1}{2}\rho\left\|\vec{v}\right\|^2 A \cdot C_d \tag{A.20}$$

where $\rho = \rho(RH, T)$ is the air density and $A = \frac{\pi}{4}d_b^2$ is the projected area of the ball. The aerodynamic drag coefficient C_d is very much dependent upon the surface roughness of the ball and the Reynolds number. The Reynolds number is the ratio of inertial forces to viscous forces and is given by

$$Re = \frac{\rho\left\|\vec{v}\right\| d_b}{\mu} \tag{A.21}$$

where $\mu = \mu(RH, T)$ is the dynamic viscosity of the air. To determine the working range of Reynolds numbers in the full factorial, we compute it for the ball at both launch and landing, finding that $(1.0 \times 10^4 < Re < 2.2 \times 10^4)$. Knowledge of this working range is important, since the aerodynamic drag coefficient C_d can vary dramatically above the transition number of about 3×10^5. Since we are operating well below this value, we can use a formula for drag coefficient proposed by Cheng (2008) for smooth spheres in the subcritical region.

$$C_d = \frac{24}{Re}\left(1 + 0.27Re\right)^{0.43} + 0.47\left(1 - e^{-0.04Re^{0.38}}\right) \tag{A.22}$$

The Magnus force is orthogonal to both the translational velocity vector and the rotational velocity vector, and for a smooth rotating sphere is modeled as a typical lift force:

$$\left\|\vec{F}_M\right\| = \frac{1}{2}\rho\left\|\vec{v}\right\|^2 A \cdot C_M \tag{A.23}$$

where C_M is the Magnus lift coefficient. Regrettably, there is not universal agreement on the correct expression for C_M for a smooth sphere. Many researchers have studied the Magnus effect for golf balls (Davies, 1949; Bearman and Harvey, 1976), baseballs (Briggs, 1959; Watts and Ferrer, 1987; Nathan, 2008) and other sports balls (Mehta, 1985; Mehta and Pallis, 2001). However, there is a curious

phenomenon unique to smooth spheres rotating at relatively low speeds: the direction of the Magnus force is reversed. This was first mentioned in the pioneering work by Maccoll (1928) and independently confirmed by both Davies (1949) and Briggs (1959). Loth (2008) reviewed the results of these and other authors, then proposed the approximation

$$C_M \approx \min\left[0.25\tanh\left(\Omega^\star - 0.5\right) - 0.1, \frac{0.5}{\Omega^\star}\right] \tag{A.24}$$

where

$$\Omega^\star = \frac{\Omega d_b}{\left\|\vec{v}\right\|} \tag{A.25}$$

is called the *spin number* and is the ratio of the velocity gradient across the sphere to the translational velocity of its center. Equation A.24 is intended for smooth spheres at *subcritical* Reynolds numbers greater than about 4×10^4, since the data upon which it is based was collected at Reynolds numbers in this range. Our case is slightly outside this region $\left(1 \times 10^4 < Re < 2.2 \times 10^4\right)$, but in the absence of an alternative we will use the above expression for the Magnus lift coefficient. However, we can make a simplification to Equation A.24 by noting that there are two regions of behavior divided by the point where $0.25\tanh\left(\Omega^\star - 0.5\right) - 0.1 = {}^{0.5}/_{\Omega^\star}$. This occurs at $\Omega^\star \approx 3.37$, so that

$$C_M \approx \begin{cases} 0.25\tanh\left(\Omega^\star - 0.5\right) - 0.1 & : \quad \Omega^\star < 3.37 \\ \frac{0.5}{\Omega^\star} & : \quad \Omega^\star \geq 3.37 \end{cases} \tag{A.26}$$

If we again look at the results over the full factorial, we see that the spin number $\Omega^\star$ ranges from about 0.18 to 0.30. Since we only operate in the "lower" region, we can simply use the expression

$$C_M \approx 0.25\tanh\left(\Omega^\star - 0.5\right) - 0.1 \tag{A.27}$$

Finally, the magnitude of the gravitational force is

$$\left\|\vec{F}_{gb}\right\| = m_b g \tag{A.28}$$

Substituting everything back into Equation A.19 and collecting terms, we get

$$\begin{aligned} \frac{dv_x}{dt} = \frac{\pi}{8}\frac{\rho\left\|\vec{v}\right\|^2 d_b^2}{m_b}\Bigg\{&\left(0.25\tanh\left(\Omega^\star - 0.5\right) - 0.1\right)\sin\alpha \\ &- \left[\frac{24}{Re}\left(1 + 0.27Re\right)^{0.43} + 0.47\left(1 - e^{-0.04Re^{0.38}}\right)\right]\cos\alpha\Bigg\} \end{aligned} \tag{A.29a}$$

$$\frac{dv_y}{dt} = -\frac{\pi}{8}\frac{\rho\left\|\vec{v}\right\|^2 d_b^2}{m_b}\left\{\left(0.25\tanh\left(\Omega^\star - 0.5\right) - 0.1\right)\cos\alpha + \left[\frac{24}{Re}\left(1 + 0.27Re\right)^{0.43} + 0.47\left(1 - e^{-0.04Re^{0.38}}\right)\right]\sin\alpha\right\} - m_b g \quad \text{(A.29b)}$$

where we have used the orthogonal decomposition of velocity $\vec{v} = v_x\,\hat{\imath} + v_y\,\hat{\jmath}$.

Two key properties in the above aerodynamics of the ball in flight are the air density and dynamic viscosity. Since we have elected to use ambient temperature and relative humidity as control factors, we must consider that the density and viscosity are dependent upon the ambient conditions, and we do so in the analysis that follows.

The equation for the density of humid air is

$$\rho = \frac{p_{air}}{R_{air}T} + \frac{p_{vap}}{R_{vap}T} \quad \text{(A.30)}$$

where

p_{air} = partial pressure of dry air
R_{air} = ideal gas constant for dry air = 1716 $^{\text{ft·lb}}/_{\text{slug·R}}$
p_{vap} = partial pressure of water vapor
R_{vap} = ideal gas constant for water vapor = 2760 $^{\text{ft·lb}}/_{\text{slug·R}}$
T = ambient temperature

In order to calculate the partial pressures in the above expression, we must first calculate the saturation pressure p_{sat}, which is the pressure above which water vapor would begin to condense. The saturation pressure is a function of only the ambient temperature (Alduchov and Eskridge, 1996):

$$p_{sat} = 0.08861 \times 10^{4.253T - 136.1/225.3 + 0.5556T} \quad \text{(A.31)}$$

where the units of p_{sat} are *psi* and the units of T are degrees Fahrenheit. The partial pressures are then calculated as

$$p_{vap} = \frac{RH}{100} \cdot p_{sat} \quad \text{(A.32a)}$$

$$p_{air} = p_o - p_{vap} \quad \text{(A.32b)}$$

where RH is relative humidity expressed as per cent and can be any value in the range 0 to 100 and p_o = 14.696 psi is the ambient pressure at sea level. Figure A-5b illustrates the relationship between ambient temperature, relative humidity and air density. In this figure it is clear that humid air density

is much more strongly dependent upon temperature than upon relative humidity.

The equation for the viscosity of humid air is (Tsilingiris, 2008)

$$\mu = \frac{\left[1 - \left(\frac{RH}{100}\right)\left(\frac{p_{sat}}{p_o}\right)\right]}{\left[1 - \left(\frac{RH}{100}\right)\left(\frac{p_{sat}}{p_o}\right)\right] + \left(\frac{RH}{100}\right)\left(\frac{p_{sat}}{p_o}\right)\Phi_{av}} \cdot \mu_{air} + \frac{\left(\frac{RH}{100}\right)\left(\frac{p_{sat}}{p_o}\right)}{\left(\frac{RH}{100}\right)\left(\frac{p_{sat}}{p_o}\right) + \left[1 - \left(\frac{RH}{100}\right)\left(\frac{p_{sat}}{p_o}\right)\right]\Phi_{va}} \cdot \mu_{vap} \quad \text{(A.33)}$$

where the viscosities of dry air and pure water vapor are given by

$$\mu_{air} = \mu_{ref}\left(\frac{T + 459.67}{T_{ref}}\right)^{3/2}\left(\frac{T_{ref} + S}{T + 459.67 + S}\right) \qquad \text{(Fiore, 1966)} \qquad \text{(A.34a)}$$

$$\mu_{vap} = 1.5344 \times 10^{-7} + 4.6418 \times 10^{-10} \cdot T \qquad \text{(Tsilingiris, 2008)} \qquad \text{(A.34b)}$$

The above equation for μ_{air} is called *Sutherland's equation*, and the required constants for dry air are the reference viscosity $\mu_{ref} = 3.584 \times 10^{-7}$ slug/ft·sec, the reference temperature $T_{ref} = 492$ °R and Sutherland's constant $S = 198.6$ °R. In both this equation and the equation for μ_{vap}, the temperature T is in degrees Fahrenheit. In (A.33), Φ_{av} and Φ_{va} are called the *interaction factors* and are calculated by

$$\Phi_{av} = \frac{\sqrt{2}}{4}\left(1 + \frac{MW_{air}}{MW_{vap}}\right)^{-1/2}\left(1 + \left(\frac{\mu_{air}}{\mu_{vap}}\right)^{1/2}\left(\frac{MW_{vap}}{MW_{air}}\right)^{1/4}\right)^2 \qquad \text{(A.35a)}$$

$$\Phi_{va} = \frac{\sqrt{2}}{4}\left(1 + \frac{MW_{vap}}{MW_{air}}\right)^{-1/2}\left(1 + \left(\frac{\mu_{vap}}{\mu_{air}}\right)^{1/2}\left(\frac{MW_{air}}{MW_{vap}}\right)^{1/4}\right)^2 \qquad \text{(A.35b)}$$

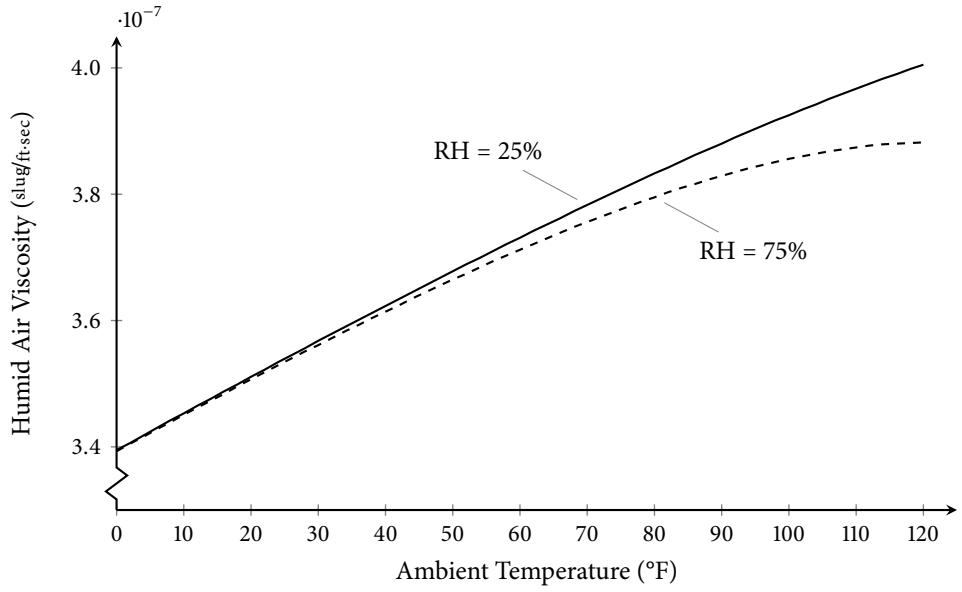
where $MW_{air} = 28.96$ kg/kmol and $MW_{vap} = 18.02$ kg/kmol are the molecular weights of dry air and pure water vapor, respectively. Figure A-5a illustrates the relationship between ambient temperature, relative humidity and air viscosity. As was the case with humid air density, we see that humid air viscosity is also much more strongly dependent upon temperature than upon relative humidity.

A.4 Implementation

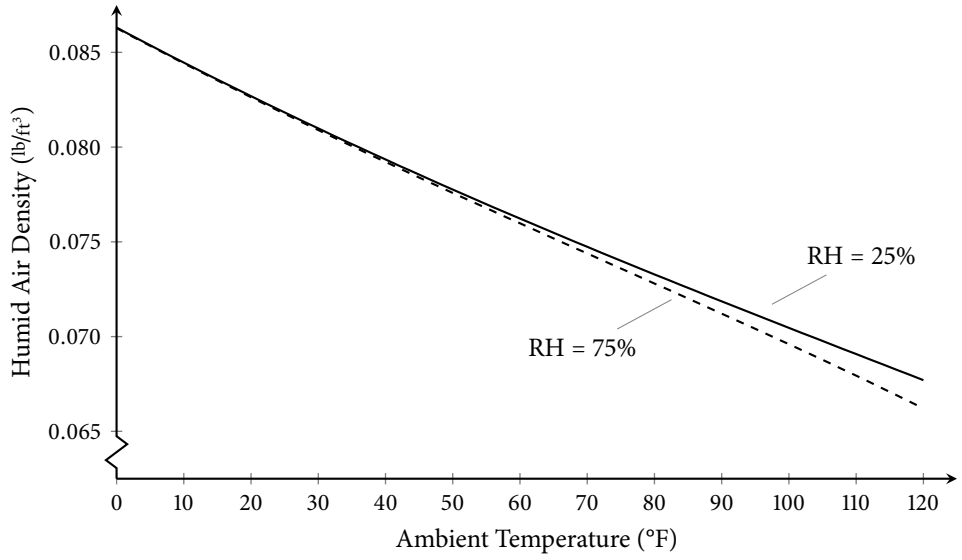
Here we explicitly consider the steps necessary to get from the input (fixed parameters and control factor settings) to the output (ball landing position). The equations below requires units of mass to be slugs, units of length to be feet and units of angles to be radians. Numerical values for all necessary parameters are given in tables A.2 and A.3. It may also be helpful to reference figure A-6. The integration algorithm used is given in Algorithm 1.

I. Precalculate the aerodynamic constants that vary with control factor settings.

   1) $\rho = \rho(RH, T)$



(a) Ambient air viscosity (μ)



(b) Ambient air density (ρ)

Figure A-5: Variation of air properties with ambient temperature and relative humidity

a. Calculate the saturation vapor pressure $p_{sat}(T)$ using Equation A.31.

b. Calculate the partial pressure for water vapor $p_{vap}(p_{sat}, RH)$ using Equation A.32a.

c. Calculate the partial pressure for dry air $p_{air}(p_{vap})$ using Equation A.32b.

d. Calculate the density of humid air $\rho(p_{air}, p_{vap}, T)$ using Equation A.30.

2) $\mu = \mu(RH, T)$

a. Calculate the viscosities of dry air $\mu_{air}(T)$ and water vapor $\mu_{vap}(T)$ using Equations A.34a-A.34b.

b. Calculate the interaction factors $\Phi_{av}(\mu_{air}, \mu_{vap})$ and $\Phi_{va}(\mu_{air}, \mu_{vap})$ using Equations A.35a-A.35b.

c. Calculate the viscosity of humid air $\mu(\mu_{air}, \mu_{vap}, \Phi_{av}, \Phi_{va}, p_{sat}, RH)$ using Equation A.33.

II. Solve the prelaunch dynamics equations to determine the linear and rotational launch speeds.

- Initial conditions: $\theta = \theta_o$, $\omega = 0$.
- Iterate the following until $\theta = 0$:
  - Calculate the x and y components of rubber band length $L_{rbx}(\theta)$ and $L_{rby}(\theta)$ using Equations A.13a-A.13b.
  - Using the updated values of θ, L_{rbx} and L_{rby}, solve the differential equation for ω given by Equation A.18 and the differential equation for θ given by $^{d\theta}/_{dt} = \omega$.
- The ball is launched at $\theta = 0$, in the direction α_ℓ with linear speed calculated by the rotational speed of the catapult arm at launch ω_ℓ times the x and y coordinates of the ball's center. The ball will continue to spin at ω_ℓ throughout the ballistic trajectory.

III. Solve the ballistics dynamics equations to determine the landing position of the ball.

- Initial conditions
  - $x_o = L_b \cdot \cos\alpha_\ell - R_{bc} \cdot \sin\alpha_\ell$
  - $y_o = L_b \cdot \sin\alpha_\ell + R_{bc} \cdot \sin\alpha_\ell$
  - $v_x o = \omega_\ell \cdot x_o$
  - $v_y o = \omega_\ell \cdot y_o$
- Iterate the following until $y = {}^{d_b}/_2$:

- Calculate the magnitude of the velocity vector $\left\|\vec{v}\right\| = \sqrt{v_x^2 + v_y^2}$.
- Calculate the Reynolds number $Re(\left\|\vec{v}\right\|)$ using Equation A.21.
- Calculate the spin ratio $\Omega^\star(\left\|\vec{v}\right\|)$ using Equation A.25, where the angular velocity $\Omega = -\omega_\ell$ is assumed constant from launch to landing.
- Calculate the direction of travel $\alpha = \tan^{-1}(v_y/v_x)$
- Using the updated values of $\left\|\vec{v}\right\|$, Re, $\Omega^\star$ and α, solve the differential equations for v_x and v_y given by Equations A.29a-A.29b and the differential equations for x and y given by $dx/dt = v_x$ and $dy/dt = v^y$.

IV. The system response is the landing position of the ball; that is, the value of x upon reaching the vertical position $y = d_b/2$.

Algorithm 1 Fixed-Step Runge-Kutta (4,5) (Dormand-Prince) ODE Solution

---

Require: $f(t,x) = dx/dt$

function RUNGEKUTTA45($x_o, t_o, t_f, \delta t$)
  $i \leftarrow 0$
  while $t_i < t_f$ **do**
    // Runge-Kutta (4,5) Dormand-Prince constants
    $K_1 \leftarrow \delta t \cdot f\left(t_i, x_i\right)$
    $K_2 \leftarrow \delta t \cdot f\left(t_i + \left(\frac{1}{5}\right)\delta t, x_i + \left(\frac{1}{5}\right)K_1\right)$
    $K_3 \leftarrow \delta t \cdot f\left(t_i + \left(\frac{3}{10}\right)\delta t, x_i + \left(\frac{3}{40}\right)K_1 + \left(\frac{9}{40}\right)K_2\right)$
    $K_4 \leftarrow \delta t \cdot f\left(t_i + \left(\frac{4}{5}\right)\delta t, x_i + \left(\frac{44}{45}\right)K_1 - \left(\frac{56}{15}\right)K_2 + \left(\frac{32}{9}\right)K_3\right)$
    $K_5 \leftarrow \delta t \cdot f\left(t_i + \left(\frac{8}{9}\right)\delta t, x_i + \left(\frac{19372}{6561}\right)K_1 - \left(\frac{25360}{2187}\right)K_2 + \left(\frac{64448}{6561}\right)K_3 - \left(\frac{212}{729}\right)K_4\right)$
    $K_6 \leftarrow \delta t \cdot f\left(t_i + \delta t, x_i + \left(\frac{9017}{3168}\right)K_1 - \left(\frac{355}{33}\right)K_2 + \left(\frac{46372}{5247}\right)K_3 + \left(\frac{49}{176}\right)K_4 - \left(\frac{5103}{18656}\right)K_5\right)$
    $K_7 \leftarrow \delta t \cdot f\left(t_i + \delta t, x_i + \left(\frac{35}{384}\right)K_1 + \left(\frac{500}{1113}\right)K_3 + \left(\frac{125}{192}\right)K_4 - \left(\frac{2187}{6784}\right)K_5 + \left(\frac{11}{84}\right)K_6\right)$
    // 5^{th}-order Runge-Kutta (4,5) Dormand-Prince solution
    $x_{i+1}^{(5)} \leftarrow x_i + \left(\frac{5179}{57600}\right)K_1 + \left(\frac{7571}{216695}\right)K_3 + \left(\frac{393}{640}\right)K_4 - \left(\frac{92097}{339200}\right)K_5 + \left(\frac{187}{2100}\right)K_6 + \left(\frac{1}{40}\right)K_7$
    // update variables
    $t_{i+1} \leftarrow t_i + \delta t$
    $x_{i+1} \leftarrow x_{i+1}^{(5)}$
    $i \leftarrow i + 1$
  end while
  return $x(t) \forall t \in \{t_o, t_f\}$
end function

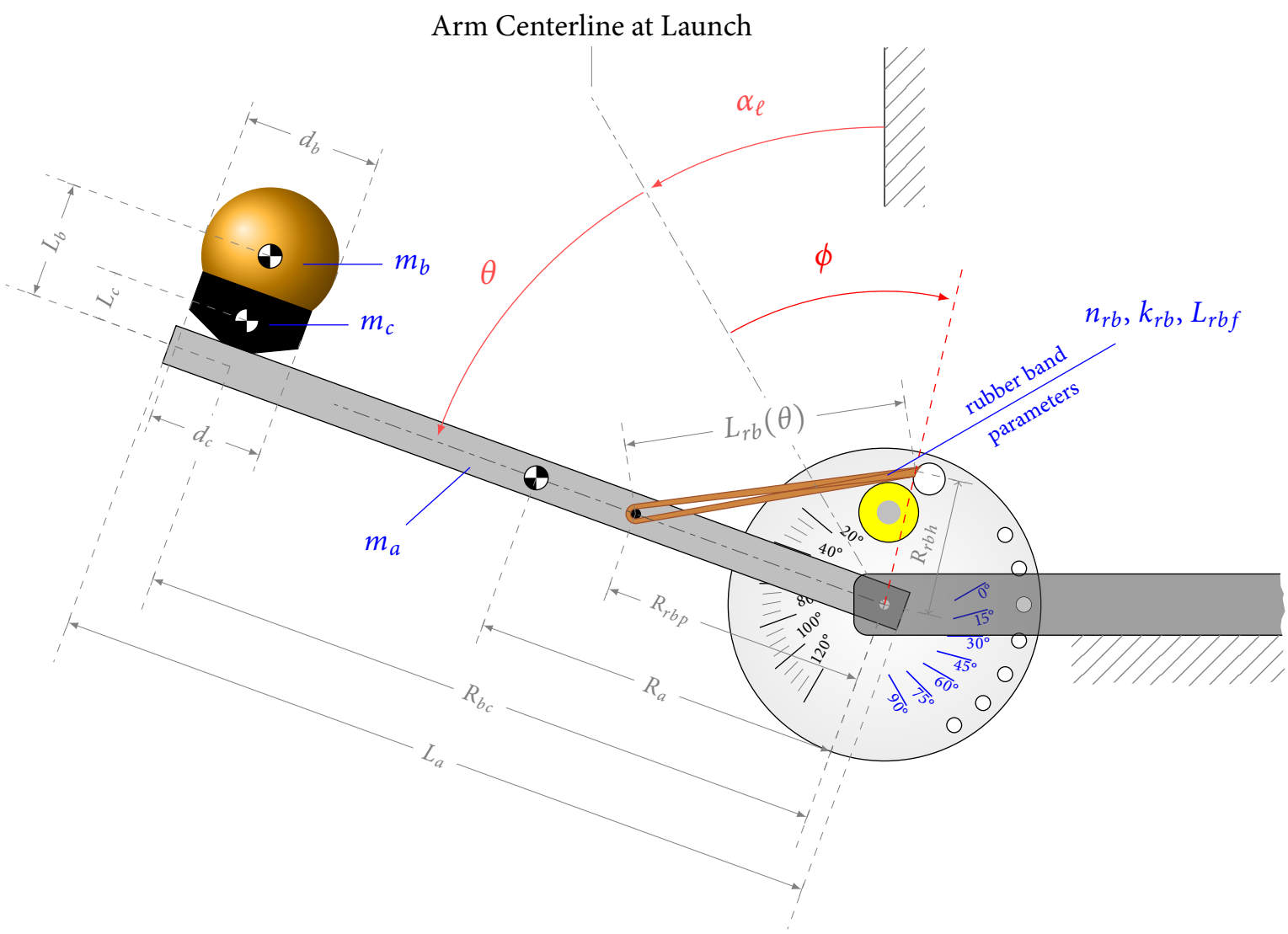


Figure A-6: Engineering constants in the modified catapult model

Table A.2: Control Factors for the Modified Catapult

Control Factor	"-1" Level		"+1" Level	
	setting	parameters	setting	parameters
launch angle	60°	$\alpha_\ell = 60°$	45°	$\alpha_\ell = 45°$
ambient humidity	25%	$RH = 25\%$	75%	$RH = 75\%$
type of	regulation	$d_b = 1.575$ in	"large ball"	$d_b = 1.732$ in
table tennis ball		$m_b = 0.094$ oz		$m_b = 0.081$ oz
		$L_b = 1.432$ in		$L_b = 1.599$ in
arm material	aluminum	$m_a = 1.840$ oz	magnesium	$m_a = 1.186$ oz
number of rubber bands	3	$n_{rb} = 3$	2	$n_{rb} = 2$
ambient temperature	72°F	$T = 72$°F	32°F	$T = 32$°F
initial pullback	40°	$\theta_o = 40°$	30°	$\theta_o = 30°$

Table A.3: Parameter Specifications for the Modified Catapult

Parameter	Description	Value
α_ℓ	launch angle	varies[a]
θ	pullback	varies[a]
μ	dynamic viscosity of air	function of RH and T
ρ	density of air	function of RH and T
C_d	aerodynamic drag coefficient	0.5[b]
C_M	Magnus effect lift coefficient	varies[a b]
d_b	ball diameter	varies[a]
d_c	cup diameter	1.45 in
k_{rb}	rubber band "spring constant"	0.5 $^{\text{lbf}}/_{\text{in}}$
L_a	length of catapult arm	9.81 in
L_b	ball c.g. offset length	varies[a]
L_c	cup c.g. offset length	0.516 in
L_{rb}	length of rubber band(s)	function of θ
L_{rbf}	"free" length of the rubber band	2.93 in[b]
m_a	mass of catapult arm	varies[a]
m_b	mass of ball	varies[a]
m_c	mass of cup	0.275 oz
n_{rb}	number of rubber bands	varies[a]
R_a	radial position of arm c.g.	4.66 in
R_{bc}	radial position of ball and cup	8.75 in
R_{rbp}	radial position of rubber band pin	3.33 in
R_{rbh}	radial position of rubber band hole	1.72 in
RH	ambient relative humidity	varies[a b]
T	ambient temperature	varies[a b]

[a] This value is either a control factor or directly dependent upon the value of a control factor.
[b] This value is not shown in Figure A-6, but it is used in the simulation of the post launch dynamics.

A.5 Full Factorial Results with Chosen Control Factors

Using the model presented here, the control factors and their levels given in table A.2, and the parameter specifications given in table A.3, we may calculate the response of the system (i.e., landing position) for each configuration in the full factorial design space. It may be helpful to reference the labeled engineering sketch of the catapult given in figure A-6. The numerical results of the full factorial response are given in table A.4. We may use the results to generate "main effects" plots for each of the control factors, as shown in figure A-7. Here we see that the overall average landing position is about 92 inches, and the relative importance of the control factors in decreasing order is (1) Number of Rubber Bands, (2) Pullback, (3) Arm Material, (4) Launch Angle, (5) Type of Ball, (6) Ambient Temperature and (7) Relative Humidity. Another important consideration that may be visualized is the two-factor

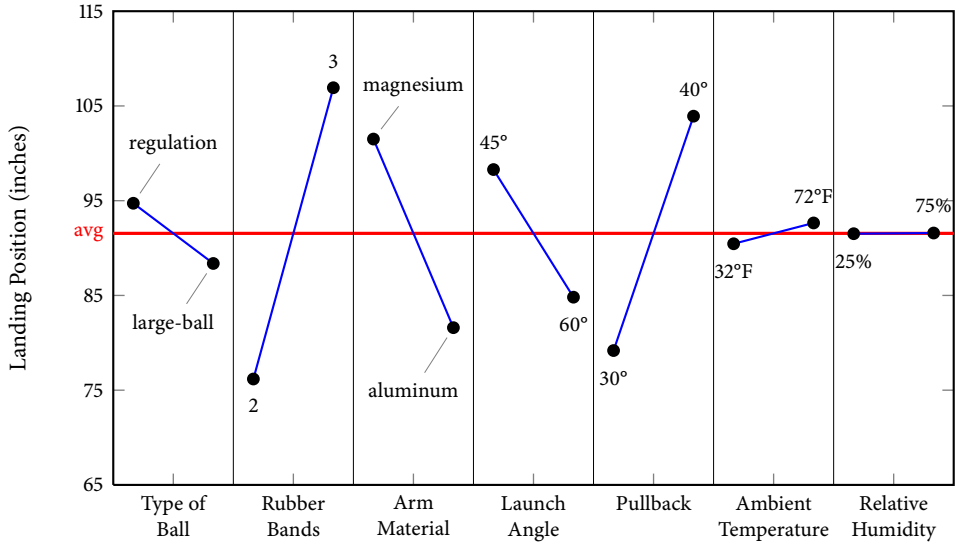
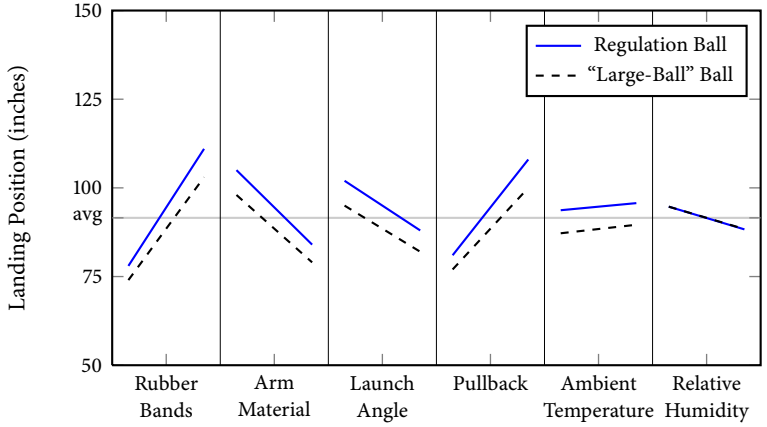
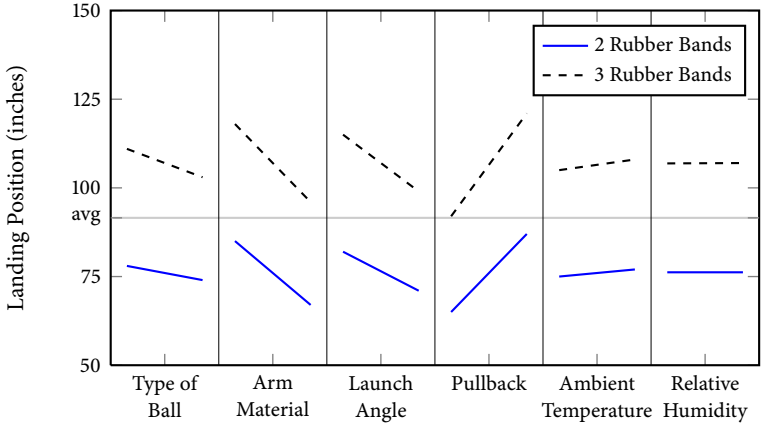


Figure A-7: Main effects from full factorial results

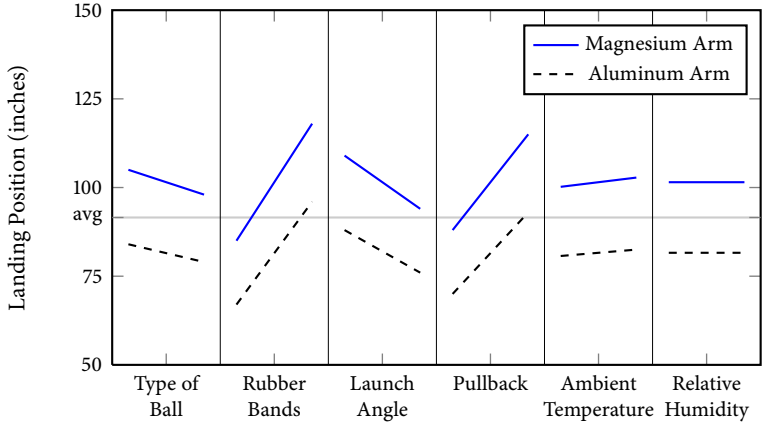
interactions in the system. Again using the numerical results, we may construct “interaction effects” plots for each pair of control factors. These are shown in figure A-8, grouped by each of the seven factors. Although convenient, there is obviously much redundant information presented here. In these interaction plots, four points are charted for each pairing of factors. If the factors are A and B with two levels each, we view the interactions by plotting the pair of lines $(A_1, B_1) - (A_1, B_2)$ and $(A_2, B_1) - (A_2, B_2)$, or the pair of lines $(A_1, B_1) - (A_2, B_1)$ and $(A_1, B_2) - (A_2, B_2)$. If no interaction effect is present, either of these pairs of lines will be parallel to each other. If an interaction effect is present, the stronger the interaction, the greater the degree to which the lines are not parallel. In reviewing the ensemble of two-factor interaction plots shown in figure A-8, we can quickly see that all pairs of lines appear to be close to parallel. Thus, there are no strong two-factor interaction effects in this system. This being the case, we can also assume that higher-order interaction effects and any confounding between interaction effects and main effects is negligible.



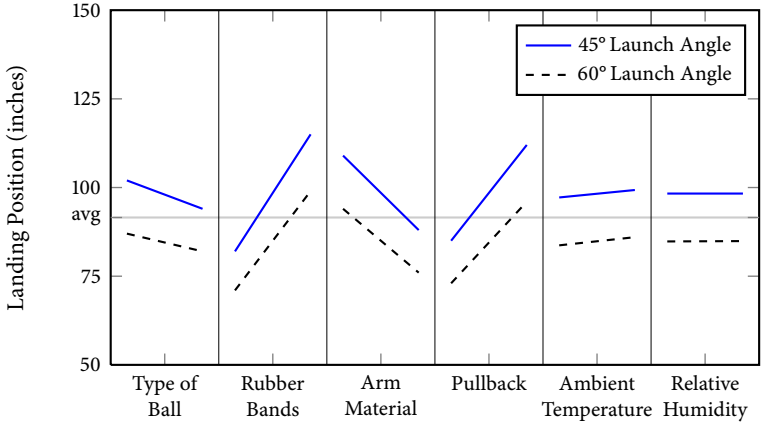
(a) Type of ball (blue=regulation, black=large-ball)



(b) Number of rubber bands (blue=3, black=2)

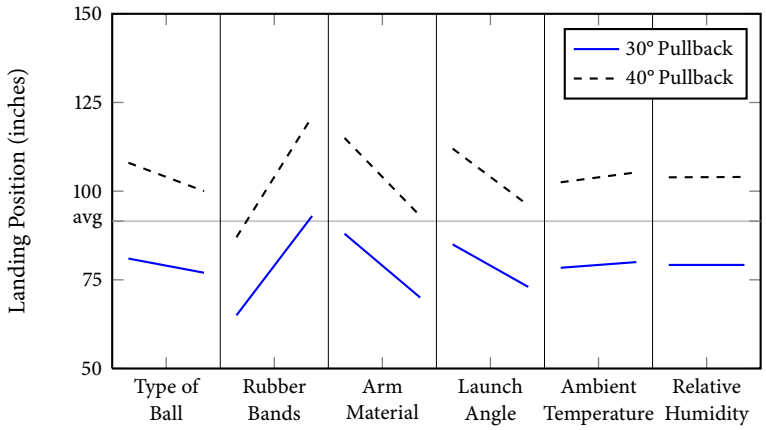


(c) Arm material (blue=aluminum, black=magnesium)

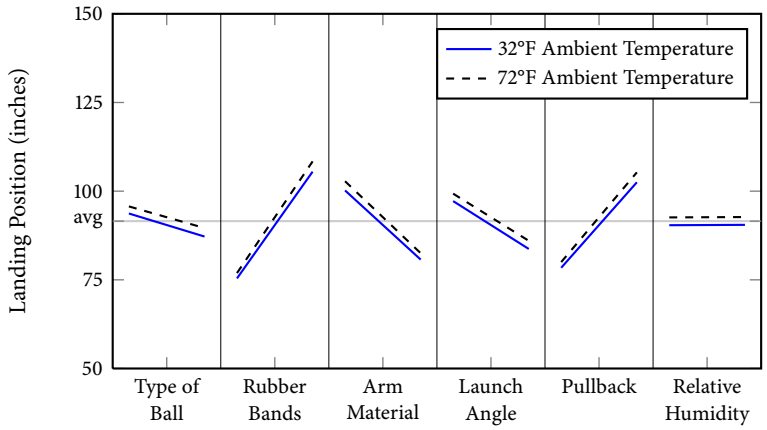


(d) Launch angle (blue=60°, black=45°)

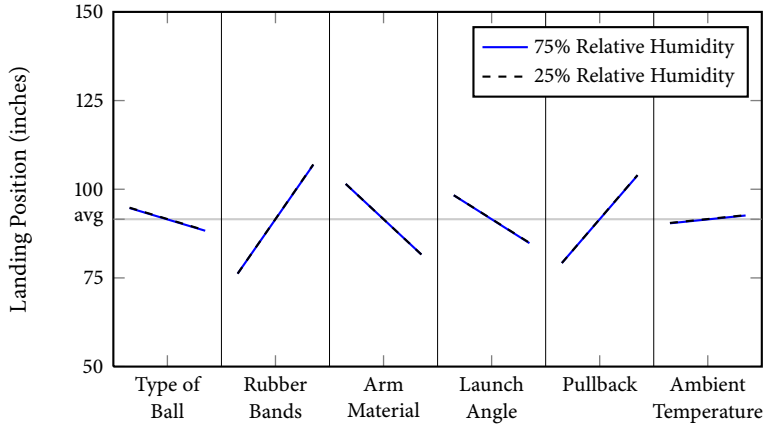
Figure A-8: Two-factor interactions from full factorial results



(e) Pullback (blue=40°, black=30°)



(f) Ambient temperature (blue=72°F, black=32°F)



(g) Relative humidity (blue=25%, black=75%)

Figure A-8: Two-factor interactions from full factorial results (continued)

Table A.4: Full Factorial Results for Catapult Simulation

Trial	Type of Table Tennis Ball	Rubber Bands	Arm Material	Launch Angle	Pullback	Ambient Temperature	Relative Humidity	Landing Position
1	Regulation (white)	2	Magnesium	45°	30°	32°F	25%	79.8 in
2	Large-Ball (orange)	2	Magnesium	45°	30°	32°F	25%	76.1 in
3	Regulation (white)	3	Magnesium	45°	30°	32°F	25%	113.5 in
4	Large-Ball (orange)	3	Magnesium	45°	30°	32°F	25%	105.3 in
5	Regulation (white)	2	Aluminum	45°	30°	32°F	25%	62.1 in
6	Large-Ball (orange)	2	Aluminum	45°	30°	32°F	25%	60.0 in
7	Regulation (white)	3	Aluminum	45°	30°	32°F	25%	90.9 in
8	Large-Ball (orange)	3	Aluminum	45°	30°	32°F	25%	85.7 in
9	Regulation (white)	2	Magnesium	60°	30°	32°F	25%	68.9 in
10	Large-Ball (orange)	2	Magnesium	60°	30°	32°F	25%	65.6 in
11	Regulation (white)	3	Magnesium	60°	30°	32°F	25%	98.2 in
12	Large-Ball (orange)	3	Magnesium	60°	30°	32°F	25%	90.6 in
13	Regulation (white)	2	Aluminum	60°	30°	32°F	25%	53.0 in
14	Large-Ball (orange)	2	Aluminum	60°	30°	32°F	25%	51.3 in
15	Regulation (white)	3	Aluminum	60°	30°	32°F	25%	78.7 in
16	Large-Ball (orange)	3	Aluminum	60°	30°	32°F	25%	73.9 in
17	Regulation (white)	2	Magnesium	45°	40°	32°F	25%	106.6 in
18	Large-Ball (orange)	2	Magnesium	45°	40°	32°F	25%	99.5 in
19	Regulation (white)	3	Magnesium	45°	40°	32°F	25%	147.4 in
20	Large-Ball (orange)	3	Magnesium	45°	40°	32°F	25%	133.2 in
21	Regulation (white)	2	Aluminum	45°	40°	32°F	25%	84.3 in
22	Large-Ball (orange)	2	Aluminum	45°	40°	32°F	25%	79.9 in
23	Regulation (white)	3	Aluminum	45°	40°	32°F	25%	120.3 in
24	Large-Ball (orange)	3	Aluminum	45°	40°	32°F	25%	110.8 in
25	Regulation (white)	2	Magnesium	60°	40°	32°F	25%	92.3 in
26	Large-Ball (orange)	2	Magnesium	60°	40°	32°F	25%	85.6 in
27	Regulation (white)	3	Magnesium	60°	40°	32°F	25%	126.5 in
28	Large-Ball (orange)	3	Magnesium	60°	40°	32°F	25%	113.3 in
29	Regulation (white)	2	Aluminum	60°	40°	32°F	25%	72.8 in
30	Large-Ball (orange)	2	Aluminum	60°	40°	32°F	25%	68.8 in
31	Regulation (white)	3	Aluminum	60°	40°	32°F	25%	104.0 in
32	Large-Ball (orange)	3	Aluminum	60°	40°	32°F	25%	95.1 in

Full Factorial Results for Catapult Simulation (continued)

Trial	Type of Table Tennis Ball	Rubber Bands	Arm Material	Launch Angle	Pullback	Ambient Temperature	Relative Humidity	Landing Position
33	Regulation (white)	2	Magnesium	45°	30°	72°F	25%	80.9 in
34	Large-Ball (orange)	2	Magnesium	45°	30°	72°F	25%	77.5 in
35	Regulation (white)	3	Magnesium	45°	30°	72°F	25%	115.8 in
36	Large-Ball (orange)	3	Magnesium	45°	30°	72°F	25%	108.0 in
37	Regulation (white)	2	Aluminum	45°	30°	72°F	25%	62.8 in
38	Large-Ball (orange)	2	Aluminum	45°	30°	72°F	25%	60.8 in
39	Regulation (white)	3	Aluminum	45°	30°	72°F	25%	92.3 in
40	Large-Ball (orange)	3	Aluminum	45°	30°	72°F	25%	87.5 in
41	Regulation (white)	2	Magnesium	60°	30°	72°F	25%	70.2 in
42	Large-Ball (orange)	2	Magnesium	60°	30°	72°F	25%	67.2 in
43	Regulation (white)	3	Magnesium	60°	30°	72°F	25%	100.7 in
44	Large-Ball (orange)	3	Magnesium	60°	30°	72°F	25%	93.4 in
45	Regulation (white)	2	Aluminum	60°	30°	72°F	25%	53.8 in
46	Large-Ball (orange)	2	Aluminum	60°	30°	72°F	25%	52.2 in
47	Regulation (white)	3	Aluminum	60°	30°	72°F	25%	80.3 in
48	Large-Ball (orange)	3	Aluminum	60°	30°	72°F	25%	75.9 in
49	Regulation (white)	2	Magnesium	45°	40°	72°F	25%	108.6 in
50	Large-Ball (orange)	2	Magnesium	45°	40°	72°F	25%	101.9 in
51	Regulation (white)	3	Magnesium	45°	40°	72°F	25%	151.3 in
52	Large-Ball (orange)	3	Magnesium	45°	40°	72°F	25%	137.6 in
53	Regulation (white)	2	Aluminum	45°	40°	72°F	25%	85.5 in
54	Large-Ball (orange)	2	Aluminum	45°	40°	72°F	25%	81.4 in
55	Regulation (white)	3	Aluminum	45°	40°	72°F	25%	122.9 in
56	Large-Ball (orange)	3	Aluminum	45°	40°	72°F	25%	113.8 in
57	Regulation (white)	2	Magnesium	60°	40°	72°F	25%	94.5 in
58	Large-Ball (orange)	2	Magnesium	60°	40°	72°F	25%	88.2 in
59	Regulation (white)	3	Magnesium	60°	40°	72°F	25%	130.5 in
60	Large-Ball (orange)	3	Magnesium	60°	40°	72°F	25%	117.8 in
61	Regulation (white)	2	Aluminum	60°	40°	72°F	25%	74.2 in
62	Large-Ball (orange)	2	Aluminum	60°	40°	72°F	25%	70.5 in
63	Regulation (white)	3	Aluminum	60°	40°	72°F	25%	106.7 in
64	Large-Ball (orange)	3	Aluminum	60°	40°	72°F	25%	98.3 in

Full Factorial Results for Catapult Simulation (continued)

Trial	Type of Table Tennis Ball	Rubber Bands	Arm Material	Launch Angle	Pullback	Ambient Temperature	Relative Humidity	Landing Position
65	Regulation (white)	2	Magnesium	45°	30°	32°F	75%	79.9 in
66	Large-Ball (orange)	2	Magnesium	45°	30°	32°F	75%	76.2 in
67	Regulation (white)	3	Magnesium	45°	30°	32°F	75%	113.6 in
68	Large-Ball (orange)	3	Magnesium	45°	30°	32°F	75%	105.3 in
69	Regulation (white)	2	Aluminum	45°	30°	32°F	75%	62.1 in
70	Large-Ball (orange)	2	Aluminum	45°	30°	32°F	75%	60.0 in
71	Regulation (white)	3	Aluminum	45°	30°	32°F	75%	90.9 in
72	Large-Ball (orange)	3	Aluminum	45°	30°	32°F	75%	85.7 in
73	Regulation (white)	2	Magnesium	60°	30°	32°F	75%	69.0 in
74	Large-Ball (orange)	2	Magnesium	60°	30°	32°F	75%	65.6 in
75	Regulation (white)	3	Magnesium	60°	30°	32°F	75%	98.3 in
76	Large-Ball (orange)	3	Magnesium	60°	30°	32°F	75%	90.6 in
77	Regulation (white)	2	Aluminum	60°	30°	32°F	75%	53.0 in
78	Large-Ball (orange)	2	Aluminum	60°	30°	32°F	75%	51.3 in
79	Regulation (white)	3	Aluminum	60°	30°	32°F	75%	78.7 in
80	Large-Ball (orange)	3	Aluminum	60°	30°	32°F	75%	73.9 in
81	Regulation (white)	2	Magnesium	45°	40°	32°F	75%	106.7 in
82	Large-Ball (orange)	2	Magnesium	45°	40°	32°F	75%	99.5 in
83	Regulation (white)	3	Magnesium	45°	40°	32°F	75%	147.5 in
84	Large-Ball (orange)	3	Magnesium	45°	40°	32°F	75%	133.3 in
85	Regulation (white)	2	Aluminum	45°	40°	32°F	75%	84.3 in
86	Large-Ball (orange)	2	Aluminum	45°	40°	32°F	75%	79.9 in
87	Regulation (white)	3	Aluminum	45°	40°	32°F	75%	120.4 in
88	Large-Ball (orange)	3	Aluminum	45°	40°	32°F	75%	110.8 in
89	Regulation (white)	2	Magnesium	60°	40°	32°F	75%	92.3 in
90	Large-Ball (orange)	2	Magnesium	60°	40°	32°F	75%	85.7 in
91	Regulation (white)	3	Magnesium	60°	40°	32°F	75%	126.5 in
92	Large-Ball (orange)	3	Magnesium	60°	40°	32°F	75%	113.4 in
93	Regulation (white)	2	Aluminum	60°	40°	32°F	75%	72.8 in
94	Large-Ball (orange)	2	Aluminum	60°	40°	32°F	75%	68.9 in
95	Regulation (white)	3	Aluminum	60°	40°	32°F	75%	104.0 in
96	Large-Ball (orange)	3	Aluminum	60°	40°	32°F	75%	95.1 in

Full Factorial Results for Catapult Simulation (continued)

Trial	Type of Table Tennis Ball	Rubber Bands	Arm Material	Launch Angle	Pullback	Ambient Temperature	Relative Humidity	Landing Position
97	Regulation (white)	2	Magnesium	45°	30°	72°F	75%	81.0 in
98	Large-Ball (orange)	2	Magnesium	45°	30°	72°F	75%	77.6 in
99	Regulation (white)	3	Magnesium	45°	30°	72°F	75%	115.9 in
100	Large-Ball (orange)	3	Magnesium	45°	30°	72°F	75%	108.2 in
101	Regulation (white)	2	Aluminum	45°	30°	72°F	75%	62.8 in
102	Large-Ball (orange)	2	Aluminum	45°	30°	72°F	75%	60.9 in
103	Regulation (white)	3	Aluminum	45°	30°	72°F	75%	92.4 in
104	Large-Ball (orange)	3	Aluminum	45°	30°	72°F	75%	87.6 in
105	Regulation (white)	2	Magnesium	60°	30°	72°F	75%	70.2 in
106	Large-Ball (orange)	2	Magnesium	60°	30°	72°F	75%	67.3 in
107	Regulation (white)	3	Magnesium	60°	30°	72°F	75%	100.8 in
108	Large-Ball (orange)	3	Magnesium	60°	30°	72°F	75%	93.6 in
109	Regulation (white)	2	Aluminum	60°	30°	72°F	75%	53.8 in
110	Large-Ball (orange)	2	Aluminum	60°	30°	72°F	75%	52.3 in
111	Regulation (white)	3	Aluminum	60°	30°	72°F	75%	80.4 in
112	Large-Ball (orange)	3	Aluminum	60°	30°	72°F	75%	76.0 in
113	Regulation (white)	2	Magnesium	45°	40°	72°F	75%	108.7 in
114	Large-Ball (orange)	2	Magnesium	45°	40°	72°F	75%	102.0 in
115	Regulation (white)	3	Magnesium	45°	40°	72°F	75%	151.5 in
116	Large-Ball (orange)	3	Magnesium	45°	40°	72°F	75%	137.8 in
117	Regulation (white)	2	Aluminum	45°	40°	72°F	75%	85.5 in
118	Large-Ball (orange)	2	Aluminum	45°	40°	72°F	75%	81.5 in
119	Regulation (white)	3	Aluminum	45°	40°	72°F	75%	123.0 in
120	Large-Ball (orange)	3	Aluminum	45°	40°	72°F	75%	114.0 in
121	Regulation (white)	2	Magnesium	60°	40°	72°F	75%	94.6 in
122	Large-Ball (orange)	2	Magnesium	60°	40°	72°F	75%	88.4 in
123	Regulation (white)	3	Magnesium	60°	40°	72°F	75%	130.7 in
124	Large-Ball (orange)	3	Magnesium	60°	40°	72°F	75%	118.0 in
125	Regulation (white)	2	Aluminum	60°	40°	72°F	75%	74.2 in
126	Large-Ball (orange)	2	Aluminum	60°	40°	72°F	75%	70.6 in
127	Regulation (white)	3	Aluminum	60°	40°	72°F	75%	106.8 in
128	Large-Ball (orange)	3	Aluminum	60°	40°	72°F	75%	98.4 in

A.6 Catapult Model References

Alduchov, O. A., and R. E. Eskridge, "Improved Magnus Form Approximation of Saturation Vapor Pressure," *Journal of Applied Meteorology*, Vol. 35, No. 4, 1996, pp. 601–609.

Bearman, P. W., and J. K. Harvey, "Golf Ball Aerodynamics," *The Aeronautical Quarterly*, Vol. 27, 1976, pp. 112–122.

Briggs, L. J., "Effect of Spin and Speed on the Lateral Deflection (Curve) of a Baseball; and the Magnus Effect for Smooth Spheres," *American Journal of Physics*, Vol. 27, 1959, pp. 589–596.

Cheng, N.-S., "Comparison of formulas for drag coefficient and settling velocity of spherical particles," *Powder Technology*, Vol. 189, No. 3, 2008, pp. 395–398.

Davies, J. M., "The Aerodynamics of Golf Balls," *Journal of Applied Physics*, Vol. 20, No. 9, 1949, pp. 821–828.

Fiore, A. W., "Viscosity of air," *Journal of Spacecraft and Rockets*, Vol. 3, No. 5, 1966, pp. 756–758.

Loth, E., "Lift of a Solid Spherical Particle Subject to Vorticity and/or Spin," *AIAA Journal*, Vol. 46, No. 4, 2008, pp. 801–809.

Maccoll, J. W., "Aerodynamics of a spinning sphere," *The Journal of the Royal Aeronautical Society*, Vol. 32, 1928, pp. 777–798.

Mehta, R. D., "Aerodynamics of Sports Balls," *Annual Review of Fluid Mechanics*, Vol. 17, 1985, pp. 151–189.

Mehta, R. D., and J. M. Pallis, "Sports Ball Aerodynamics: Effects of Velocity, Spin and Surface Roughness," in *Materials and Science in Sports*, edited by F. H. Froes and S. J. Haake, The Minerals, Metals and Materials Society, San Diego, CA, 2001, pp. 185–197.

Nathan, A. M., "The effect of spin on the flight of a baseball," *American Journal of Physics*, Vol. 76, No. 2, 2008, pp. 119–124.

Tsilingiris, P. T., "Thermophysical and transport properties of humid air at temperature range between 0 and 100°C," *Energy Conversion and Management*, Vol. 49, 2008, pp. 1098–1110.

Watts, R. G., and R. Ferrer, "The lateral force on a spinning sphere: Aerodynamics of a curveball," *American Journal of Physics*, Vol. 55, No. 1, 1987, pp. 40–44.

This page intentionally left blank.

Appendix B

Forms and Supplementary Materials for All Experiments

This appendix presents all paper materials used in conducting the pilot and main studies described in this thesis. Forms that did not change in later versions of the experiment are not included more than once, but the following list of contents may be used as a guide to which forms were used in which of the experiments.

Contents

Pilot Study
Administrator Script
(2 pages)

This page intentionally left blank.

1 Introduction

1.1 Listen to a description of the experiment format

In this experiment, youThll be performing a short design task for a simple configurable physical system, to determine which configuration will best satisfy the design objective.

After the necessary paperwork has been completed, I'll describe the configurable system, the design objective and the design method, then we'll get started with the task.

1.2 Read and sign the consent form

The consent form broadly describes the purpose and format of this experiment, and your rights and protections as a test subject. It is a requirement of the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) that each test subject read and sign the consent form before proceeding.

A few highlights from the consent form are that

1. As part of the data set, I will be recording everything we discuss following the initial paperwork. I'm using this digital recorder that generates and stores .mp3 files of the audio, which I will then transcribe into a written version. The audio files will be kept secure by me until they are no longer needed, then destroyed. The text transcript will not be connected to your name, only a random subject ID number that cannot be traced back to you. The text file is archived for five years then destroyed per COUHES regulations.
2. Results will be published only in aggregate form based on the statistics of all test subjects' results.
3. The results are confidential. I am not allowed to speak about your performance with anyone.
4. In order to use the data in any other way, I would first contact you to request your permission.

1.3 Provide information about your educational and work background (area of technical expertise, degree(s) earned, years of full-time working experience) that may assist us in analyzing the results.

I'm collecting demographic data that could potentially help to explain trends in the experimental results. Please provide as much detail as you feel comfortable doing.

2 Training

2.1 Listen to a description of the physical device, its configuration options and the performance goal(s) to be achieved.

(*Give the catapult overview sketch to the test subject, to reference while speaking about it.*) The catapult is operated to launch a table-tennis-sized ball:

1. The ball is placed on the rim of the cup at the end of the catapult arm.
2. The catapult arm is grasped just behind the cup and pulled back.
3. The arm is released from rest, and the stretched rubber band(s) force the arm to be rotated until it hits the yellow stopper.
4. The arm is prevented from rotating further, but the ball leaves the cup.
5. After a short flight, the ball lands at some distance from the catapult.

We may control the settings of five configuration options for the catapult and two for the ambient environment in which it operates:

Launch Angle	Relative Humidity	Type of Ball	Arm Material	Rubber Bands	Ambient Temperature	Pullback
60°	25%	Regulation Table Tennis	Aluminum	3	72°F	40°
45°	75%	Large-Ball Table Tennis	Magnesium	2	32°F	30°

To gauge your comfort level with the physics involved here, I'm going to go through each of these control factors one at a time, asking you to give a rough estimate of what you think will happen to the landing distance when going from one setting to the other. Your options are

Large Decrease Small Decrease Negligible Effect Small Increase Large Increase

Please briefly give your reason for your answer, and how confident you are that it is correct, on a scale from 1 to 5, 5 being most confident.

Pilot Study First Group
Consent Form
(4 pages)

This page intentionally left blank.

CONSENT TO PARTICIPATE IN
NON-BIOMEDICAL RESEARCH

THE EFFECT OF EXPERIMENTAL DESIGN METHOD ON ENGINEERING JUDGMENT IN USING COMPUTER SIMULATIONS FOR DESIGN, Part II

You are asked to participate in a research study conducted by Dan Frey and Troy Savoie, from the Mechanical Engineering Department at the Massachusetts Institute of Technology (M.I.T.). The results of this study will contribute to research papers and a doctoral thesis. You were selected as a possible participant in this study because you are either an engineer or an engineering student with the appropriate technical background. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

• PARTICIPATION AND WITHDRAWAL

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

• PURPOSE OF THE STUDY

The purpose of this study is to understand certain aspects of using computer simulations with structured design methods in optimizing a physical system. Further details about the nature of the study will be revealed at the end of your participation.

• PROCEDURES

If you volunteer to participate in this study, we would ask you to do the following things:

I. Introduction (10 min)
   a) Listen to a description of the experiment format.
   b) Read and sign this consent form.
   c) Provide information about your educational and work background (area of technical expertise, degree(s) earned, years of full-time working experience) that may assist us in analyzing the results.

II. Training (10 min)

a) Listen to a description of the physical device, its configuration options and the performance goal(s) to be achieved.

b) Listen to a description of the design space sampling method to be used.

III. Design Problem (30 min) - For each experiment in the design method:

a) Study the system configuration,

b) Predict the performance in this configuration,

c) Explain the reasoning supporting the prediction,

d) Disclose your level of confidence in the prediction,

e) Observe the outcome of the simulation results,

f) Discuss possible reasons for any differences between your prediction and the simulation results.

IV. Conclusion (10 min)

a) Provide feedback regarding your experience in this study.

b) Discuss any questions you may have about the study.

c) Learn further details about the nature of the study.

Your participation in this study should take about one hour to complete.

• POTENTIAL RISKS AND DISCOMFORTS

None.

• POTENTIAL BENEFITS

It is hoped you will find your participation in this experiment interesting and educational. You will follow a specified method of experimental design and interact with a computer simulation of a physical system, therefore you may learn about these tools commonly used by contemporary design engineers.

It is also hoped that this and subsequent experiments will help organizations (companies, governments, etc.) to make more effective use of simulations in making design decisions.

• PAYMENT FOR PARTICIPATION

None.

• CONFIDENTIALITY

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

Any discussion between you and the test administrator will be recorded using a portable digital recorder and saved as an audio file. This recording will begin after this consent form has been signed, and will continue through the end of your participation in the study. You have the right to listen to this audio recording and to edit parts or all of it at your discretion. The audio file will be used by the test administrator to create a text transcript of the discussion, and this text transcript will become part of the data set to be analyzed. The audio file will be deleted after the text transcript is created.

The data from this study will be associated with a subject number which helps us to link the experimental results with your survey information (education, experience, etc). The records will be archived in a filing cabinet behind two locked doors for five years and then will be destroyed. The data will be used in research papers and a doctoral thesis, but only in the aggregate after statistical analysis.

• IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact Daniel D. Frey who is the Principal Investigator. He can be reached at room 3-449D, (617) 324-6133, and danfrey@mit.edu.

• EMERGENCY CARE AND COMPENSATION FOR INJURY

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to the M.I.T. Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

• RIGHTS OF RESEARCH SUBJECTS

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone (617) 253-6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

_____
Name of Subject

_____ _____
Signature of Subject or Legal Representative Date

_____
Email address or telephone number
(in case we need to contact you later regarding your participation in this study)

SIGNATURE OF INVESTIGATOR

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

_____ _____
Signature of Investigator Date

Demographic Data Form
(1 page)

This page intentionally left blank.

Demographic Data Form

Population Demographics. We would like to collect the following demographic information that may help explain trends in the data. This information is confidential and will only be associated with the random subject identification number, which is not connected with your name. Any analysis based on this data will be published only in aggregate form.

1. Do you have any limitation(s) in operating a basic calculator?

☐ Yes ☐ No

2. Is your vision 20/20 or correctible to 20/20 in both eyes?

☐ Yes ☐ No

3. What is your gender?

☐ Female ☐ Male

4. Into which of the following age groups do you fit? Select the older group if you are on the boundary.

☐ <25 ☐ 25-35 ☐ 35-45 ☐ 45-55 ☐ >55

5. Education.

(a) Bachelor's, please specify field _____

☐ In progress, percent complete: _____ ☐ Completed, year: _____

(b) Master's, please specify field _____

☐ In progress, percent complete: _____ ☐ Completed, year: _____

(c) Doctorate, please specify field _____

☐ In progress, percent complete: _____ ☐ Completed, year: _____

(d) Other, please specify _____

☐ In progress, percent complete: _____ ☐ Completed, year: _____

6. Experience.

(a) How many years of full time engineering or science experience do you have?

☐ < 1 ☐ 1-2 ☐ 2-5 ☐ 5-10 ☐ 10-15 ☐ 15-20 ☐ > 20

(b) In what field is most of your full-time working experience? _____

7. Experience using design of experiments.

(a) How would you describe yourself in terms of experience using design of experiments?

☐ Novice ☐ Intermediate ☐ Expert

(b) Which of the following design methods have you used?

☐ Screening Design ☐ Response Surface Design
☐ Full Factorial Design ☐ Mixture Design
☐ Space Filling Design ☐ Nonlinear Design
☐ Taguchi Arrays ☐ Other, please specify _____

8. Experience using computer simulations.

(a) How would you describe yourself in terms of experience using computer simulation?

☐ Novice ☐ Intermediate ☐ Experienced ☐ Developer

(b) How many years of computer simulation experience do you have?

☐ < 1 ☐ 1-2 ☐ 2-5 ☐ 5-10 ☐ 10-15 ☐ 15-20 ☐ > 20

(c) Which application(s) have you used for simulation?

☐ Spreadsheet (e.g. Microsoft Excel)
☐ Commercial engineering software, please specify _____
☐ High-level programming language, please specify _____
☐ Other, please specify _____

9. In which system of units are you more comfortable working?

☐ English/American ☐ SI/metric ☐ equally comfortable in both

This page intentionally left blank.

Pilot Study
Administrator Graphical Aids
(2 pages)

This page intentionally left blank.

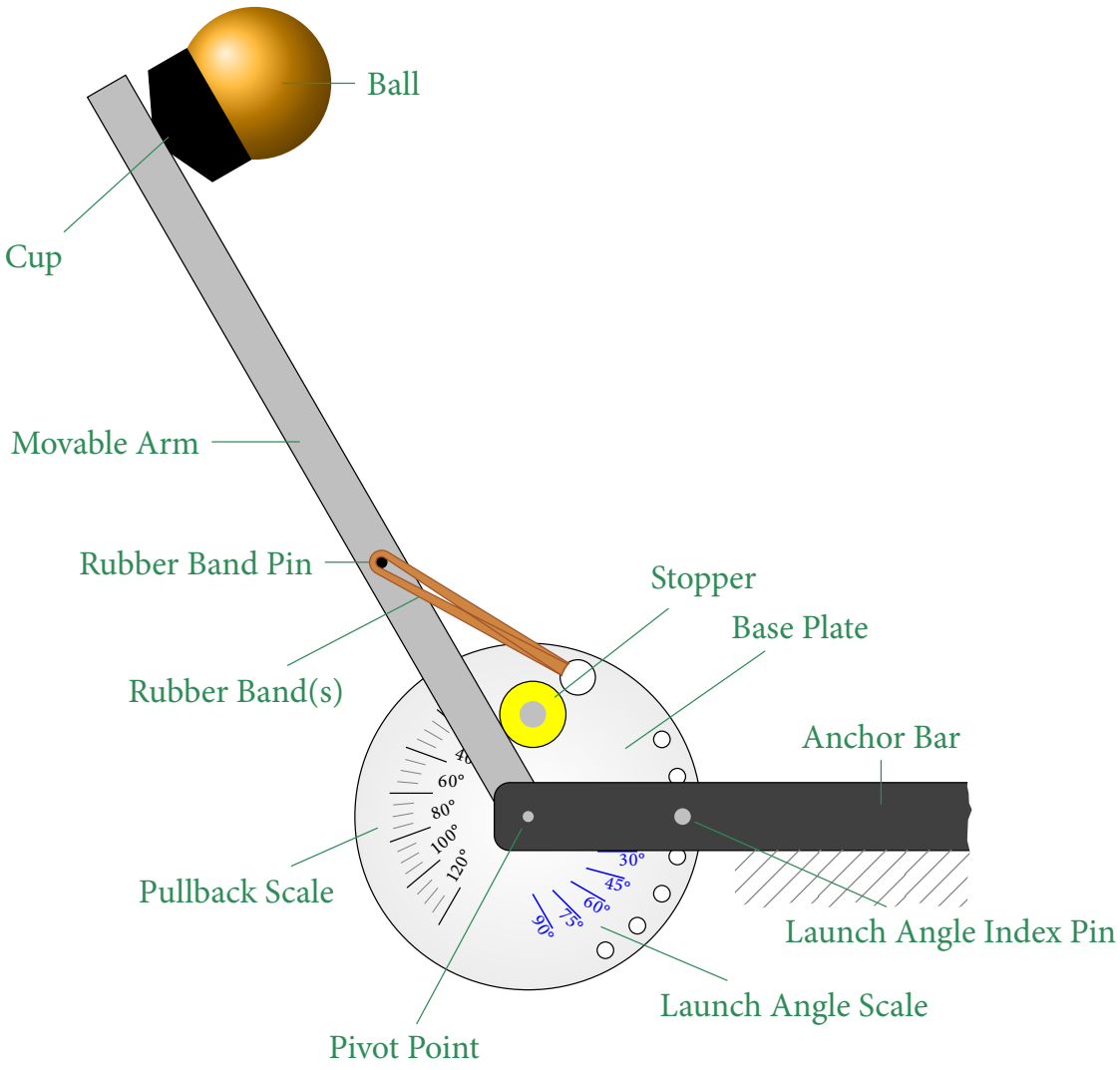
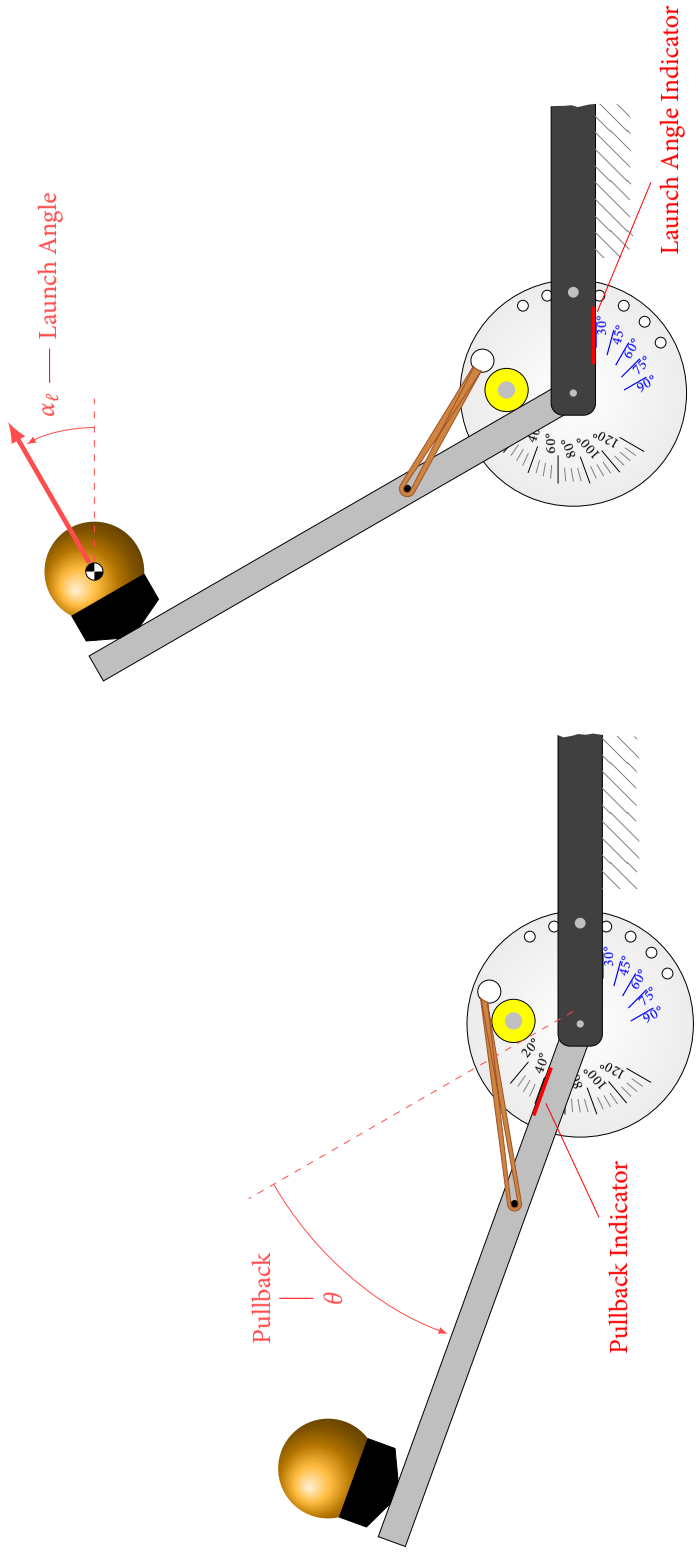


Table 1: Salient Features of the Simulation Model

	from release to launch	from launch to (first) landing
Initial State	released from rest at initial pullback	launch speed, position & direction
Kinetic Energy	arm (rigid body); ball (point mass)	ball (point mass)
Potential Energy	rubber band(s) (ideal spring); gravity	gravity
Losses	none	aerodynamic drag; Magnus lift due to ball spin
Final State	launch speed, position & direction	landing position



(b) Catapult Device in the 30°-Launch, 0°-Pullback Position

(a) Catapult Device in the 30°-Launch, 40°-Pullback Position

Figure 1: Catapult angle definitions. Note that pullback is relative to the stopper and launch angle is relative to the horizon.

Pilot Study First Group
Participant Graphical Aids
(2 pages)

This page intentionally left blank.

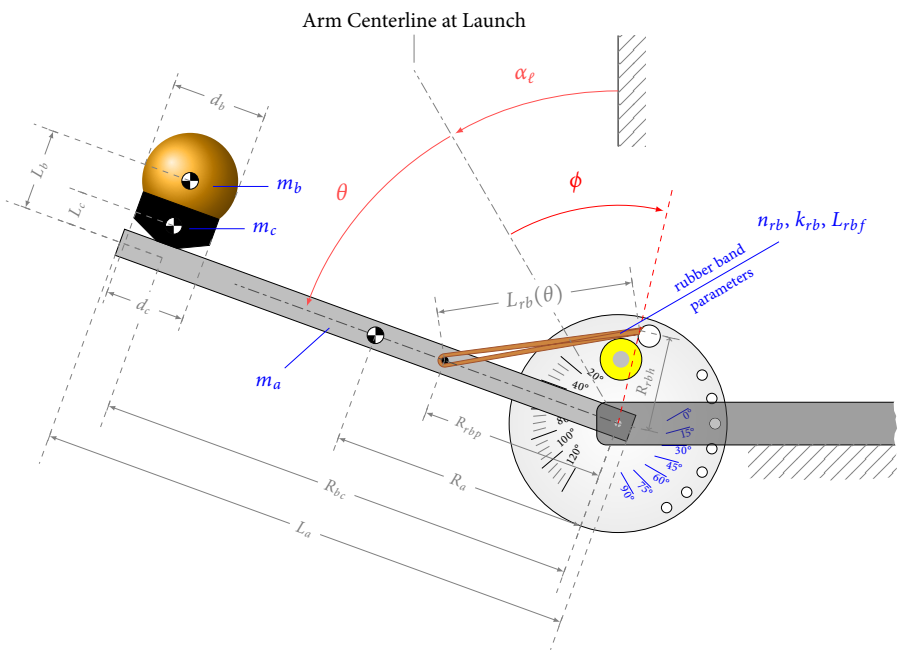


Figure 1: Engineering Schematic

Table 1: Parameter Specifications

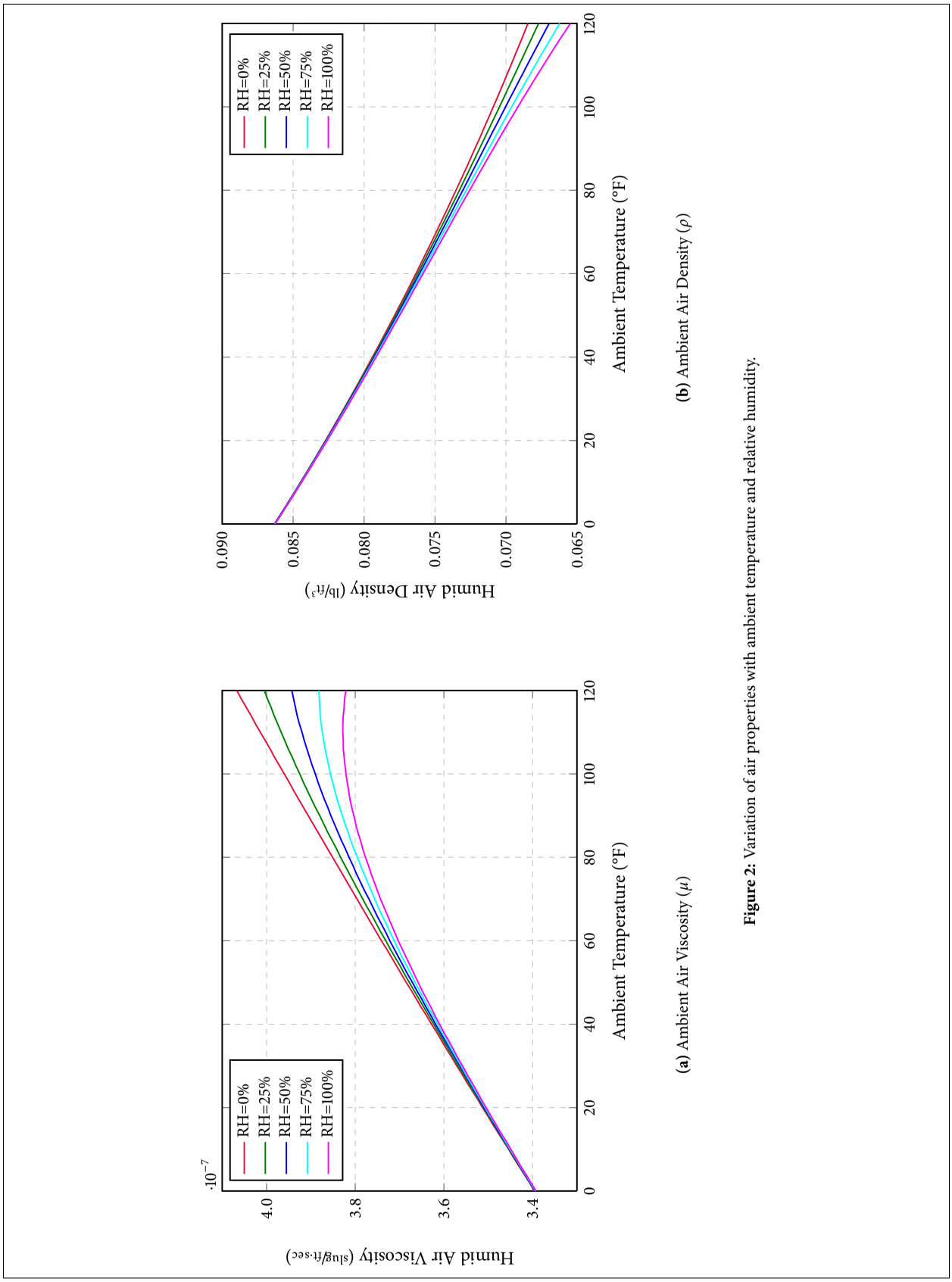
Parameter	Description	Value
α_ℓ	launch angle	varies[a]
θ	pullback	varies[a]
μ	dynamic viscosity of air	function of RH and T
ρ	density of air	function of RH and T
ϕ	rubber band hole offset angle	43°
C_d	aerodynamic drag coefficient	0.5[b]
C_M	Magnus effect lift coefficient	varies[a,b]
d_b	ball diameter	varies[a]
d_c	cup diameter	1.45 in
k_{rb}	rubber band "spring constant"	0.514 $\frac{\text{lbf}}{\text{in}}$
L_a	length of catapult arm	9.81 in
L_b	ball c.g. offset length	varies[a]
L_c	cup c.g. offset length	0.516 in
L_{rb}	length of rubber band(s)	function of θ
L_{rbf}	"free" length of the rubber band	2.93 in[b]
m_a	mass of catapult arm	varies[a]
m_b	mass of ball	varies[a]
m_c	mass of cup	0.275 oz
n_{rb}	number of rubber bands	varies[a]
R_a	radial position of arm c.g.	4.66 in
R_{bc}	radial position of ball and cup	8.75 in
R_{rbp}	radial position of rubber band pin	3.33 in
R_{rbh}	radial position of rubber band hole	1.72 in
RH	ambient relative humidity	varies[a,b]
T	ambient temperature	varies[a,b]

[a]This value is either a control factor or directly dependent upon the value of a control factor.

[b]This value is not shown in the figure at left, but it is used in the simulation of the post launch dynamics.

Table 2: Control Factors

Control Factor	"-1" Level		"+1" Level	
	setting	parameters	setting	parameters
relative humidity	25%	RH = 25%	75%	RH = 75%
initial pullback	30°	θ_o = 30°	40°	θ_o = 40°
type of ball	"large ball" table tennis	d_b = 1.732 in m_b = 0.081 oz L_b = 1.599 in	regulation table tennis	d_b = 1.575 in m_b = 0.094 oz L_b = 1.432 in
arm material	magnesium	m_a = 1.186 oz	aluminum	m_a = 1.840 oz
launch angle	60°	α_ℓ = 60°	45°	α_ℓ = 45°
number of rubber bands	3	n_{rb} = 3	2	n_{rb} = 2
ambient temperature	72°F	T = 72°F	32°F	T = 32°F



(a) Ambient Air Viscosity (μ)

(b) Ambient Air Density (ρ)

Figure 2: Variation of air properties with ambient temperature and relative humidity.

Pilot Study
Instructions & Design Table for Participants using aOFAT
(1 page)

This page intentionally left blank.

Instructions

In the adaptive one factor at a time (aOFAT) method, given the set of control factors and their possible settings, we:

1. Select a starting configuration either at random or based upon *a priori* knowledge of the system.
2. Evaluate the system response at the starting configuration.
3. For each control factor in the system
   (a) Select a new configuration by using the *best* previous configuration and changing only this factor's setting to its alternate value.
   (b) Evaluate the system response at the new configuration.
   (c) If the performance improves at the new setting of this factor, keep it at this setting for the remainder of the experiment; otherwise, keep it at the original value for the remainder of the experiment.
4. The configuration obtained after stepping through each control factor is the optimized result for this design approach.

For the selected system, the Xpult catapult, we have identified seven control factors, each with two level settings of interest (i.e., a 2^7 system). Using the aOFAT algorithm in this case requires eight trials (initial configuration plus one iteration for each control factor).

In this experiment, each trial is an evaluation of the system response (i.e., the ball's landing position) using the computer simulation of the catapult. Use the aOFAT Design Table below with the above algorithm to find an optimal configuration for the catapult if the objective is to hit a target of 96 inches.

aOFAT Design Table

>>> Target Landing Position = 96 inches <<<

Trial	Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position
1	25%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	
2	75%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	
3		40°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	
4	⋮		Regulation Table Tennis	Magnesium	60°	3	72°F	
5	⋮	⋮		Aluminum	60°	3	72°F	
6	⋮	⋮	⋮		45°	3	72°F	
7	⋮	⋮	⋮	⋮		2	72°F	
8	⋮	⋮	⋮	⋮	⋮		32°F	
Best								

This page intentionally left blank.

Pilot Study
Instructions & Design Table for Participants using PB-L_8
(2 pages)

This page intentionally left blank.

Instructions

The objective here is to find the combination of control factor settings that will result in the ball landing as close as possible to the target distance of 96 inches. This is a 2^7 system (seven factors, each of which can be set at two possible values), so the total number of unique combinations of control factors is $2^7 = 128$. Instead of testing the device at every possible combination, we test a small subset using *orthogonal arrays* to sample the design space. Orthogonality results in unique, non-redundant information about the system being learned in each trial. An appropriate set of arrays for a 2^7 system is the Plackett-Burman L_8 matrix, which is shown below with random ordering of columns.

Trial	Control Factor						
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	−1	−1	−1	−1	−1	−1	−1
2	+1	−1	+1	−1	−1	+1	+1
3	+1	+1	+1	−1	+1	−1	−1
4	−1	+1	−1	−1	+1	+1	+1
5	+1	−1	−1	+1	+1	−1	+1
6	−1	−1	+1	+1	+1	+1	−1
7	−1	+1	+1	+1	−1	−1	+1
8	+1	+1	−1	+1	−1	+1	−1

Here the "-1" and "+1" values indicate the "low" and "high" settings at which the control factors can be set. Note that "low" and "high" do not imply relative numerical values of the settings. In fact, each setting for the catapult system has been randomly assigned to a "low" or "high" designation as shown in the table below.

Coded Setting	Relative Humidity (X_1)	Pullback (X_2)	Type of Table Tennis Ball (X_3)	Arm Material (X_4)	Launch Angle (X_5)	Rubber Bands (X_6)	Ambient Temperature (X_7)
−1	25%	30°	Orange	Magnesium	60°	3	72°F
+1	75%	40°	White	Aluminum	45°	2	32°F

Now if we evaluate the system response at the configuration shown in each row of the design matrix, we can create a simple linear approximation of the system response

$$\widehat{Y} = \underbrace{\beta_0}_{\text{AVERAGE}} + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7}_{\text{MAIN EFFECTS}} \tag{1}$$

where $\widehat{Y}$ is the estimate of the landing position Y in inches, and the X_i values are each set at either −1 or +1. Because the orthogonal arrays have the property of being *balanced* (each control factor setting appears in the same number of trials, 4), we may calculate the coefficients in the above equation as

$$\beta_0 = \frac{1}{8}\sum_{j=1}^{8} Y_j \qquad \text{(Grand Mean)} \tag{2}$$

$$\beta_i = \frac{1}{2}\left(\frac{1}{4}\sum_{j\left|X_i=+1\right.} Y_j - \frac{1}{4}\sum_{j\left|X_i=-1\right.} Y_j\right) \qquad \text{(Main Effects)} \tag{3}$$

Note that subscript i refers to control factor number and subscript j refers to trial number. Once the coefficients $\beta_0 - \beta_7$ are known, we can use Equation 1 to estimate the response at the system configurations that were *not tested*, allowing us to identify one with performance close to the design goal.

The task in this experiment is therefore to

1. Evaluate the system response at each of the required configurations in the design table, using the Predict-Observe-Explain format for each trial.
2. Use Equations 2 and 3 to calculate the coefficients of the linear approximation of the system response. Note that a worksheet and basic calculator are provided to help with this step.
3. Use Equation 1, repeated below the design table for convenience, to find a combination of control factors that will result in a system response close to the target value. Again, the calculator is provided for this.
4. When an appropriate system configuration has been identified, record it at the bottom of the design table and evaluate the system response at this configuration to confirm that the actual response is close to the value predicted by the linear model.

PB-L_8 Design Table

>>> Target Landing Position = 96 inches <<<

Trial	Relative Humidity (X_1)	Pullback (X_2)	Type of Ball (X_3)	Arm Material (X_4)	Launch Angle (X_5)	Rubber Bands (X_6)	Ambient Temperature (X_7)	Landing Position (Y)
1	25%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	
2	75%	30°	Regulation Table Tennis	Magnesium	60°	2	32°F	
3	75%	40°	Regulation Table Tennis	Magnesium	45°	3	72°F	
4	25%	40°	Large-Ball Table Tennis	Magnesium	45°	2	32°F	
5	75%	30°	Large-Ball Table Tennis	Aluminum	45°	3	32°F	
6	25%	30°	Regulation Table Tennis	Aluminum	45°	2	72°F	
7	25%	40°	Regulation Table Tennis	Aluminum	60°	3	32°F	
8	75%	40°	Large-Ball Table Tennis	Aluminum	60°	2	72°F	
Best								

$$\widehat{Y} = \underbrace{\square}_{\beta_0} + \underbrace{\square}_{\beta_1} X_1 + \underbrace{\square}_{\beta_2} X_2 + \underbrace{\square}_{\beta_3} X_3 + \underbrace{\square}_{\beta_4} X_4 + \underbrace{\square}_{\beta_5} X_5 + \underbrace{\square}_{\beta_6} X_6 + \underbrace{\square}_{\beta_7} X_7$$

Pilot Study First Group
POE Worksheet for Participants using aOFAT
(4 pages)

This page intentionally left blank.

***Planned* Data Collection** Test Subject ID: 01022

1. Evaluate the system response for the configuration required for Trial 1 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
25%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F

   (a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

   Please explain your reasoning for the above prediction.

   Please circle your level of confidence in the above prediction on a 5-point scale:

   no confidence 1 2 3 4 5 highest confidence

   (b) **Observe.** Use the computer simulation to determine the response for this configuration.

   Landing position: ☐

   (c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

2. Evaluate the system response for the configuration required for Trial 2 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
75%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F

   (a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

   Please explain your reasoning for the above prediction.

   Please circle your level of confidence in the above prediction on a 5-point scale:

   no confidence 1 2 3 4 5 highest confidence

   (b) **Observe.** Use the computer simulation to determine the response for this configuration.

   Landing position: ☐

   (c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

***Planned* Data Collection** Test Subject ID: 01022

3. Evaluate the system response for the configuration required for Trial 3 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
	40°	Large-Ball Table Tennis	Magnesium	60°	3	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position:

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

4. Evaluate the system response for the configuration required for Trial 4 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
		Regulation Table Tennis	Magnesium	60°	3	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position:

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

***Planned* Data Collection** Test Subject ID: 01022

5. Evaluate the system response for the configuration required for Trial 5 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
			Aluminum	60°	3	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position:

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

6. Evaluate the system response for the configuration required for Trial 6 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
				45°	3	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position:

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

***Planned* Data Collection** Test Subject ID: 01022

7. Evaluate the system response for the configuration required for Trial 7 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
					2	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position:

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

8. Evaluate the system response for the configuration required for Trial 8 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
						32°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position:

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

Pilot Study First Group
POE Worksheet for Participants using PB-L_8
(4 pages)

This page intentionally left blank.

***Planned* Data Collection** Test Subject ID: 11636

1. Evaluate the system response for the configuration required for Trial 1 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
25%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F

   (a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

   Please explain your reasoning for the above prediction.

   Please circle your level of confidence in the above prediction on a 5-point scale:

   no confidence 1 2 3 4 5 highest confidence

   (b) **Observe.** Use the computer simulation to determine the response for this configuration.

   Landing position: ____

   (c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

2. Evaluate the system response for the configuration required for Trial 2 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
75%	30°	Regulation Table Tennis	Magnesium	60°	2	32°F

   (a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

   Please explain your reasoning for the above prediction.

   Please circle your level of confidence in the above prediction on a 5-point scale:

   no confidence 1 2 3 4 5 highest confidence

   (b) **Observe.** Use the computer simulation to determine the response for this configuration.

   Landing position: ____

   (c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

***Planned* Data Collection** Test Subject ID: 11636

3. Evaluate the system response for the configuration required for Trial 3 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
75%	40°	Regulation Table Tennis	Magnesium	45°	3	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position: ☐

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

4. Evaluate the system response for the configuration required for Trial 4 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
25%	40°	Large-Ball Table Tennis	Magnesium	45°	2	32°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position: ☐

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

***Planned* Data Collection** Test Subject ID: 11636

5. Evaluate the system response for the configuration required for Trial 5 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
75%	30°	Large-Ball Table Tennis	Aluminum	45°	3	32°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position: ☐

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

6. Evaluate the system response for the configuration required for Trial 6 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
25%	30°	Regulation Table Tennis	Aluminum	45°	2	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position: ☐

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

***Planned* Data Collection** Test Subject ID: 11636

7. Evaluate the system response for the configuration required for Trial 7 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
25%	40°	Regulation Table Tennis	Aluminum	60°	3	32°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position: ☐

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

8. Evaluate the system response for the configuration required for Trial 8 in the design plan:

Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature
75%	40°	Large-Ball Table Tennis	Aluminum	60°	2	72°F

(a) **Predict.** Based on what you have learned about this system so far, what would you predict the response to be for this configuration? Your prediction can be absolute (e.g. the landing position will be 77.5 inches) or relative (e.g. the landing position will be greater than in the configuration for a previous step).

Please explain your reasoning for the above prediction.

Please circle your level of confidence in the above prediction on a 5-point scale:

no confidence 1 2 3 4 5 highest confidence

(b) **Observe.** Use the computer simulation to determine the response for this configuration.

Landing position: ☐

(c) **Explain.** If the response does not agree with your prediction, can you offer an explanation for the discrepancy?

Pilot Study
Tabulated Simulation Results
(3 pages)

This page intentionally left blank.

Table 2: Full Factorial Results for Catapult Simulation

Trial	Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position
44	25%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	75.87 in
12	25%	30°	Large-Ball Table Tennis	Magnesium	60°	3	32°F	73.91 in
42	25%	30°	Large-Ball Table Tennis	Magnesium	60°	2	72°F	52.23 in
10	25%	30°	Large-Ball Table Tennis	Magnesium	60°	2	32°F	51.25 in
36	25%	30°	Large-Ball Table Tennis	Magnesium	45°	3	72°F	87.47 in
4	25%	30°	Large-Ball Table Tennis	Magnesium	45°	3	32°F	85.68 in
34	25%	30°	Large-Ball Table Tennis	Magnesium	45°	2	72°F	60.85 in
2	25%	30°	Large-Ball Table Tennis	Magnesium	45°	2	32°F	59.98 in
48	25%	30°	Large-Ball Table Tennis	Aluminum	60°	3	72°F	93.44 in
16	25%	30°	Large-Ball Table Tennis	Aluminum	60°	3	32°F	90.55 in
46	25%	30°	Large-Ball Table Tennis	Aluminum	60°	2	72°F	67.17 in
14	25%	30°	Large-Ball Table Tennis	Aluminum	60°	2	32°F	65.61 in
40	25%	30°	Large-Ball Table Tennis	Aluminum	45°	3	72°F	108.02 in
8	25%	30°	Large-Ball Table Tennis	Aluminum	45°	3	32°F	105.28 in
38	25%	30°	Large-Ball Table Tennis	Aluminum	45°	2	72°F	77.54 in
6	25%	30°	Large-Ball Table Tennis	Aluminum	45°	2	32°F	76.13 in
43	25%	30°	Regulation Table Tennis	Magnesium	60°	3	72°F	80.29 in
11	25%	30°	Regulation Table Tennis	Magnesium	60°	3	32°F	78.70 in
41	25%	30°	Regulation Table Tennis	Magnesium	60°	2	72°F	53.75 in
9	25%	30°	Regulation Table Tennis	Magnesium	60°	2	32°F	53.01 in
35	25%	30°	Regulation Table Tennis	Magnesium	45°	3	72°F	92.34 in
3	25%	30°	Regulation Table Tennis	Magnesium	45°	3	32°F	90.91 in
33	25%	30°	Regulation Table Tennis	Magnesium	45°	2	72°F	62.77 in
1	25%	30°	Regulation Table Tennis	Magnesium	45°	2	32°F	62.12 in
47	25%	30°	Regulation Table Tennis	Aluminum	60°	3	72°F	100.68 in
15	25%	30°	Regulation Table Tennis	Aluminum	60°	3	32°F	98.22 in
45	25%	30°	Regulation Table Tennis	Aluminum	60°	2	72°F	70.17 in
13	25%	30°	Regulation Table Tennis	Aluminum	60°	2	32°F	68.94 in
39	25%	30°	Regulation Table Tennis	Aluminum	45°	3	72°F	115.78 in
7	25%	30°	Regulation Table Tennis	Aluminum	45°	3	32°F	113.53 in
37	25%	30°	Regulation Table Tennis	Aluminum	45°	2	72°F	80.94 in
5	25%	30°	Regulation Table Tennis	Aluminum	45°	2	32°F	79.85 in
60	25%	40°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	98.27 in
28	25%	40°	Large-Ball Table Tennis	Magnesium	60°	3	32°F	95.09 in
58	25%	40°	Large-Ball Table Tennis	Magnesium	60°	2	72°F	70.54 in
26	25%	40°	Large-Ball Table Tennis	Magnesium	60°	2	32°F	68.84 in
52	25%	40°	Large-Ball Table Tennis	Magnesium	45°	3	72°F	113.80 in
20	25%	40°	Large-Ball Table Tennis	Magnesium	45°	3	32°F	110.77 in
50	25%	40°	Large-Ball Table Tennis	Magnesium	45°	2	72°F	81.41 in
18	25%	40°	Large-Ball Table Tennis	Magnesium	45°	2	32°F	79.86 in
64	25%	40°	Large-Ball Table Tennis	Aluminum	60°	3	72°F	117.78 in
32	25%	40°	Large-Ball Table Tennis	Aluminum	60°	3	32°F	113.32 in
62	25%	40°	Large-Ball Table Tennis	Aluminum	60°	2	72°F	88.24 in

Table 2: Full Factorial Results for Catapult Simulation (continued)

Trial	Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position
30	25%	40°	Large-Ball Table Tennis	Aluminum	60°	2	32°F	85.64 in
56	25%	40°	Large-Ball Table Tennis	Aluminum	45°	3	72°F	137.60 in
24	25%	40°	Large-Ball Table Tennis	Aluminum	45°	3	32°F	133.20 in
54	25%	40°	Large-Ball Table Tennis	Aluminum	45°	2	72°F	101.90 in
22	25%	40°	Large-Ball Table Tennis	Aluminum	45°	2	32°F	99.47 in
59	25%	40°	Regulation Table Tennis	Magnesium	60°	3	72°F	106.70 in
27	25%	40°	Regulation Table Tennis	Magnesium	60°	3	32°F	103.95 in
57	25%	40°	Regulation Table Tennis	Magnesium	60°	2	72°F	74.17 in
25	25%	40°	Regulation Table Tennis	Magnesium	60°	2	32°F	72.80 in
51	25%	40°	Regulation Table Tennis	Magnesium	45°	3	72°F	122.86 in
19	25%	40°	Regulation Table Tennis	Magnesium	45°	3	32°F	120.32 in
49	25%	40°	Regulation Table Tennis	Magnesium	45°	2	72°F	85.47 in
17	25%	40°	Regulation Table Tennis	Magnesium	45°	2	32°F	84.25 in
63	25%	40°	Regulation Table Tennis	Aluminum	60°	3	72°F	130.50 in
31	25%	40°	Regulation Table Tennis	Aluminum	60°	3	32°F	126.47 in
61	25%	40°	Regulation Table Tennis	Aluminum	60°	2	72°F	94.47 in
29	25%	40°	Regulation Table Tennis	Aluminum	60°	2	32°F	92.30 in
55	25%	40°	Regulation Table Tennis	Aluminum	45°	3	72°F	151.29 in
23	25%	40°	Regulation Table Tennis	Aluminum	45°	3	32°F	147.45 in
53	25%	40°	Regulation Table Tennis	Aluminum	45°	2	72°F	108.62 in
21	25%	40°	Regulation Table Tennis	Aluminum	45°	2	32°F	106.64 in
108	75%	30°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	75.97 in
76	75%	30°	Large-Ball Table Tennis	Magnesium	60°	3	32°F	73.94 in
106	75%	30°	Large-Ball Table Tennis	Magnesium	60°	2	72°F	52.28 in
74	75%	30°	Large-Ball Table Tennis	Magnesium	60°	2	32°F	51.27 in
100	75%	30°	Large-Ball Table Tennis	Magnesium	45°	3	72°F	87.57 in
68	75%	30°	Large-Ball Table Tennis	Magnesium	45°	3	32°F	85.70 in
98	75%	30°	Large-Ball Table Tennis	Magnesium	45°	2	72°F	60.89 in
66	75%	30°	Large-Ball Table Tennis	Magnesium	45°	2	32°F	59.99 in
112	75%	30°	Large-Ball Table Tennis	Aluminum	60°	3	72°F	93.60 in
80	75%	30°	Large-Ball Table Tennis	Aluminum	60°	3	32°F	90.59 in
110	75%	30°	Large-Ball Table Tennis	Aluminum	60°	2	72°F	67.25 in
78	75%	30°	Large-Ball Table Tennis	Aluminum	60°	2	32°F	65.63 in
104	75%	30°	Large-Ball Table Tennis	Aluminum	45°	3	72°F	108.16 in
72	75%	30°	Large-Ball Table Tennis	Aluminum	45°	3	32°F	105.32 in
102	75%	30°	Large-Ball Table Tennis	Aluminum	45°	2	72°F	77.62 in
70	75%	30°	Large-Ball Table Tennis	Aluminum	45°	2	32°F	76.15 in
107	75%	30°	Regulation Table Tennis	Magnesium	60°	3	72°F	80.38 in
75	75%	30°	Regulation Table Tennis	Magnesium	60°	3	32°F	78.72 in
105	75%	30°	Regulation Table Tennis	Magnesium	60°	2	72°F	53.79 in
73	75%	30°	Regulation Table Tennis	Magnesium	60°	2	32°F	53.02 in
99	75%	30°	Regulation Table Tennis	Magnesium	45°	3	72°F	92.41 in
67	75%	30°	Regulation Table Tennis	Magnesium	45°	3	32°F	90.93 in

Table 2: Full Factorial Results for Catapult Simulation (continued)

Trial	Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position
97	75%	30°	Regulation Table Tennis	Magnesium	45°	2	72°F	62.81 in
65	75%	30°	Regulation Table Tennis	Magnesium	45°	2	32°F	62.12 in
111	75%	30°	Regulation Table Tennis	Aluminum	60°	3	72°F	100.81 in
79	75%	30°	Regulation Table Tennis	Aluminum	60°	3	32°F	98.25 in
109	75%	30°	Regulation Table Tennis	Aluminum	60°	2	72°F	70.24 in
77	75%	30°	Regulation Table Tennis	Aluminum	60°	2	32°F	68.96 in
103	75%	30°	Regulation Table Tennis	Aluminum	45°	3	72°F	115.89 in
71	75%	30°	Regulation Table Tennis	Aluminum	45°	3	32°F	113.55 in
101	75%	30°	Regulation Table Tennis	Aluminum	45°	2	72°F	81.00 in
69	75%	30°	Regulation Table Tennis	Aluminum	45°	2	32°F	79.86 in
124	75%	40°	Large-Ball Table Tennis	Magnesium	60°	3	72°F	98.44 in
92	75%	40°	Large-Ball Table Tennis	Magnesium	60°	3	32°F	95.13 in
122	75%	40°	Large-Ball Table Tennis	Magnesium	60°	2	72°F	70.64 in
90	75%	40°	Large-Ball Table Tennis	Magnesium	60°	2	32°F	68.86 in
116	75%	40°	Large-Ball Table Tennis	Magnesium	45°	3	72°F	113.96 in
84	75%	40°	Large-Ball Table Tennis	Magnesium	45°	3	32°F	110.81 in
114	75%	40°	Large-Ball Table Tennis	Magnesium	45°	2	72°F	81.49 in
82	75%	40°	Large-Ball Table Tennis	Magnesium	45°	2	32°F	79.88 in
128	75%	40°	Large-Ball Table Tennis	Aluminum	60°	3	72°F	118.03 in
96	75%	40°	Large-Ball Table Tennis	Aluminum	60°	3	32°F	113.37 in
126	75%	40°	Large-Ball Table Tennis	Aluminum	60°	2	72°F	88.38 in
94	75%	40°	Large-Ball Table Tennis	Aluminum	60°	2	32°F	85.68 in
120	75%	40°	Large-Ball Table Tennis	Aluminum	45°	3	72°F	137.83 in
88	75%	40°	Large-Ball Table Tennis	Aluminum	45°	3	32°F	133.25 in
118	75%	40°	Large-Ball Table Tennis	Aluminum	45°	2	72°F	102.03 in
86	75%	40°	Large-Ball Table Tennis	Aluminum	45°	2	32°F	99.50 in
123	75%	40°	Regulation Table Tennis	Magnesium	60°	3	72°F	106.84 in
91	75%	40°	Regulation Table Tennis	Magnesium	60°	3	32°F	103.99 in
121	75%	40°	Regulation Table Tennis	Magnesium	60°	2	72°F	74.25 in
89	75%	40°	Regulation Table Tennis	Magnesium	60°	2	32°F	72.82 in
115	75%	40°	Regulation Table Tennis	Magnesium	45°	3	72°F	122.99 in
83	75%	40°	Regulation Table Tennis	Magnesium	45°	3	32°F	120.35 in
113	75%	40°	Regulation Table Tennis	Magnesium	45°	2	72°F	85.54 in
81	75%	40°	Regulation Table Tennis	Magnesium	45°	2	32°F	84.27 in
127	75%	40°	Regulation Table Tennis	Aluminum	60°	3	72°F	130.72 in
95	75%	40°	Regulation Table Tennis	Aluminum	60°	3	32°F	126.53 in
125	75%	40°	Regulation Table Tennis	Aluminum	60°	2	72°F	94.59 in
93	75%	40°	Regulation Table Tennis	Aluminum	60°	2	32°F	92.32 in
119	75%	40°	Regulation Table Tennis	Aluminum	45°	3	72°F	151.49 in
87	75%	40°	Regulation Table Tennis	Aluminum	45°	3	32°F	147.49 in
117	75%	40°	Regulation Table Tennis	Aluminum	45°	2	72°F	108.72 in
85	75%	40°	Regulation Table Tennis	Aluminum	45°	2	32°F	106.66 in

This page intentionally left blank.

Pilot Study Second Group
Consent Form
(4 pages)

This page intentionally left blank.

CONSENT TO PARTICIPATE IN
NON-BIOMEDICAL RESEARCH

THE EFFECT OF EXPERIMENTAL DESIGN METHOD ON ENGINEERING JUDGMENT IN USING COMPUTER SIMULATIONS FOR DESIGN, Part II

You are asked to participate in a research study conducted by Dan Frey and Troy Savoie, from the Mechanical Engineering Department at the Massachusetts Institute of Technology (M.I.T.). The results of this study will contribute to research papers and a doctoral thesis. You were selected as a possible participant in this study because you are either an engineer or an engineering student with the appropriate technical background. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

• PARTICIPATION AND WITHDRAWAL

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

• PURPOSE OF THE STUDY

The purpose of this study is to understand certain aspects of using computer simulations with structured design methods in optimizing a physical system. Further details about the nature of the study will be revealed at the end of your participation.

• PROCEDURES

If you volunteer to participate in this study, we would ask you to do the following things:

I. Introduction (10 min)
   a) Listen to a description of the experiment format.
   b) Read and sign this consent form.
   c) Provide information about your educational and work background (area of technical expertise, degree(s) earned, years of full-time working experience) that may assist us in analyzing the results.

II. Training (10 min)

a) Listen to a description of the physical device, its configuration options and the performance goal(s) to be achieved.

b) Listen to a description of the design space sampling method to be used.

III. Design Problem (30 min) - For each experiment in the design method:

a) Study the system configuration,

b) Predict the performance in this configuration,

c) Explain the reasoning supporting the prediction,

d) Disclose your level of confidence in the prediction,

e) Observe the outcome of the simulation results,

f) Discuss possible reasons for any differences between your prediction and the simulation results.

IV. Conclusion (10 min)

a) Provide feedback regarding your experience in this study.

b) Discuss any questions you may have about the study.

c) Learn further details about the nature of the study.

Your participation in this study should take about one hour to complete.

• POTENTIAL RISKS AND DISCOMFORTS

None.

• POTENTIAL BENEFITS

It is hoped you will find your participation in this experiment interesting and educational. You will follow a specified method of experimental design and interact with a computer simulation of a physical system, therefore you may learn about these tools commonly used by contemporary design engineers.

It is also hoped that this and subsequent experiments will help organizations (companies, governments, etc.) to make more effective use of simulations in making design decisions.

• PAYMENT FOR PARTICIPATION

None.

• CONFIDENTIALITY

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

Both video and audio of your participation will be recorded using a digital video camera and a digital audio recorder. These recordings will begin after this consent form has been signed, and will continue through the end of your participation in the study. You have the right to view the video recording and listen to the audio recording and to edit parts or all of them at your discretion. These recordings, together with any writings by you during your participation, constitute the raw data for your participation in the experiment and will remain confidential.

The data from this study will be associated with a subject number which helps us to link the experimental results with your survey information (education, experience, etc). The records will be archived in a filing cabinet behind two locked doors for five years and then will be destroyed. The data will be used in research papers and a doctoral thesis, but only in the aggregate after statistical analysis.

• IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact Daniel D. Frey who is the Principal Investigator. He can be reached at room 3-449D, (617) 324-6133, and danfrey@mit.edu.

• EMERGENCY CARE AND COMPENSATION FOR INJURY

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to the M.I.T. Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

• RIGHTS OF RESEARCH SUBJECTS

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone (617) 253-6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

_____
Name of Subject

_____ _____
Signature of Subject or Legal Representative Date

_____
Email address or telephone number
(in case we need to contact you later regarding your participation in this study)

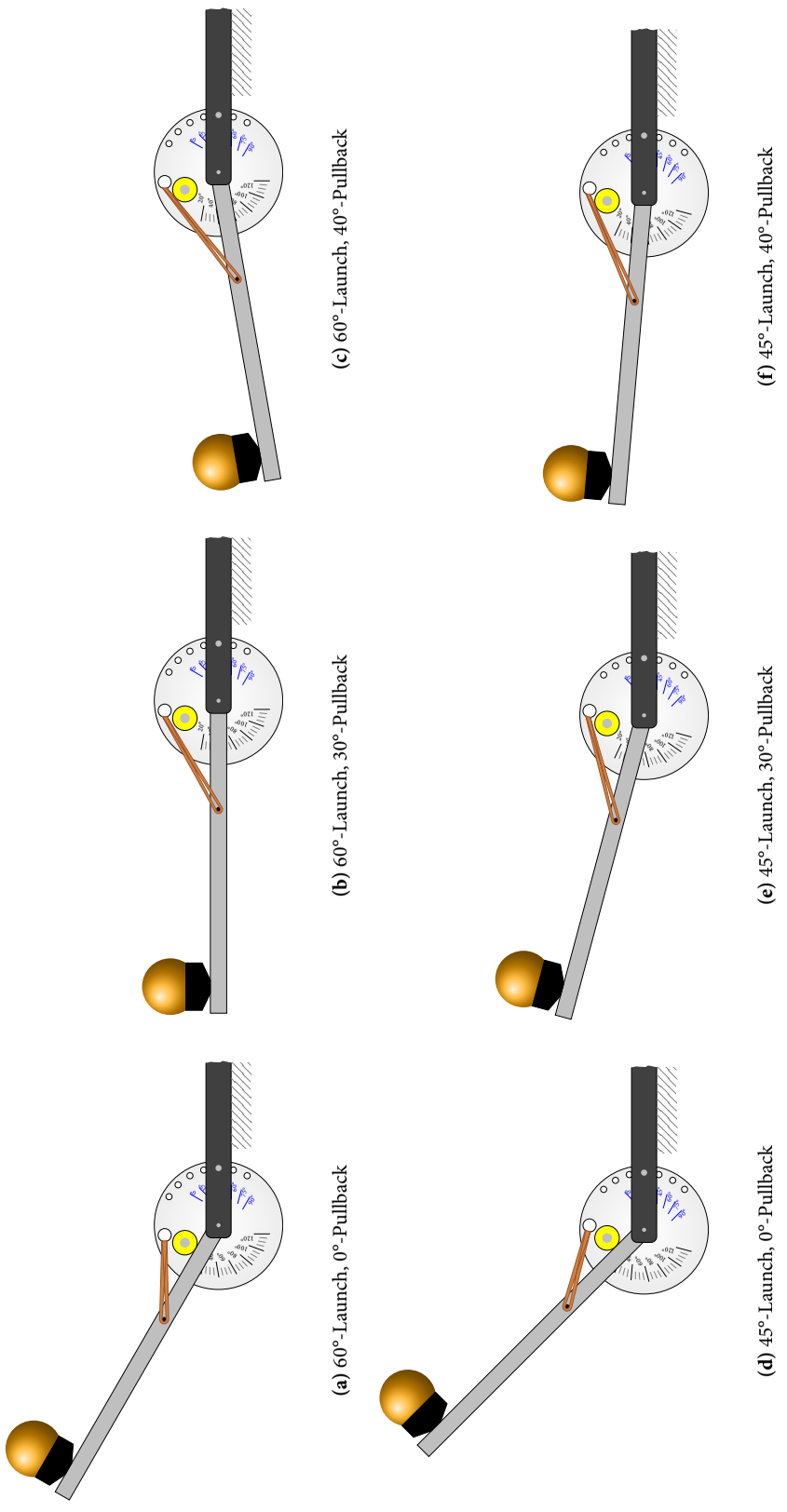
SIGNATURE OF INVESTIGATOR

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

_____ _____
Signature of Investigator Date

Pilot Study Second Group
Participant Graphical Aids
(2 pages)

This page intentionally left blank.



(a) 60°-Launch, 0°-Pullback

(b) 60°-Launch, 30°-Pullback

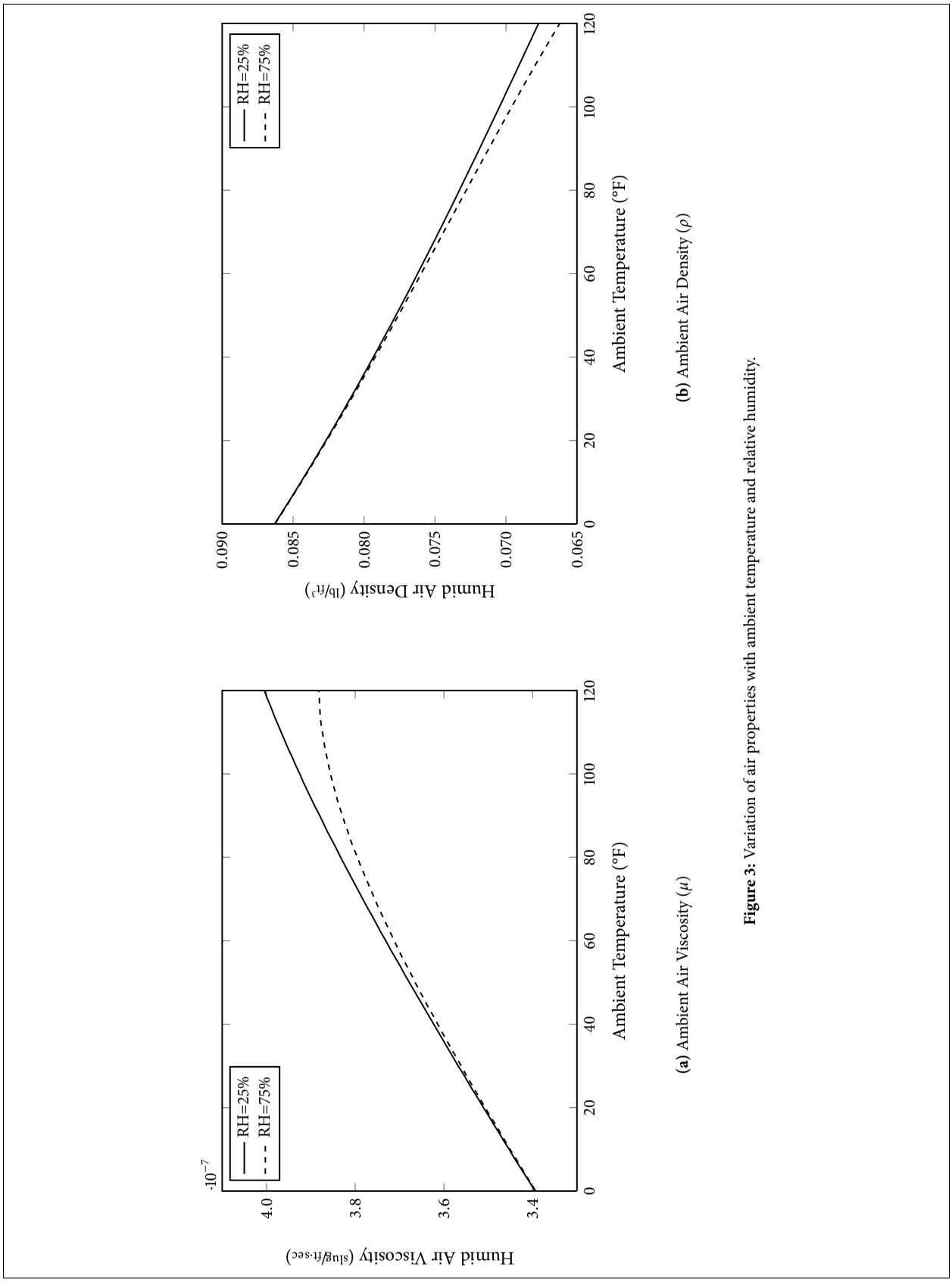
(c) 60°-Launch, 40°-Pullback

(d) 45°-Launch, 0°-Pullback

(e) 45°-Launch, 30°-Pullback

(f) 45°-Launch, 40°-Pullback

Figure 2: Catapult in various positions of interest (SCALE = 1:5).



(a) Ambient Air Viscosity (μ)

(b) Ambient Air Density (ρ)

Figure 3: Variation of air properties with ambient temperature and relative humidity.

Pilot Study Second Group
Worksheet for Participants using PB-L_8
(2 pages)

This page intentionally left blank.

PB-L_8 Worksheet Test Subject ID: 11703

Overall Average Response.

$$\beta_0 = \underbrace{\frac{1}{8}\left(\square + \square + \square + \square + \square + \square + \square + \square\right)}_{\text{MEAN VALUE OVER ALL TRIALS}}$$

$$\beta_0 = \square$$

1. Relative Humidity

$$\beta_1 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (75\%)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (25\%)}}\right)$$

$$= \frac{1}{2}\left(\square - \square\right)$$

$$\beta_1 = \square$$

2. Pullback

$$\beta_2 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (40°)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (30°)}}\right)$$

$$= \frac{1}{2}\left(\square - \square\right)$$

$$\beta_2 = \square$$

3. Type of Ball

$$\beta_3 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (WHITE)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (ORANGE)}}\right)$$

$$= \frac{1}{2}\left(\square - \square\right)$$

$$\beta_3 = \square$$

PB-L_8 Worksheet Test Subject ID: 11703

4. Arm Material

$$\beta_4 = \frac{1}{2}\left[\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (ALUMINUM)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (MAGNESIUM)}}\right]$$

$$= \frac{1}{2}\left[\square - \square\right]$$

$$\beta_4 = \square$$

5. Launch Angle

$$\beta_5 = \frac{1}{2}\left[\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (45°)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (60°)}}\right]$$

$$= \frac{1}{2}\left[\square - \square\right]$$

$$\beta_5 = \square$$

6. Number of Rubber Bands

$$\beta_6 = \frac{1}{2}\left[\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (2)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (3)}}\right]$$

$$= \frac{1}{2}\left[\square - \square\right]$$

$$\beta_6 = \square$$

7. Ambient Temperature

$$\beta_7 = \frac{1}{2}\left[\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (32°F)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (72°F)}}\right]$$

$$= \frac{1}{2}\left[\square - \square\right]$$

$$\beta_7 = \square$$

Main Study
Administrator Script
(7 pages)

This page intentionally left blank.

Research Study: The Effect of Experimental Design Method on Engineering Judgment in Using Computer Simulations for Design

Troy Savoie

Spring 2009

1 Introduction

1.1 Listen to a description of the experiment format

Thank you for volunteering to participate in this experiment, in which you'll be performing a short design task for a simple configurable physical system, to determine a configuration that will satisfy the design objective.

For consistency between test subjects, I will be reading from a script for much of the experiment. Please feel free to interrupt me to ask questions if something is not clear.

After the necessary paperwork has been completed, I'll describe the configurable system, the design objective and the design method, then we'll get started with the task.

1.2 Read and sign the consent form

This is the consent form. It broadly describes the purpose and format of the experiment, and your rights and protections as a test subject. It is a requirement of the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) that each test subject read and sign the consent form before proceeding.

(Wait for test subject to complete consent form, and then sign and date the consent form.)

1.3 Provide information about your educational and work background (area of technical expertise, degree(s) earned, years of full-time working experience) that may assist us in analyzing the results.

I'm collecting demographic data that could potentially help to explain trends in the experimental results. Again this is only connected to the randomly generated ID number. Please provide as much detail as you feel comfortable doing. *(Wait for test subject to complete demographic data form.)*

1.4 (Turn on the audio and video recorders.)

I am now turning on both the audio and video recorders. *(Pause to turn on video, manually focus it and check framing, then turn on audio.)*

1.5 (Read opening statement for record keeping on recordings.)

This is test subject XXXXX, the date is XXX XX, 2009, and the time is XX:XX AM/PM. Test subject consent has been granted for participation in the experiment, including recording audio and video beginning now. As a condition for me to conduct this experiment here, the Security Department requires that I remind you to not discuss anything that is either classified or export controlled during this session.

2 Training

2.1 Listen to a description of the physical device, its configuration options and the performance goal(s) to be achieved.

The simple physical system that is the focus of the design task is this catapult device.

(Give the catapult overview sketch to the test subject, to reference while speaking about it.)

The important parts of the catapult are as follows.

1. The anchor is a metal bar that is clamped to ground; for example, a table.
2. The base plate is a metal plate that is pinned to the anchor in two places but can be repositioned through rotation about its center by removing ...
3. the launch angle index pin, which goes through the anchor bar and one of seven holes in the base plate spaced 15°apart.
4. The movable arm is a metal bar that freely rotates about the same pivot as the base plate,
5. The stopper is a thick plastic washer that is bolted to the baseplate to stop the movable arm,
6. One or more standard office-supply-store rubber bands are threaded through a hole in the base plate and connected to the movable arm by looping each end around ...
7. the rubber band pin, which protrudes about a half inch on either side of the arm.
8. The cup is a plastic, hollow half-cylinder bolted to the movable arm. Its inner diameter is slightly smaller than the diameter of ...
9. the ball, which is a table tennis ball, also known as a "ping pong" ball.

At rest, the catapult arm is held against the stopper by the rubber bands. In order to launch the ball:

1. First, select the launch angle by pinning the base plate through one of the seven holes. The launch angle is defined counterclockwise with respect to the horizon as shown.
2. Next, the catapult arm is grasped just behind the cup and pulled back to the desired initial pullback. This angle is also measured counterclockwise, but it is relative to the stopper, whereas the launch angle is relative to ground. Remember that the stopper will be repositioned along with the base plate if a different launch angle is set.
3. When the arm is pulled back to the desired position, the ball is placed on the rim of the cup.
4. The arm is released from rest, and the stretched rubber band(s) pull the arm back against the stopper.
5. The arm is prevented from rotating further, but the ball leaves the cup.
6. After a short flight, the ball lands at some distance from the catapult.
7. The landing distance, measured from the catapult pivot, is the metric that we use to quantify the performance.

The idea is that we have this metric, landing distance, and we want to hit a certain target. Here we are not focused on maximizing performance, we just want to hit the target. The target in this case is 96 inches from the pivot. We have a number of configuration options for the system, and the design task is to find one of those configuration options that will get us close to the target, ideally without having to sample the entire design space. The design space is defined as all permutations of the system with the given control factors and their possible settings.

We can control the settings of five configuration options for the catapult and two for the ambient environment in which it operates. For the catapult we can change the type of ball, the number of rubber bands, the type of material used for the catapult arm, the launch angle, and the initial pullback. For the ambient environment we can change the air temperature and relative humidity. These seven configurable parameters are called the control factors.

To assess your familiarity with the physics involved, I'll go through each of the control factors one at a time, asking you to give a rough estimate of what you think will happen to the landing distance of the ball when you change the control factor from its nominal setting to its alternate setting.

(Point to control factor table.)

This table gives the control factor names in the first column, nominal setting for the control factors in the second column and alternate setting for the control factors in the third column. Blue labels are used here and in the relevant parts of the engineering sketch for quick reference. For this first task I'd like you to consider the case where all of the control factors are at the nominal setting and we only consider one control factor - what happens when you change that control factor from its nominal to its alternate setting? What effect do you think that will have on the landing distance

of the ball? Here I am looking for qualitative answers: If you don't think it will have much effect, "negligible effect" is a valid answer. Similarly, if you think the ball will go farther, whether you think it will go a little farther or a lot farther would be a good thing to know. Also, please provide the rationale behind your judgment and a level of confidence on a scale of 1 to 5, with 5 being most confident.

Before we proceed, let me point out a few things in the engineering sketch. The drawing is dimensionally accurate to one-half scale; if you want to know any specific lengths in the drawing that are not given, use the provided ruler to measure the length and multiply the result by two to get the value. There are two different ping-pong balls here: an orange one and a white one. Not only are they different colors, they also have different properties. The diameter of the white ball is about 10 percent smaller than the diameter of the orange ball, but the mass of the white ball is about 15 percent larger than the mass of the orange ball. The specific values of these properties are printed on each of the balls for reference. Note that the mass of the cup is significant as well. The mass of the arm for the two different material options is written on the arm for each of the options. This is the total mass of the arm; geometry remains exactly the same - only the material used differs. In your consideration of the arm, you may assume it behaves as a rigid body and that its elasticity does not influence the outcome. For the rubber bands, it's easiest to consider them as ideal springs. This *(point to text written next to spring)* would be the equivalent spring model for a single rubber band: there is a stiffness and a free length. This should be all that you need to know in order to make judgments about the effects of the control factors on the system response. Do you have any questions before getting started? When you are ready, you may begin your predictions for each of the control factors.

(Proceed through predictions for all seven control factors.)

2.1.1 The Adaptive One Factor at a Time (aOFAT) Approach

Next I will introduce the specific design algorithm to be used, which in this case is called Adaptive One Factor at a Time (aOFAT). This algorithm is a very simple approach, and one that is often used by designers by default without specifically calling it this. The approach is to start with some configuration of the system, which could be the designer's guess at the optimum or a completely random configuration. The system response is then obtained for this configuration; this is Trial 1. Next, for each control factor

1. Start with the configuration in the set of previous trials that results in the best performance.
2. Change the setting of the control factor under consideration to its alternate value.
3. Evaluate the system response for this configuration.

When all control factors have been evaluated in this manner, the optimum result for this algorithm is the configuration specifically tested that results in the best performance. Note that this algorithm requires $n + 1$ trials, where n is the number of two-level control factors in the system.

2.1.2 The Plackett-Burman L8 (PBL8) Approach

Now I will talk about the design method to be used. By design method, I mean a specific algorithm for sampling the design space that will allow you to build a simple model of what is happening in the design space. This will allow you to estimate what is happening at the points that you don't explicitly sample. Specifically, we'll sample eight configurations in the design space. When I say design space here, in design of experiments this system would be called a 2^7 system, because there are seven control factors, each of which can be set at one of two levels. If there were three levels, it would be a 3^7 system, or if there were 8 factors it would be a 2^8 system. This is just nomenclature used in design of experiments. This nomenclature is also practical in that it tells us the number of permutations possible in the system. In a 2^7 system, there are 128 possible configurations. We could evaluate the system over all 128 configurations to determine the best one - this would be the brute force approach. Using a design of experiments approach allows us to build a simple model of the system by strategic sampling of the design space. In this case, we will evaluate the system response at 8 points, and the resulting model will allow us to estimate the system response at the 120 other points in the design space. The system response in this case is the landing position of the ball as calculated by a computer simulation.

The model used in the simulation has the assumptions given in the top box of the information sheet. There are two distinct parts of the simulation. From the release of the arm until it hits the stopper we model the dynamics of the entire catapult device. When the catapult arm hits the stopper, the ball is launched and the dynamics of interest shifts to the ball alone in free flight.

During the prelaunch simulation, the assumption for the initial state is that the catapult arm is released from rest at the initial pullback. The kinetic energy modeled is the arm as a rigid body and the ball and cup as point masses. The potential energy modeled are the rubber bands as ideal springs and also gravity. The model assumes no losses in this part of the simulation. The final state is the launch speed, position and direction of the ball.

During the postlaunch simulation, the initial state is simply equal to the final state of the prelaunch simulation. The kinetic energy modeled is the ball as a point mass in free flight. The potential energy modeled is due to gravity. Losses here are due to aerodynamic drag and the Magnus lift from ball spin. The final state is the landing position of the ball. The landing position of the ball is the system response needed to evaluate performance.

I have developed a simulation of this system based on the described assumptions and using the parameters shown in the sketch. Instead of bringing a computer into the room and having you run the simulation, which would increase the amount of time needed to complete this task, I have run the simulation over all 128 possible cases for this system. The results are in a lookup table, which I will use to give you the required result as needed to complete the design task. I would like for you to consider this data as if you were generating it yourself by running this unfamiliar simulation for the first time.

The design method to be used is called fractional factorial. This is a generic term that simply means we are not evaluating the full factorial which would be all 128 points in this case. Here we

are going to evaluate the system at 8 of the 128 points, so the design is a 1/16 fractional factorial. To determine which points to evaluate, we will use a set of orthogonal arrays that is very popular in Taguchi methods, the Plackett-Burman L8 design matrix. The design matrix is shown as a table here *(point to matrix)*, where control factors are mapped to columns, table entries specify the control factor setting ("-1" is nominal and "+1" is alternate), and each row corresponds to one evaluation of the system response (or trial in DoE nomenclature). For example, in trial 1 all control factors will be set at the nominal setting since all of the entries are "-1" in the table. This design matrix is essentially a recipe for efficiently gathering data regarding the average system response over a broad range of control factor settings.

In order to use this information to estimate the system response at other points in the design space, we need to choose a mathematical model for the estimate. Here we will use a simple linear equation in which we only consider the main effects in the system. This equation is shown here *(point to equation)*, where the X's are control factor settings and Y-hat is the estimate of the system response Y. In our application, the X's can be either "-1" or "+1"; we are not working in units associated with each control factor, only the nondimensional terms that correspond to nominal or alternate settings. Y and Y-hat are in the units of the system response, which here is inches and corresponds to landing distance of the ball with respect to the catapult pivot point. Beta-naught is a constant with the same units as Y-hat. It corresponds to the overall average response of the system. Ideally we would calculate this value by averaging over all 128 values in the full factorial. In our fractional factorial method, we are counting on the properties of balance and orthogonality in the design matrix to allow us to average over our 8 trial cases and still get a reasonable estimate of the overall average response. The coefficients beta-one through beta-seven are called the main effect sizes and are in units of inches as well. The physical meaning of these coefficients is that each represents half of the estimated change in the system response if the corresponding control factor is changed from its nominal to its alternate configuration. Why half? Because changing X from "-1" to "+1" introduces a factor of 2 into the response estimate. The main effect beta coefficients are calculated as follows: for each factor, its beta value is equal to one-half times the average response at the alternate (+1) setting minus the average response at the nominal (-1) setting.

Technically speaking, we could calculate eight beta coefficients exactly using any eight points in the design space. However, we are using the points specified by the Plackett-Burman L8 design matrix because of two nice properties: orthogonality and balance. First, each row in the design matrix is orthogonal to the others. This ensures that there is no overlap in information gathering in the experiment, as there would be if for example one row was a linear combination of two or more of the other rows. Second, the matrix is balanced with respect to the control factor levels: each control factor is at the nominal setting in four rows and the alternate setting in four rows. The benefit of this is a reduction in bias in calculating the main effect sizes (beta-one through beta-seven). If a control factor appeared in more trials at one setting than at the other, there would be a bias toward this setting in the calculated coefficient.

Once the eight beta coefficients are calculated, the linear response estimate may be used to explore the system response at all combinations of control factor settings. Now the entire design task is clear: To find a configuration of the catapult, from the specified set of control factors and their

settings, that will result in the ball hitting close to the target landing position, we evaluate the system response at the 8 catapult configurations specified by the design matrix, using the computer simulation to calculate the landing distance in each case. Next we calculate the coefficients in the linear equation for estimating the landing position. Finally, we use this linear equation to explore the full 128-point design space to find a configuration close to the target landing position. Performance of the selected configuration might then be verified using the computer simulation, since it is possible that the linear estimate of the system response is not a good estimate.

3 Design Problem

This page intentionally left blank.

Main Study
Consent Form
(4 pages)

This page intentionally left blank.

CONSENT TO PARTICIPATE IN
NON-BIOMEDICAL RESEARCH

THE EFFECT OF EXPERIMENTAL DESIGN METHOD ON ENGINEERING JUDGMENT IN USING COMPUTER SIMULATIONS FOR DESIGN, Part II

You are asked to participate in a research study conducted by Dan Frey and Troy Savoie, from the Mechanical Engineering Department at the Massachusetts Institute of Technology (M.I.T.). The results of this study will contribute to research papers and a doctoral thesis. You were selected as a possible participant in this study because you are either an engineer or an engineering student with the appropriate technical background. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

• PARTICIPATION AND WITHDRAWAL

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

• PURPOSE OF THE STUDY

The purpose of this study is to understand certain aspects of using computer simulations with structured design methods in optimizing a physical system. Further details about the nature of the study will be revealed at the end of your participation.

• PROCEDURES

If you volunteer to participate in this study, we would ask you to do the following things:

I. Introduction (10 min)
   a) Listen to a description of the experiment format.
   b) Read and sign this consent form.
   c) Provide information about your educational and work background (area of technical expertise, degree(s) earned, years of full-time working experience) that may assist us in analyzing the results.

II. Training (10 min)
   a) Listen to a description of the physical device, its configuration options and the performance goal(s) to be achieved.

b) Answer questions to assess your comprehension of the device's operation.

c) Listen to a description of the design space sampling method to be used.

d) Answer questions to assess your comprehension of the design space sampling method.

III. Design Problem (30 min) - For each experiment in the design method:

a) Study the system configuration,

b) Predict the performance in this configuration,

c) Explain the reasoning supporting the prediction,

d) Disclose your level of confidence in the prediction,

e) Observe the outcome of the simulation results,

f) Discuss possible reasons for any differences between your prediction and the simulation results.

IV. Conclusion (10 min)

a) Provide feedback regarding your experience in this study.

b) Discuss any questions you may have about the study.

c) Learn further details about the nature of the study.

Your participation in this study should take about one hour to complete.

• POTENTIAL RISKS AND DISCOMFORTS

None.

• POTENTIAL BENEFITS

It is hoped you will find your participation in this experiment interesting and educational. You will follow a specified method of experimental design and interact with a computer simulation of a physical system, therefore you may learn about these tools commonly used by contemporary design engineers.

It is also hoped that this and subsequent experiments will help organizations (companies, governments, etc.) to make more effective use of simulations in making design decisions.

• PAYMENT FOR PARTICIPATION

None.

• CONFIDENTIALITY

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

Both video and audio of your participation will be recorded using a digital video camera and a digital audio recorder. These recordings will begin after this consent form has been signed, and will continue through the end of your participation in the study. You have the right to view the video recording and listen to the audio recording and to edit parts or all of them at your discretion. These recordings, together with any writing by you during your participation, constitute the raw data for your participation in the experiment and will remain confidential.

Third-party access to the data is as follows. The Director of Environmental Health and Safety at Draper Laboratory will have access to the data for the purpose of auditing to ensure that this experiment complies with applicable safety regulations. The professional videographer in Draper Laboratory Media Services will have access to the video recording for the purpose of editing and then transferring onto a DVD-ROM disc. In addition, two paid assistants trained in protocol analysis will have access to this DVD-ROM disc and a written transcript of the audio recording for the purpose of independently analyzing the results. The data will be used in research papers and a doctoral thesis, but only in the aggregate after statistical analysis.

When not in use as described above, the records will be locked in a desk drawer at Draper Laboratory while the analysis and its documentation are underway. At the conclusion of the study, the records will be archived in a locked cabinet in the Principal Investigator's office at M.I.T. for a period of five years and then will be destroyed.

• IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact Daniel D. Frey who is the Principal Investigator. He can be reached at room 3-449D, (617) 324-6133, and danfrey@mit.edu.

• EMERGENCY CARE AND COMPENSATION FOR INJURY

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to the M.I.T. Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

• RIGHTS OF RESEARCH SUBJECTS

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone (617) 253-6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

_____
Name of Subject

_____ _____
Signature of Subject or Legal Representative Date

_____
Email address or telephone number
(in case we need to contact you later regarding your participation in this study)

SIGNATURE OF INVESTIGATOR

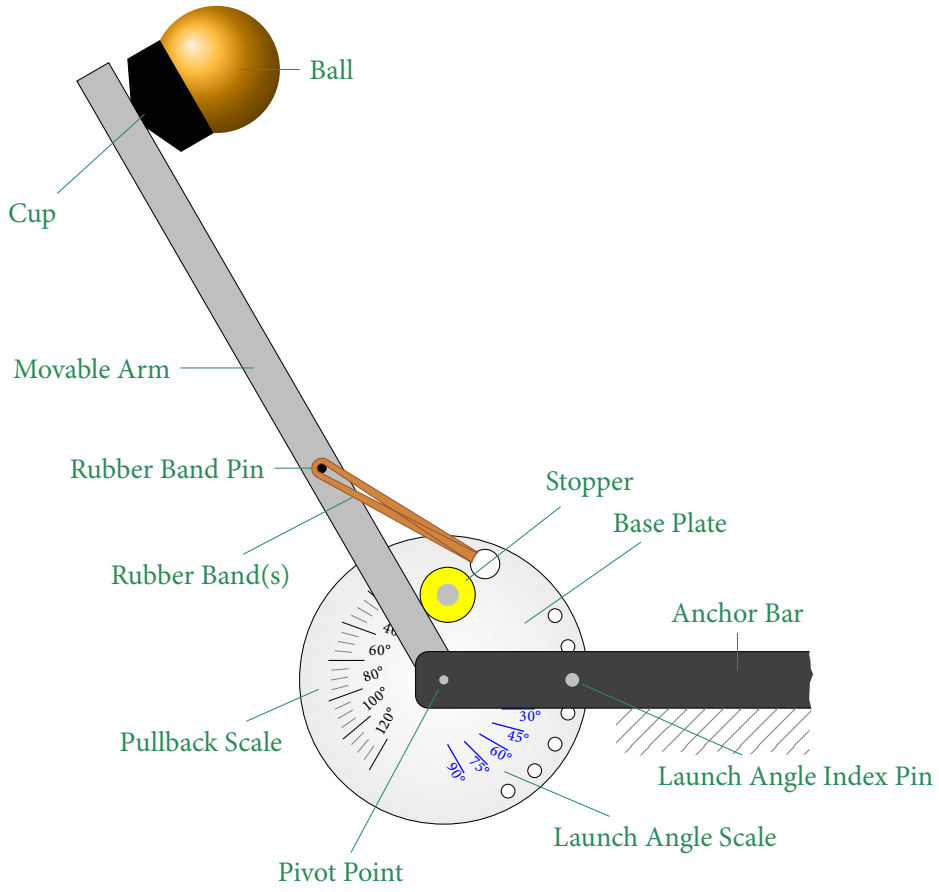
In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

_____ _____
Signature of Investigator Date

Main Study
Participant Graphical Aids
(3 pages)

This page intentionally left blank.

Components of the Catapult Device

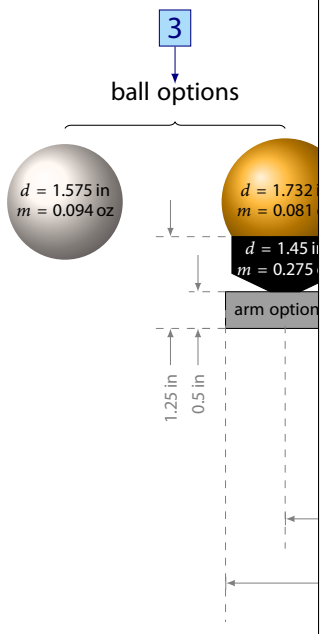


Salient Features of the Computer Simulation Model

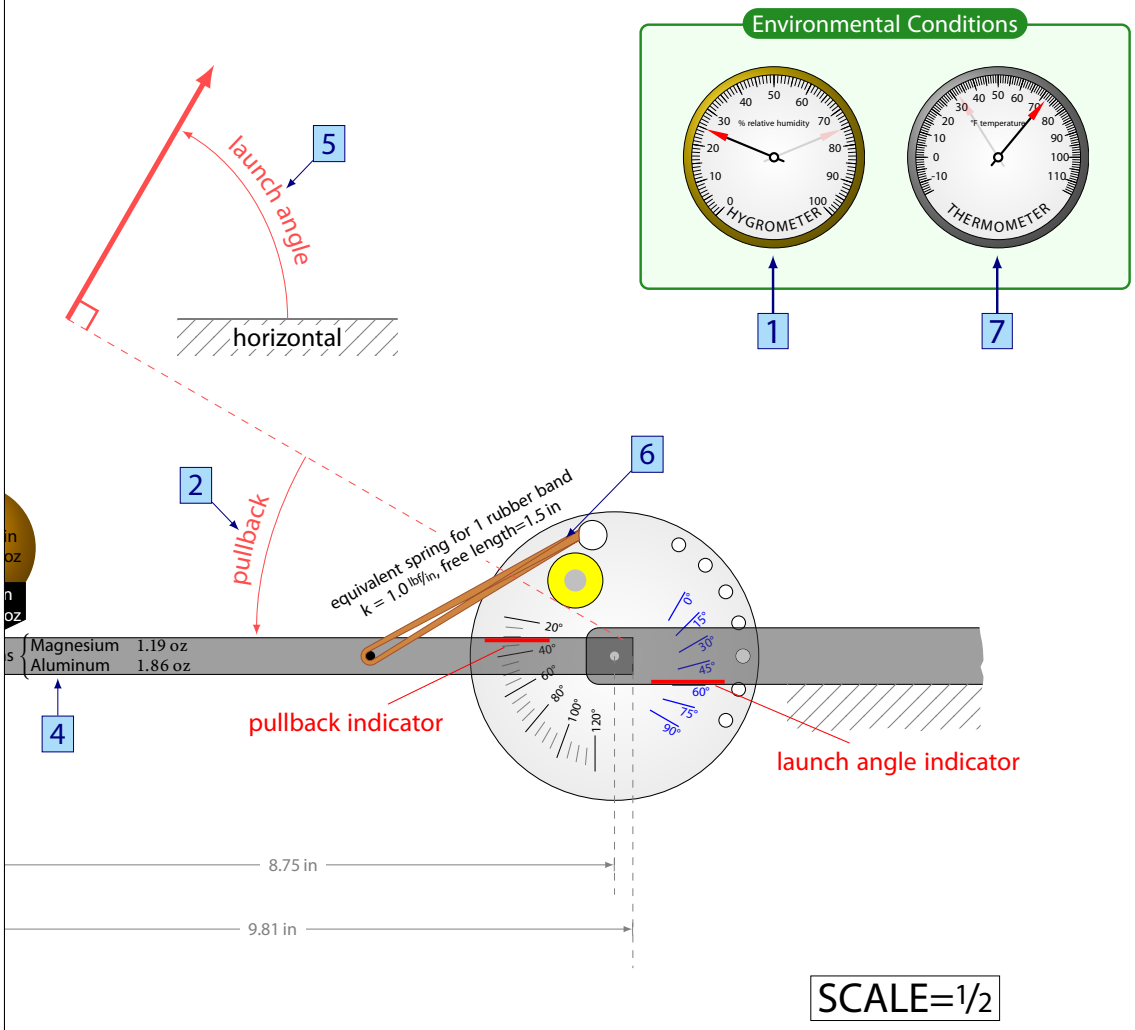
	Initial State	Kinetic Energy
Part 1: Release to Launch	released from rest at initial pullback	arm (rigid body); ball & cup (point masses)
Part 2: Launch to Landing	launch speed, position & direction from Part 1	ball (point mass)

System Shown in Nominal Configuration with Factors Labeled

Control Factor	Nominal Setting	Alternate Setting
1 Relative Humidity	25%	75%
2 Pullback	30°	40°
3 Type of Ball	Orange Table Tennis	White Table Tennis
4 Arm Material	Magnesium	Aluminum
5 Launch Angle	60°	45°
6 Rubber Bands	3	2
7 Ambient Temperature	72°F	32°F



Potential Energy	Losses	Final State
rubber band(s) (ideal spring); gravity	none	launch speed, position & direction
gravity	aerodynamic drag; Magnus lift due to ball spin	landing position



This page intentionally left blank.

Main Study
Instructions & Design Table for Participants using aOFAT
(1 page)

This page intentionally left blank.

Instructions

In the adaptive one factor at a time (aOFAT) method, given the set of control factors and their possible settings, we:

1. Select a starting configuration either at random or based upon *a priori* knowledge of the system.
2. Evaluate the system response at the starting configuration.
3. For each control factor in the system
   (a) Select a new configuration by using the *best* previous configuration and changing only this factor's setting to its alternate value.
   (b) Evaluate the system response at the new configuration.
   (c) If the performance improves at the new setting of this factor, keep it at this setting for the remainder of the experiment; otherwise, keep it at the original value for the remainder of the experiment.
4. The configuration obtained after stepping through each control factor is the optimized result for this design approach.

For the selected system, the Xpult catapult, we have identified seven control factors, each with two level settings of interest (i.e., a 2^7 system). Using the aOFAT algorithm in this case requires eight trials (initial configuration plus one iteration for each control factor).

In this experiment, each trial is an evaluation of the system response (i.e., the ball's landing position) using the computer simulation of the catapult. Use the aOFAT Design Table below with the above algorithm to find an optimal configuration for the catapult if the objective is to hit a target of 96 inches.

aOFAT Design Table

>>> Target Landing Position = 96 inches <<<

Trial	Relative Humidity	Pullback	Type of Table Tennis Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position (predicted)	Landing Position (simulation)
1	25%	30°	Orange	Magnesium	60°	3	72°F		75.9 in
2	75%	30°	Orange	Magnesium	60°	3	72°F		
3		40°	Orange	Magnesium	60°	3	72°F		
4	⋮		White	Magnesium	60°	3	72°F		
5	⋮	⋮		Aluminum	60°	3	72°F		
6	⋮	⋮	⋮		45°	3	72°F		
7	⋮	⋮	⋮	⋮		2	72°F		
8	⋮	⋮	⋮	⋮	⋮		32°F		
Best									

This page intentionally left blank.

Main Study
Instructions & Design Table for Participants using PB-L_8
(2 pages)

This page intentionally left blank.

Instructions

The objective here is to find the combination of control factor settings that will result in the ball landing as close as possible to the target distance of 96 inches. This is a 2^7 system (seven factors, each of which can be set at two possible values), so the total number of unique combinations of control factors is $2^7 = 128$. Instead of testing the device at every possible combination, we test a small subset using *orthogonal arrays* to sample the design space. Orthogonality results in unique, non-redundant information about the system being learned in each trial. An appropriate set of arrays for a 2^7 system is the Plackett-Burman L_8 matrix, which is shown below with random ordering of columns.

Trial	Control Factor						
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	−1	−1	−1	−1	−1	−1	−1
2	+1	−1	+1	−1	−1	+1	+1
3	+1	+1	+1	−1	+1	−1	−1
4	−1	+1	−1	−1	+1	+1	+1
5	+1	−1	−1	+1	+1	−1	+1
6	−1	−1	+1	+1	+1	+1	−1
7	−1	+1	+1	+1	−1	−1	+1
8	+1	+1	−1	+1	−1	+1	−1

Here the "-1" and "+1" values indicate the "low" and "high" settings at which the control factors can be set. Note that "low" and "high" do not imply relative numerical values of the settings. In fact, each setting for the catapult system has been randomly assigned to a "low" or "high" designation as shown in the table below.

Coded Setting	Relative Humidity (X_1)	Pullback (X_2)	Type of Table Tennis Ball (X_3)	Arm Material (X_4)	Launch Angle (X_5)	Rubber Bands (X_6)	Ambient Temperature (X_7)
−1	25%	30°	Orange	Magnesium	60°	3	72°F
+1	75%	40°	White	Aluminum	45°	2	32°F

Now if we evaluate the system response at the configuration shown in each row of the design matrix, we can create a simple linear approximation of the system response

$$\widehat{Y} = \underbrace{\beta_0}_{\text{AVERAGE}} + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7}_{\text{MAIN EFFECTS}} \tag{1}$$

where $\widehat{Y}$ is the estimate of the landing position Y in inches, and the X_i values are each set at either −1 or +1. Because the orthogonal arrays have the property of being *balanced* (each control factor setting appears in the same number of trials, 4), we may calculate the coefficients in the above equation as

$$\beta_0 = \frac{1}{8}\sum_{j=1}^{8} Y_j \qquad \text{(Grand Mean)} \tag{2}$$

$$\beta_i = \frac{1}{2}\left(\frac{1}{4}\sum_{j\left|X_i=+1\right.} Y_j - \frac{1}{4}\sum_{j\left|X_i=-1\right.} Y_j\right) \qquad \text{(Main Effects)} \tag{3}$$

Note that subscript i refers to control factor number and subscript j refers to trial number. Once the coefficients $\beta_0 - \beta_7$ are known, we can use Equation 1 to estimate the response at the system configurations that were *not tested*, allowing us to identify one with performance close to the design goal.

The task in this experiment is therefore to

1. Evaluate the system response at each of the required configurations in the design table, using the Predict-Observe-Explain format for each trial.
2. Use Equations 2 and 3 to calculate the coefficients of the linear approximation of the system response. Note that a worksheet and basic calculator are provided to help with this step.
3. Use Equation 1, repeated below the design table for convenience, to find a combination of control factors that will result in a system response close to the target value. Again, the calculator is provided for this.
4. When an appropriate system configuration has been identified, record it at the bottom of the design table and evaluate the system response at this configuration to confirm that the actual response is close to the value predicted by the linear model.

PB-L_8 Design Table

>>> Target Landing Position = 96 inches <<<

Trial	Relative Humidity (X_1)	Pullback (X_2)	Type of Ball (X_3)	Arm Material (X_4)	Launch Angle (X_5)	Rubber Bands (X_6)	Ambient Temperature (X_7)	Landing Position (prediction)	Landing Position (simulation) (Y)
1	$^{-1}$25%	$^{-1}$30°	$^{-1}$Orange	$^{-1}$Magnesium	$^{-1}$60°	$^{-1}$3	$^{-1}$72°F		75.9 in
2	$^{+1}$75%	$^{-1}$30°	$^{+1}$ White	$^{-1}$Magnesium	$^{-1}$60°	$^{+1}$2	$^{+1}$32°F		
3	$^{+1}$75%	$^{+1}$40°	$^{+1}$ White	$^{-1}$Magnesium	$^{+1}$45°	$^{-1}$3	$^{-1}$72°F		
4	$^{-1}$25%	$^{+1}$40°	$^{-1}$Orange	$^{-1}$Magnesium	$^{+1}$45°	$^{+1}$2	$^{+1}$32°F		
5	$^{+1}$75%	$^{-1}$30°	$^{-1}$Orange	$^{+1}$ Aluminum	$^{+1}$45°	$^{-1}$3	$^{+1}$32°F		
6	$^{-1}$25%	$^{-1}$30°	$^{+1}$ White	$^{+1}$ Aluminum	$^{+1}$45°	$^{+1}$2	$^{-1}$72°F		
7	$^{-1}$25%	$^{+1}$40°	$^{+1}$ White	$^{+1}$ Aluminum	$^{-1}$60°	$^{-1}$3	$^{+1}$32°F		
8	$^{+1}$75%	$^{+1}$40°	$^{-1}$Orange	$^{+1}$ Aluminum	$^{-1}$60°	$^{+1}$2	$^{-1}$72°F		
Best									

$$\widehat{Y} = \underbrace{\square}_{\beta_0} + \underbrace{\square}_{\beta_1} X_1 + \underbrace{\square}_{\beta_2} X_2 + \underbrace{\square}_{\beta_3} X_3 + \underbrace{\square}_{\beta_4} X_4 + \underbrace{\square}_{\beta_5} X_5 + \underbrace{\square}_{\beta_6} X_6 + \underbrace{\square}_{\beta_7} X_7$$

Main Study
Tabulated Simulation Results
(3 pages)

This page intentionally left blank.

Table 1: Full Factorial Results for Catapult Simulation

Trial	Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position	
44	25%	30°	Orange	Magnesium	60°	3	72°F	75.87 in	BOTH-1
12	25%	30°	Orange	Magnesium	60°	3	32°F	73.91 in	
42	25%	30°	Orange	Magnesium	60°	2	72°F	52.23 in	
10	25%	30°	Orange	Magnesium	60°	2	32°F	51.25 in	
36	25%	30°	Orange	Magnesium	45°	3	72°F	87.47 in	
4	25%	30°	Orange	Magnesium	45°	3	32°F	85.68 in	
34	25%	30°	Orange	Magnesium	45°	2	72°F	60.85 in	
2	25%	30°	Orange	Magnesium	45°	2	32°F	59.98 in	
48	25%	30°	Orange	Aluminum	60°	3	72°F	93.44 in	
16	25%	30°	Orange	Aluminum	60°	3	32°F	90.55 in	
46	25%	30°	Orange	Aluminum	60°	2	72°F	67.17 in	
14	25%	30°	Orange	Aluminum	60°	2	32°F	65.61 in	
40	25%	30°	Orange	Aluminum	45°	3	72°F	108.02 in	
8	25%	30°	Orange	Aluminum	45°	3	32°F	105.28 in	
38	25%	30°	Orange	Aluminum	45°	2	72°F	77.54 in	
6	25%	30°	Orange	Aluminum	45°	2	32°F	76.13 in	
43	25%	30°	White	Magnesium	60°	3	72°F	80.29 in	
11	25%	30°	White	Magnesium	60°	3	32°F	78.70 in	
41	25%	30°	White	Magnesium	60°	2	72°F	53.75 in	
9	25%	30°	White	Magnesium	60°	2	32°F	53.01 in	
35	25%	30°	White	Magnesium	45°	3	72°F	92.34 in	
3	25%	30°	White	Magnesium	45°	3	32°F	90.91 in	
33	25%	30°	White	Magnesium	45°	2	72°F	62.77 in	
1	25%	30°	White	Magnesium	45°	2	32°F	62.12 in	
47	25%	30°	White	Aluminum	60°	3	72°F	100.68 in	
15	25%	30°	White	Aluminum	60°	3	32°F	98.22 in	
45	25%	30°	White	Aluminum	60°	2	72°F	70.17 in	
13	25%	30°	White	Aluminum	60°	2	32°F	68.94 in	
39	25%	30°	White	Aluminum	45°	3	72°F	115.78 in	
7	25%	30°	White	Aluminum	45°	3	32°F	113.53 in	
37	25%	30°	White	Aluminum	45°	2	72°F	80.94 in	PBL8-6
5	25%	30°	White	Aluminum	45°	2	32°F	79.85 in	
60	25%	40°	Orange	Magnesium	60°	3	72°F	98.27 in	
28	25%	40°	Orange	Magnesium	60°	3	32°F	95.09 in	
58	25%	40°	Orange	Magnesium	60°	2	72°F	70.54 in	
26	25%	40°	Orange	Magnesium	60°	2	32°F	68.84 in	
52	25%	40°	Orange	Magnesium	45°	3	72°F	113.80 in	
20	25%	40°	Orange	Magnesium	45°	3	32°F	110.77 in	
50	25%	40°	Orange	Magnesium	45°	2	72°F	81.41 in	
18	25%	40°	Orange	Magnesium	45°	2	32°F	79.86 in	PBL8-4
64	25%	40°	Orange	Aluminum	60°	3	72°F	117.78 in	
32	25%	40°	Orange	Aluminum	60°	3	32°F	113.32 in	
62	25%	40°	Orange	Aluminum	60°	2	72°F	88.24 in	

Table 1: Full Factorial Results for Catapult Simulation (continued)

Trial	Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position	
30	25%	40°	Orange	Aluminum	60°	2	32°F	85.64 in	
56	25%	40°	Orange	Aluminum	45°	3	72°F	137.60 in	
24	25%	40°	Orange	Aluminum	45°	3	32°F	133.20 in	
54	25%	40°	Orange	Aluminum	45°	2	72°F	101.90 in	
22	25%	40°	Orange	Aluminum	45°	2	32°F	99.47 in	
59	25%	40°	White	Magnesium	60°	3	72°F	106.70 in	
27	25%	40°	White	Magnesium	60°	3	32°F	103.95 in	
57	25%	40°	White	Magnesium	60°	2	72°F	74.17 in	
25	25%	40°	White	Magnesium	60°	2	32°F	72.80 in	
51	25%	40°	White	Magnesium	45°	3	72°F	122.86 in	
19	25%	40°	White	Magnesium	45°	3	32°F	120.32 in	
49	25%	40°	White	Magnesium	45°	2	72°F	85.47 in	
17	25%	40°	White	Magnesium	45°	2	32°F	84.25 in	
63	25%	40°	White	Aluminum	60°	3	72°F	130.50 in	
31	25%	40°	White	Aluminum	60°	3	32°F	126.47 in	PBL8-7
61	25%	40°	White	Aluminum	60°	2	72°F	94.47 in	
29	25%	40°	White	Aluminum	60°	2	32°F	92.30 in	
55	25%	40°	White	Aluminum	45°	3	72°F	151.29 in	
23	25%	40°	White	Aluminum	45°	3	32°F	147.45 in	
53	25%	40°	White	Aluminum	45°	2	72°F	108.62 in	
21	25%	40°	White	Aluminum	45°	2	32°F	106.64 in	
108	75%	30°	Orange	Magnesium	60°	3	72°F	75.97 in	OFAT-2
76	75%	30°	Orange	Magnesium	60°	3	32°F	73.94 in	
106	75%	30°	Orange	Magnesium	60°	2	72°F	52.28 in	
74	75%	30°	Orange	Magnesium	60°	2	32°F	51.27 in	
100	75%	30°	Orange	Magnesium	45°	3	72°F	87.57 in	
68	75%	30°	Orange	Magnesium	45°	3	32°F	85.70 in	
98	75%	30°	Orange	Magnesium	45°	2	72°F	60.89 in	
66	75%	30°	Orange	Magnesium	45°	2	32°F	59.99 in	
112	75%	30°	Orange	Aluminum	60°	3	72°F	93.60 in	
80	75%	30°	Orange	Aluminum	60°	3	32°F	90.59 in	
110	75%	30°	Orange	Aluminum	60°	2	72°F	67.25 in	
78	75%	30°	Orange	Aluminum	60°	2	32°F	65.63 in	
104	75%	30°	Orange	Aluminum	45°	3	72°F	108.16 in	
72	75%	30°	Orange	Aluminum	45°	3	32°F	105.32 in	PBL8-5
102	75%	30°	Orange	Aluminum	45°	2	72°F	77.62 in	
70	75%	30°	Orange	Aluminum	45°	2	32°F	76.15 in	
107	75%	30°	White	Magnesium	60°	3	72°F	80.38 in	
75	75%	30°	White	Magnesium	60°	3	32°F	78.72 in	
105	75%	30°	White	Magnesium	60°	2	72°F	53.79 in	
73	75%	30°	White	Magnesium	60°	2	32°F	53.02 in	PBL8-2
99	75%	30°	White	Magnesium	45°	3	72°F	92.41 in	
67	75%	30°	White	Magnesium	45°	3	32°F	90.93 in	

Table 1: Full Factorial Results for Catapult Simulation (continued)

Trial	Relative Humidity	Pullback	Type of Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Landing Position	
97	75%	30°	White	Magnesium	45°	2	72°F	62.81 in	
65	75%	30°	White	Magnesium	45°	2	32°F	62.12 in	
111	75%	30°	White	Aluminum	60°	3	72°F	100.81 in	
79	75%	30°	White	Aluminum	60°	3	32°F	98.25 in	
109	75%	30°	White	Aluminum	60°	2	72°F	70.24 in	
77	75%	30°	White	Aluminum	60°	2	32°F	68.96 in	
103	75%	30°	White	Aluminum	45°	3	72°F	115.89 in	
71	75%	30°	White	Aluminum	45°	3	32°F	113.55 in	
101	75%	30°	White	Aluminum	45°	2	72°F	81.00 in	
69	75%	30°	White	Aluminum	45°	2	32°F	79.86 in	
124	75%	40°	Orange	Magnesium	60°	3	72°F	98.44 in	OFAT-3
92	75%	40°	Orange	Magnesium	60°	3	32°F	95.13 in	OFAT-8
122	75%	40°	Orange	Magnesium	60°	2	72°F	70.64 in	OFAT-7
90	75%	40°	Orange	Magnesium	60°	2	32°F	68.86 in	
116	75%	40°	Orange	Magnesium	45°	3	72°F	113.96 in	OFAT-6
84	75%	40°	Orange	Magnesium	45°	3	32°F	110.81 in	
114	75%	40°	Orange	Magnesium	45°	2	72°F	81.49 in	
82	75%	40°	Orange	Magnesium	45°	2	32°F	79.88 in	
128	75%	40°	Orange	Aluminum	60°	3	72°F	118.03 in	OFAT-5
96	75%	40°	Orange	Aluminum	60°	3	32°F	113.37 in	
126	75%	40°	Orange	Aluminum	60°	2	72°F	88.38 in	PBL8-8
94	75%	40°	Orange	Aluminum	60°	2	32°F	85.68 in	
120	75%	40°	Orange	Aluminum	45°	3	72°F	137.83 in	
88	75%	40°	Orange	Aluminum	45°	3	32°F	133.25 in	
118	75%	40°	Orange	Aluminum	45°	2	72°F	102.03 in	
86	75%	40°	Orange	Aluminum	45°	2	32°F	99.50 in	
123	75%	40°	White	Magnesium	60°	3	72°F	106.84 in	OFAT-4
91	75%	40°	White	Magnesium	60°	3	32°F	103.99 in	
121	75%	40°	White	Magnesium	60°	2	72°F	74.25 in	
89	75%	40°	White	Magnesium	60°	2	32°F	72.82 in	
115	75%	40°	White	Magnesium	45°	3	72°F	122.99 in	PBL8-3
83	75%	40°	White	Magnesium	45°	3	32°F	120.35 in	
113	75%	40°	White	Magnesium	45°	2	72°F	85.54 in	
81	75%	40°	White	Magnesium	45°	2	32°F	84.27 in	
127	75%	40°	White	Aluminum	60°	3	72°F	130.72 in	
95	75%	40°	White	Aluminum	60°	3	32°F	126.53 in	
125	75%	40°	White	Aluminum	60°	2	72°F	94.59 in	
93	75%	40°	White	Aluminum	60°	2	32°F	92.32 in	
119	75%	40°	White	Aluminum	45°	3	72°F	151.49 in	
87	75%	40°	White	Aluminum	45°	3	32°F	147.49 in	
117	75%	40°	White	Aluminum	45°	2	72°F	108.72 in	
85	75%	40°	White	Aluminum	45°	2	32°F	106.66 in	

This page intentionally left blank.

Main Study
Worksheet for Participants using PB-L_8
(2 pages)

This page intentionally left blank.

PB-L_8 Worksheet Test Subject ID: 11900

Overall Average Response.

$$\beta_0 = \underbrace{\frac{1}{8}\left(\square + \square + \square + \square + \square + \square + \square + \square\right)}_{\text{MEAN VALUE OVER ALL TRIALS}}$$

$$\beta_0 = \square$$

1. Relative Humidity

$$\beta_1 = \frac{1}{2}\left[\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (75\%)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (25\%)}}\right]$$

$$= \frac{1}{2}\left[\square - \square\right]$$

$$\beta_1 = \square$$

2. Pullback

$$\beta_2 = \frac{1}{2}\left[\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (40°)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (30°)}}\right]$$

$$= \frac{1}{2}\left[\square - \square\right]$$

$$\beta_2 = \square$$

3. Type of Ball

$$\beta_3 = \frac{1}{2}\left[\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (White)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (Orange)}}\right]$$

$$= \frac{1}{2}\left[\square - \square\right]$$

$$\beta_3 = \square$$

PB-L_8 Worksheet Test Subject ID: 11900

4. Arm Material

$$\beta_4 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (ALUMINUM)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (MAGNESIUM)}}\right)$$

$$= \frac{1}{2}\left(\square - \square\right)$$

$$\beta_4 = \square$$

5. Launch Angle

$$\beta_5 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (45°)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (60°)}}\right)$$

$$= \frac{1}{2}\left(\square - \square\right)$$

$$\beta_5 = \square$$

6. Number of Rubber Bands

$$\beta_6 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (2)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (3)}}\right)$$

$$= \frac{1}{2}\left(\square - \square\right)$$

$$\beta_6 = \square$$

7. Ambient Temperature

$$\beta_7 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (32°F)}} - \underbrace{\frac{1}{4}\left(\square + \square + \square + \square\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (72°F)}}\right)$$

$$= \frac{1}{2}\left(\square - \square\right)$$

$$\beta_7 = \square$$

Main Study
Filled-In Worksheet for Participants using PB-L_8
(2 pages)

This page intentionally left blank.

Overall Average Response.

$$\beta_0 = \underbrace{\frac{1}{8}\left(\boxed{75.9} + \boxed{53.0} + \boxed{123.0} + \boxed{79.9} + \boxed{105.3} + \boxed{80.9} + \boxed{126.5} + \boxed{88.4}\right)}_{\text{MEAN VALUE OVER ALL TRIALS}}$$

$$\beta_0 = \boxed{91.6}$$

1. Relative Humidity

$$\beta_1 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\boxed{53.0} + \boxed{123.0} + \boxed{105.3} + \boxed{88.4}\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (75\%)}} - \underbrace{\frac{1}{4}\left(\boxed{75.9} + \boxed{123.0} + \boxed{80.9} + \boxed{88.4}\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (25\%)}}\right)$$

$$= \frac{1}{2}\left(\boxed{92.4} - \boxed{90.8}\right)$$

$$\beta_1 = \boxed{0.8}$$

2. Pullback

$$\beta_2 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\boxed{123.0} + \boxed{79.9} + \boxed{126.5} + \boxed{88.4}\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (40°)}} - \underbrace{\frac{1}{4}\left(\boxed{75.9} + \boxed{53.0} + \boxed{126.5} + \boxed{88.4}\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (30°)}}\right)$$

$$= \frac{1}{2}\left(\boxed{104.4} - \boxed{78.8}\right)$$

$$\beta_2 = \boxed{12.8}$$

3. Type of Ball

$$\beta_3 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(\boxed{53.0} + \boxed{123.0} + \boxed{80.9} + \boxed{126.5}\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (WHITE)}} - \underbrace{\frac{1}{4}\left(\boxed{75.9} + \boxed{79.9} + \boxed{105.3} + \boxed{88.4}\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (ORANGE)}}\right)$$

$$= \frac{1}{2}\left(\boxed{95.9} - \boxed{87.4}\right)$$

$$\beta_3 = \boxed{4.2}$$

4. Arm Material

$$\beta_4 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(105.3 + 80.9 + 126.5 + 88.4\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (ALUMINUM)}} - \underbrace{\frac{1}{4}\left(75.9 + 53.0 + 123.0 + 79.9\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (MAGNESIUM)}}\right)$$

$$= \frac{1}{2}\left(\boxed{100.3} - \boxed{82.9}\right)$$

$$\beta_4 = \boxed{8.7}$$

5. Launch Angle

$$\beta_5 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(123.0 + 79.9 + 105.3 + 80.9\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (45°)}} - \underbrace{\frac{1}{4}\left(75.9 + 53.0 + 123.0 + 79.9\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (60°)}}\right)$$

$$= \frac{1}{2}\left(\boxed{97.3} - \boxed{85.9}\right)$$

$$\beta_5 = \boxed{5.7}$$

6. Number of Rubber Bands

$$\beta_6 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(53.0 + 79.9 + 80.9 + 88.4\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (2)}} - \underbrace{\frac{1}{4}\left(75.9 + 53.0 + 105.3 + 80.9\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (3)}}\right)$$

$$= \frac{1}{2}\left(\boxed{75.5} - \boxed{107.7}\right)$$

$$\beta_6 = \boxed{-16.1}$$

7. Ambient Temperature

$$\beta_7 = \frac{1}{2}\left(\underbrace{\frac{1}{4}\left(53.0 + 79.9 + 105.3 + 126.5\right)}_{\text{MEAN RESPONSE AT "+1" SETTING (32°F)}} - \underbrace{\frac{1}{4}\left(75.9 + 123.0 + 105.3 + 126.5\right)}_{\text{MEAN RESPONSE AT "-1" SETTING (72°F)}}\right)$$

$$= \frac{1}{2}\left(\boxed{91.2} - \boxed{92.0}\right)$$

$$\beta_7 = \boxed{-0.4}$$

Main Study
Filled-In Design Table for Participants using PB-L_8
(1 page)

This page intentionally left blank.

Main Study - Filled-In Design Table for Participants using PB-L_8

Trial	Relative Humidity (X_1)	Pullback (X_2)	Type of Ball (X_3)	Arm Material (X_4)	Launch Angle (X_5)	Rubber Bands (X_6)	Ambient Temperature (X_7)	Landing Position (Y)
1	$^{-1}$25%	$^{-1}$30°	$^{-1}$Orange	$^{-1}$Magnesium	$^{-1}$60°	$^{-1}$3	$^{-1}$72°F	75.9 in
2	$^{+1}$75%	$^{-1}$30°	$^{+1}$ White	$^{-1}$Magnesium	$^{-1}$60°	$^{+1}$2	$^{+1}$32°F	53.0 in
3	$^{+1}$75%	$^{+1}$40°	$^{+1}$ White	$^{-1}$Magnesium	$^{+1}$45°	$^{-1}$3	$^{-1}$72°F	123.0 in
4	$^{-1}$25%	$^{+1}$40°	$^{-1}$Orange	$^{-1}$Magnesium	$^{+1}$45°	$^{+1}$2	$^{+1}$32°F	79.9 in
5	$^{+1}$75%	$^{-1}$30°	$^{-1}$Orange	$^{+1}$ Aluminum	$^{+1}$45°	$^{-1}$3	$^{+1}$32°F	105.3 in
6	$^{-1}$25%	$^{-1}$30°	$^{+1}$ White	$^{+1}$ Aluminum	$^{+1}$45°	$^{+1}$2	$^{-1}$72°F	80.9 in
7	$^{-1}$25%	$^{+1}$40°	$^{+1}$ White	$^{+1}$ Aluminum	$^{-1}$60°	$^{-1}$3	$^{+1}$32°F	126.5 in
8	$^{+1}$75%	$^{+1}$40°	$^{-1}$Orange	$^{+1}$ Aluminum	$^{-1}$60°	$^{+1}$2	$^{-1}$72°F	88.4 in
Best								

$$\widehat{Y} = \underbrace{91.61}_{\beta_0} + \underbrace{0.82}_{\beta_1} X_1 + \underbrace{12.82}_{\beta_2} X_2 + \underbrace{4.25}_{\beta_3} X_3 + \underbrace{8.67}_{\beta_4} X_4 + \underbrace{5.67}_{\beta_5} X_5 + \underbrace{-16.06}_{\beta_6} X_6 + \underbrace{-0.44}_{\beta_7} X_7$$

This page intentionally left blank.

Main Study
Tabulated Estimation Results for Participants using PB-L_8
(2 pages)

This page intentionally left blank.

Main Study - Tabulated Estimation Results for Participants using PB-L_8 (Page 1 of 2)

Relative Humidity	Pullback	Type of Table Tennis Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Predicted Landing Position	Predicted Target Error
75%	30°	White	Aluminum	45°	3	72°F	96.00 in	−0.00 in
25%	40°	Orange	Magnesium	60°	2	72°F	95.71 in	0.29 in
75%	30°	Orange	Magnesium	60°	2	72°F	96.47 in	−0.47 in
25%	40°	White	Aluminum	45°	3	72°F	95.24 in	0.76 in
75%	40°	White	Aluminum	45°	3	72°F	96.88 in	−0.88 in
75%	40°	Orange	Aluminum	45°	2	32°F	94.85 in	1.15 in
25%	30°	Orange	Magnesium	60°	2	72°F	94.83 in	1.17 in
25%	30°	White	Magnesium	45°	3	32°F	97.20 in	−1.20 in
75%	40°	Orange	Magnesium	60°	2	72°F	97.35 in	−1.35 in
25%	30°	White	Aluminum	45°	3	72°F	94.36 in	1.64 in
75%	30°	Orange	Aluminum	45°	2	32°F	93.97 in	2.03 in
25%	40°	White	Magnesium	45°	3	32°F	98.08 in	−2.08 in
25%	40°	Orange	Aluminum	45°	2	32°F	93.21 in	2.79 in
75%	30°	White	Magnesium	45°	3	32°F	98.84 in	−2.84 in
25%	30°	Orange	Aluminum	45°	2	32°F	92.33 in	3.67 in
75%	40°	White	Magnesium	45°	3	32°F	99.72 in	−3.72 in
25%	30°	White	Aluminum	60°	2	32°F	100.63 in	−4.63 in
25%	30°	Orange	Aluminum	45°	2	72°F	100.83 in	−4.83 in
75%	40°	White	Magnesium	60°	3	72°F	90.88 in	5.12 in
25%	40°	White	Aluminum	60°	2	32°F	101.51 in	−5.51 in
25%	40°	Orange	Aluminum	45°	2	72°F	101.71 in	−5.71 in
75%	30°	White	Magnesium	60°	3	72°F	90.00 in	6.00 in
75%	30°	White	Aluminum	60°	2	32°F	102.27 in	−6.27 in
75%	30°	Orange	Aluminum	45°	2	72°F	102.47 in	−6.47 in
25%	40°	White	Magnesium	60°	3	72°F	89.24 in	6.76 in
75%	40°	White	Aluminum	60°	2	32°F	103.15 in	−7.15 in
75%	40°	Orange	Magnesium	60°	2	32°F	88.85 in	7.15 in
75%	40°	Orange	Aluminum	45°	2	72°F	103.35 in	−7.35 in
75%	40°	White	Aluminum	45°	3	32°F	88.38 in	7.62 in
25%	30°	White	Magnesium	60°	3	72°F	88.36 in	7.64 in
25%	30°	Orange	Magnesium	45°	2	32°F	103.68 in	−7.68 in
75%	30°	Orange	Magnesium	60°	2	32°F	87.97 in	8.03 in
75%	30°	White	Aluminum	45°	3	32°F	87.50 in	8.50 in
25%	40°	Orange	Magnesium	45°	2	32°F	104.55 in	−8.55 in
25%	40°	Orange	Magnesium	60°	2	32°F	87.21 in	8.79 in
25%	40°	White	Aluminum	45°	3	32°F	86.74 in	9.26 in
75%	30°	Orange	Magnesium	45°	2	32°F	105.32 in	−9.32 in
25%	30°	Orange	Magnesium	60°	2	32°F	86.33 in	9.67 in
25%	30°	White	Magnesium	45°	3	72°F	105.70 in	−9.70 in
75%	40°	Orange	Aluminum	60°	2	72°F	86.01 in	9.99 in
25%	30°	White	Aluminum	45°	3	32°F	85.86 in	10.14 in
75%	40°	Orange	Magnesium	45°	2	32°F	106.20 in	−10.20 in
25%	40°	White	Magnesium	45°	3	72°F	106.58 in	−10.58 in
75%	30°	Orange	Aluminum	60°	2	72°F	85.13 in	10.87 in
75%	30°	White	Magnesium	45°	3	72°F	107.34 in	−11.34 in
25%	40°	Orange	Aluminum	60°	2	72°F	84.37 in	11.63 in
75%	40°	White	Magnesium	45°	3	72°F	108.22 in	−12.22 in
25%	30°	Orange	Aluminum	60°	2	72°F	83.49 in	12.51 in
25%	30°	White	Aluminum	60°	2	72°F	109.13 in	−13.13 in
75%	40°	Orange	Magnesium	45°	3	72°F	82.58 in	13.42 in
75%	40°	White	Magnesium	60°	3	32°F	82.38 in	13.62 in
25%	40°	White	Aluminum	60°	2	72°F	110.01 in	−14.01 in
75%	30°	Orange	Magnesium	45°	3	72°F	81.70 in	14.30 in
75%	30°	White	Magnesium	60°	3	32°F	81.50 in	14.50 in
75%	30°	White	Aluminum	60°	2	72°F	110.77 in	−14.77 in
25%	40°	Orange	Magnesium	45°	3	72°F	80.94 in	15.06 in
25%	40°	White	Magnesium	60°	3	32°F	80.74 in	15.26 in
75%	40°	White	Aluminum	60°	2	72°F	111.65 in	−15.65 in
25%	30°	Orange	Magnesium	45°	3	72°F	80.06 in	15.94 in
25%	30°	White	Magnesium	60°	2	32°F	111.97 in	−15.97 in
25%	30°	White	Magnesium	60°	3	32°F	79.86 in	16.14 in
25%	30°	Orange	Magnesium	45°	2	72°F	112.18 in	−16.18 in
75%	40°	White	Aluminum	60°	3	72°F	79.53 in	16.47 in
25%	40°	White	Magnesium	60°	2	32°F	112.85 in	−16.85 in

Main Study - Tabulated Estimation Results for Participants using PB-L_8 (Page 2 of 2)

Relative Humidity	Pullback	Type of Table Tennis Ball	Arm Material	Launch Angle	Rubber Bands	Ambient Temperature	Predicted Landing Position	Predicted Target Error
25%	40°	Orange	Magnesium	45°	2	72°F	113.05 in	−17.05 in
75%	30°	White	Aluminum	60°	3	72°F	78.66 in	17.34 in
75%	30°	White	Magnesium	60°	2	32°F	113.61 in	−17.61 in
75%	30°	Orange	Magnesium	45°	2	72°F	113.82 in	−17.82 in
25%	40°	White	Aluminum	60°	3	72°F	77.89 in	18.11 in
75%	40°	White	Magnesium	60°	2	32°F	114.49 in	−18.49 in
75%	40°	Orange	Aluminum	60°	2	32°F	77.51 in	18.49 in
75%	40°	Orange	Magnesium	45°	2	72°F	114.69 in	−18.69 in
25%	30°	White	Aluminum	60°	3	72°F	77.02 in	18.98 in
75%	30°	Orange	Aluminum	60°	2	32°F	76.63 in	19.37 in
25%	40°	Orange	Aluminum	60°	2	32°F	75.87 in	20.13 in
25%	30°	Orange	Aluminum	60°	2	32°F	74.99 in	21.01 in
75%	40°	Orange	Magnesium	45°	3	32°F	74.08 in	21.92 in
25%	30°	White	Aluminum	45°	2	32°F	117.97 in	−21.97 in
75%	30°	Orange	Magnesium	45°	3	32°F	73.20 in	22.80 in
25%	40°	White	Aluminum	45°	2	32°F	118.85 in	−22.85 in
25%	40°	Orange	Magnesium	45°	3	32°F	72.44 in	23.56 in
75%	30°	White	Aluminum	45°	2	32°F	119.62 in	−23.62 in
25%	30°	Orange	Magnesium	45°	3	32°F	71.56 in	24.44 in
25%	30°	White	Magnesium	60°	2	72°F	120.47 in	−24.47 in
75%	40°	White	Aluminum	45°	2	32°F	120.49 in	−24.49 in
75%	40°	Orange	Aluminum	45°	3	72°F	71.24 in	24.76 in
75%	40°	White	Aluminum	60°	3	32°F	71.04 in	24.96 in
25%	40°	White	Magnesium	60°	2	72°F	121.35 in	−25.35 in
75%	30°	Orange	Aluminum	45°	3	72°F	70.36 in	25.64 in
75%	30°	White	Aluminum	60°	3	32°F	70.16 in	25.84 in
75%	30°	White	Magnesium	60°	2	72°F	122.11 in	−26.11 in
25%	40°	Orange	Aluminum	45°	3	72°F	69.60 in	26.40 in
25%	40°	White	Aluminum	60°	3	32°F	69.40 in	26.60 in
75%	40°	White	Magnesium	60°	2	72°F	122.99 in	−26.99 in
25%	30°	Orange	Aluminum	45°	3	72°F	68.72 in	27.28 in
25%	30°	White	Aluminum	60°	3	32°F	68.52 in	27.48 in
25%	30°	White	Aluminum	45°	2	72°F	126.47 in	−30.47 in
75%	40°	Orange	Magnesium	60°	3	72°F	65.24 in	30.76 in
25%	40°	White	Aluminum	45°	2	72°F	127.35 in	−31.35 in
75%	30°	Orange	Magnesium	60°	3	72°F	64.36 in	31.64 in
75%	30°	White	Aluminum	45°	2	72°F	128.11 in	−32.11 in
25%	40°	Orange	Magnesium	60°	3	72°F	63.60 in	32.40 in
75%	40°	White	Aluminum	45°	2	72°F	128.99 in	−32.99 in
75%	40°	Orange	Aluminum	45°	3	32°F	62.74 in	33.26 in
25%	30°	Orange	Magnesium	60°	3	72°F	62.72 in	33.28 in
25%	30°	White	Magnesium	45°	2	32°F	129.32 in	−33.32 in
75%	30°	Orange	Aluminum	45°	3	32°F	61.86 in	34.14 in
25%	40°	White	Magnesium	45°	2	32°F	130.20 in	−34.20 in
25%	40°	Orange	Aluminum	45°	3	32°F	61.10 in	34.90 in
75%	30°	White	Magnesium	45°	2	32°F	130.96 in	−34.96 in
25%	30°	Orange	Aluminum	45°	3	32°F	60.22 in	35.78 in
75%	40°	White	Magnesium	45°	2	32°F	131.84 in	−35.84 in
75%	40°	Orange	Magnesium	60°	3	32°F	56.74 in	39.26 in
75%	30°	Orange	Magnesium	60°	3	32°F	55.86 in	40.14 in
25%	40°	Orange	Magnesium	60°	3	32°F	55.10 in	40.90 in
25%	30°	Orange	Magnesium	60°	3	32°F	54.22 in	41.78 in
25%	30°	White	Magnesium	45°	2	72°F	137.82 in	−41.82 in
75%	40°	Orange	Aluminum	60°	3	72°F	53.89 in	42.11 in
25%	40°	White	Magnesium	45°	2	72°F	138.69 in	−42.69 in
75%	30°	Orange	Aluminum	60°	3	72°F	53.02 in	42.98 in
75%	30°	White	Magnesium	45°	2	72°F	139.46 in	−43.46 in
25%	40°	Orange	Aluminum	60°	3	72°F	52.25 in	43.75 in
75%	40°	White	Magnesium	45°	2	72°F	140.34 in	−44.34 in
25%	30°	Orange	Aluminum	60°	3	72°F	51.38 in	44.62 in
75%	40°	Orange	Aluminum	60°	3	32°F	45.39 in	50.61 in
75%	30°	Orange	Aluminum	60°	3	32°F	44.52 in	51.48 in
25%	40°	Orange	Aluminum	60°	3	32°F	43.75 in	52.25 in
25%	30°	Orange	Aluminum	60°	3	32°F	42.88 in	53.12 in

Appendix C

Tabulated Experimental Data

The raw data from the main experiment come from four source materials. For each participant, there is a demographic survey, an audio recording, a video recording, and the predictions written in the design table during the exercise. Those sources were used to generate the four tables of raw data in this appendix, and the data in these tables are used in the analyses for the main experiment. In each table, the randomly assigned subject identification numbers that link all of the results together are given.

Table C.1 on page 247 gives the answers to the demographic survey shown on page 143. The data in this table were not used in the analyses but are provided here for completeness.

Table C.2 on page 249 gives the answers to the domain knowledge assessment quiz, administered before the participant starts the design task. For each control factor in the 2^7 catapult, the participant was asked to predict qualitatively what would happen to the ball's landing distance if the factor is changed from its nominal to its alternate value. The answers are given with respect to the landing distance for the nominal configuration. In this simplified scoring approach, one point is given for each prediction in the correct relative direction. Prediction magnitudes are not considered in assigning points. Each participant's score is given at the end of the corresponding row, and the sums of individual scores for each factor are given in the last row of the table.

Table C.3 on page 253 gives the raw data collected during each of the 385 trials in the design task. For each trial, the participant is asked to predict the outcome of the simulation for a given configuration of the catapult. The participant may base this prediction upon one or more of the outcomes previously revealed. For participants in the aOFAT group, there is only one comparison that makes sense for the anchor-and-adjust strategy used by most: the configuration that differs by only one control factor from the one being predicted. Participants in the PB-L_8 group were asked which previous trial(s) were referenced to make the prediction. In this table, the first two columns give the subject ID and trial number. The third column gives the reference trial number(s). If only one trial was referenced,

its outcome value is given in the fourth column. If more than one trial was referenced, a simple anchor-and-adjust strategy was not used and the value in the fourth column is unknown. The fifth column gives the predicted outcome for the current trial, and the sixth column gives the provided simulation answer for the current trial. The seventh trial indicates whether or not the anomaly is elicited: if the flawed control factor is at the nominal value in the reference trial(s) but at the alternate value in the predicted trial, then the anomaly is elicited. The last four columns in this table give the result of the ratings by independent judges for signs of surprise. First are the two ratings values selected from the five choices, then whether the raters agreed according to the specified decision rule. If they did agree, the final column gives the agreed-upon surprise assessment.

Table C.4 on page 273 gives the result of the debriefing questioning for each participant. If the flaw was identified before the participant was told that there is a flaw, the outcome is classified as an "Unprompted" discovery of the flaw. If the flaw was identified after the participant was told that there is a flaw, the outcome is classified as a "Prompted" discovery. If the flaw was not correctly identified by the participant, the outcome is classified as "It wasn't". The table is divided into control group on the left and treatment group on the right, and the number of checks in each column is provided in the last row.

Table C.1: Participant Demographics

Subject ID	Gender	Age Group	Highest Degree	Post BS Work Years	Design of Expts. User Level	Simulation User Level	Simulation Exp. Years
01039	M	35-45	MS	10-15	Novice	Developer	15-20
01042	M	25-35	MS	2-5	Novice	Developer	5-10
01083	M	45-55	MS+	>20	Novice	Intermediate	5-10
01104	M	45-55	MS	>20	Intermediate	Intermediate	2-5
01153	M	45-55	BS	>20	Expert	Developer	>20
01194	F	25-35	MS	2-5	Novice	Intermediate	1-2
01225	M	25-35	MS	2-5	Novice	Developer	5-10
01231	M	45-55	MS	>20	Novice	Developer	2-5
01247	M	25-35	MS	5-10	Intermediate	Intermediate	5-10
01295	M	35-45	PhD	5-10	Novice	Developer	15-20
01317	M	<25	BS-	<1	Novice	Novice	<1
01335	M	25-35	MS+	2-5	Novice	Expert	5-10
01353	M	35-45	PhD	5-10	Novice	Expert	2-5
01391	M	45-55	PhD	>20	Expert	Developer	>20
01488	M	35-45	MS	15-20	Intermediate	Intermediate	1-2
01596	M	>55	MS	>20	Novice	Expert	5-10
01757	M	25-35	PhD	5-10	Novice	Expert	2-5
01785	M	45-55	BS	>20	Intermediate	Novice	<1
01791	M	45-55	MS	>20	Novice	Intermediate	1-2
01794	M	25-35	MS	10-15	Intermediate	Intermediate	5-10
01797	M	>55	PhD	>20	Novice	Developer	>20
01820	M	25-35	PhD	2-5	Expert	Intermediate	2-5
01858	M	35-45	BS	5-10	Expert	Expert	2-5
01893	M	>55	PhD	>20	Novice	Developer	>20
01905	M	25-35	MS	5-10	Novice	Expert	5-10
01957	M	45-55	MS	>20	Novice	Novice	<1
01988	M	>55	PhD	>20	Novice	Expert	>20
11018	M	>55	PhD	>20	Novice	Expert	>20
11065	M	45-55	PhD	10-15	Novice	Intermediate	1-2
11070	M	35-45	MS	10-15	Intermediate	Novice	<1
11094	F	25-35	MS	5-10	Novice	Expert	5-10

Table C.1: Participant Demographics (continued)

Subject ID	Gender	Age Group	Highest Degree	Post BS Work Years	Design of Expts. User Level	Simulation User Level	Simulation Exp. Years
11097	M	25-35	MS	2-5	Intermediate	Intermediate	5-10
11112	M	45-55	PhD	10-15	Novice	Intermediate	–
11123	F	45-55	MS	>20	Intermediate	Novice	<1
11130	M	35-45	MS	15-20	Novice	Intermediate	5-10
11138	M	25-35	MS	5-10	Novice	Expert	5-10
11174	M	25-35	MS	<1	Novice	Intermediate	2-5
11199	M	45-55	MS	>20	Novice	Expert	5-10
11207	M	45-55	BS	>20	Expert	Expert	5-10
11257	M	45-55	MS	>20	Novice	Expert	5-10
11294	M	25-35	MS	10-15	Intermediate	Expert	2-5
11363	M	45-55	MS	>20	Intermediate	Developer	15-20
11499	M	25-35	BS	5-10	Novice	Novice	<1
11539	M	35-45	PhD	2-5	Novice	Developer	5-10
11572	M	>55	MS	>20	Expert	Intermediate	10-15
11576	M	35-45	PhD	15-20	Novice	Developer	15-20
11588	M	25-35	MS	2-5	Novice	Intermediate	2-5
11711	M	45-55	PhD	>20	Intermediate	Developer	10-15
11881	M	<25	BS+	<1	Intermediate	Intermediate	<1
11900	M	25-35	MS	2-5	Novice	Expert	2-5
11901	M	<25	BS+	<1	Intermediate	Intermediate	<1
11972	M	>55	MS	>20	Novice	Intermediate	5-10
11975	M	<25	BS-	1-2	Novice	Novice	<1
11979	M	25-35	BS+	2-5	Novice	Novice	1-2

Table C.2: Raw Data from Quiz for Domain Knowledge Score. The correct answers are as follows and may be corroborated using the selected nominal configuration on page 212 and the main effects plot on page 120: Humidity – slightly farther, Pullback – farther, Type of Ball – farther, Arm Material – shorter, Launch Angle – farther, Rubber Bands – shorter, Temperature – slightly shorter. One point is given for each prediction in the correct direction.

Subject ID	Humidity Prediction : Score	Pullback Prediction : Score	Type of Ball Prediction : Score	Arm Material Prediction : Score	Launch Angle Prediction : Score	Rubber Bands Prediction : Score	Temperature Prediction : Score	Total Score
01039	no change : 0	farther : 1	farther : 1	shorter : 1	shorter : 0	shorter : 1	no change : 0	4
01042	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01083	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01104	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01153	shorter : 0	farther : 1	no change : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
01194	no change : 0	farther : 1	no change : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
01225	slightly shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
01231	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01247	slightly shorter : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	slightly shorter : 1	5
01295	slightly shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
01317	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01335	slightly shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	slightly farther : 0	5
01353	slightly shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01391	no change : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
01488	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	farther : 0	5
01596	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01757	shorter : 0	farther : 1	slightly farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01785	slightly shorter : 0	farther : 1	no change : 0	shorter : 1	farther : 1	shorter : 1	slightly farther : 0	4

Table C.2: Raw Data from Quiz for Domain Knowledge Score (continued)

Subject ID	Humidity Prediction : Score	Pullback Prediction : Score	Type of Ball Prediction : Score	Arm Material Prediction : Score	Launch Angle Prediction : Score	Rubber Bands Prediction : Score	Temperature Prediction : Score	Total Score
01791	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
01794	slightly shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	slightly farther : 0	5
01797	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
01820	shorter : 0	farther : 1	farther : 1	shorter : 1	shorter : 0	shorter : 1	shorter : 1	5
01858	shorter : 0	farther : 1	no change : 0	shorter : 1	shorter : 0	shorter : 1	no change : 0	3
01893	no change : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
01905	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	slightly shorter : 1	6
01957	shorter : 0	farther : 1	farther : 1	farther : 0	shorter : 0	shorter : 1	shorter : 1	4
01988	farther : 1	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	slightly shorter : 1	7
11018	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
11065	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
11070	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
11094	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
11097	shorter : 0	farther : 1	no change : 0	shorter : 1	farther : 1	shorter : 1	shorter : 1	5
11112	slightly farther : 1	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	7
11123	no change : 0	farther : 1	shorter : 0	farther : 0	shorter : 0	shorter : 1	no change : 0	2
11130	slightly shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
11138	no change : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
11174	no change : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
11199	no change : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	slightly shorter : 1	5

Table C.2: Raw Data from Quiz for Domain Knowledge Score (continued)

Subject ID	Humidity Prediction : Score	Pullback Prediction : Score	Type of Ball Prediction : Score	Arm Material Prediction : Score	Launch Angle Prediction : Score	Rubber Bands Prediction : Score	Temperature Prediction : Score	Total Score
11207	shorter : 0	farther : 1	farther : 1	shorter : 1	shorter : 0	shorter : 1	shorter : 1	5
11257	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	no change : 0	5
11294	slightly shorter : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	slightly farther : 0	4
11363	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
11499	slightly shorter : 0	farther : 1	no change : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
11539	slightly shorter : 0	farther : 1	no change : 0	shorter : 1	farther : 1	shorter : 1	slightly shorter : 1	5
11572	no change : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	slightly farther : 0	4
11576	slightly shorter : 0	farther : 1	farther : 1	n/a : 0	farther : 1	shorter : 1	shorter : 1	5
11588	slightly shorter : 0	farther : 1	farther : 1	farther : 0	farther : 1	shorter : 1	slightly farther : 0	4
11711	no change : 0	farther : 1	farther : 1	no change : 0	farther : 1	no change : 0	no change : 0	3
11881	no change : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
11900	slightly shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
11901	shorter : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	shorter : 1	5
11972	shorter : 0	farther : 1	farther : 1	shorter : 1	farther : 1	shorter : 1	shorter : 1	6
11975	no change : 0	farther : 1	shorter : 0	shorter : 1	farther : 1	shorter : 1	no change : 0	4
11979	shorter : 0	farther : 1	farther : 1	shorter : 1	shorter : 0	shorter : 1	no change : 0	4
Total	: 2	: 54	: 36	: 49	: 47	: 53	: 26	

This page intentionally left blank.

Table C.3: Raw Data from Distance Predictions and Surprise Ratings

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01039	2	1	75.87	76.00	75.97	no	Agree	Agree	yes	yes
01039	3	2	75.97	86.00	98.44	no	Strongly Agree	Strongly Agree	yes	yes
01039	4	3	98.44	115.00	106.84	no	Neutral	Disagree	yes	no
01039	5	3	98.44	70.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01039	6	3	98.44	70.00	113.96	no	Strongly Agree	Agree	yes	yes
01039	7	3	98.44	60.00	70.64	no	Disagree	Disagree	yes	no
01039	8	3	98.44	98.40	95.13	no	Neutral	Neutral	yes	no
01042	2	1	75.87	75.00	75.97	no	Agree	Agree	yes	yes
01042	3	2	75.97	88.00	98.44	no	Agree	Agree	yes	yes
01042	4	3	98.44	118.00	106.84	no	Neutral	Neutral	yes	no
01042	5	3	98.44	63.00	118.03	yes	Neutral	Strongly Disagree	yes	no
01042	6	3	98.44	110.00	113.96	no	Neutral	Strongly Disagree	yes	no
01042	7	3	98.44	66.00	70.64	no	Disagree	Disagree	yes	no
01042	8	3	98.44	96.00	95.13	no	Strongly Disagree	Disagree	yes	no
01083	2	1	75.87	77.20	75.97	no	Neutral	Disagree	yes	no
01083	3	2	75.97	82.00	98.44	no	Agree	Strongly Agree	yes	yes
01083	4	3	98.44	105.00	106.84	no	Disagree	Neutral	yes	no
01083	5	3	98.44	85.00	118.03	yes	Strongly Agree	Agree	yes	yes
01083	6	3	98.44	118.00	113.96	no	Neutral	Agree	no	–

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01083	7	3	98.44	73.00	70.64	no	Agree	Strongly Agree	yes	yes
01083	8	3	98.44	98.20	95.13	no	Neutral	Agree	no	–
01104	2	1	75.87	74.00	75.97	no	Agree	Agree	yes	yes
01104	3	2	75.97	84.00	98.44	no	Neutral	Neutral	yes	no
01104	4	3	98.44	103.00	106.84	no	Neutral	Strongly Disagree	yes	no
01104	5	3	98.44	90.00	118.03	yes	Agree	Agree	yes	yes
01104	6	3	98.44	120.00	113.96	no	Neutral	Disagree	yes	no
01104	7	3	98.44	85.00	70.64	no	Neutral	Disagree	yes	no
01104	8	3	98.44	96.00	95.13	no	Disagree	Disagree	yes	no
01153	2	1	75.87	72.00	75.97	no	Strongly Agree	Strongly Agree	yes	yes
01153	3	2	75.97	87.40	98.44	no	Agree	Neutral	no	–
01153	4	3	98.44	100.00	106.84	no	Agree	Neutral	no	–
01153	5	3	98.44	65.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01153	6	3	98.44	147.00	113.96	no	Agree	Neutral	no	–
01153	7	3	98.44	70.00	70.64	no	Neutral	Disagree	yes	no
01153	8	3	98.44	98.00	95.13	no	Neutral	Strongly Disagree	yes	no
01194	2	1	75.87	75.90	75.97	no	Disagree	Neutral	yes	no
01194	3	2	75.97	91.20	98.44	no	Neutral	Agree	no	–
01194	4	3	98.44	105.00	106.84	no	Agree	Strongly Agree	yes	yes
01194	5	3	98.44	65.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01194	6	3	98.44	118.00	113.96	no	Neutral	Disagree	yes	no
01194	7	3	98.44	65.00	70.64	no	Neutral	Disagree	yes	no
01194	8	3	98.44	98.00	95.13	no	Agree	Disagree	no	–
01225	2	1	75.87	75.00	75.97	no	Neutral	Disagree	yes	no
01225	3	2	75.97	87.40	98.44	no	Neutral	Strongly Agree	no	–
01225	4	3	98.44	119.00	106.84	no	Neutral	Neutral	yes	no
01225	5	3	98.44	63.00	118.03	yes	Disagree	Neutral	yes	no
01225	6	3	98.44	117.00	113.96	no	Strongly Agree	Agree	yes	yes
01225	7	3	98.44	65.60	70.64	no	Disagree	Neutral	yes	no
01225	8	3	98.44	97.80	95.13	no	Strongly Disagree	Neutral	yes	no
01231	2	1	75.87	74.00	75.97	no	Agree	Agree	yes	yes
01231	3	2	75.97	110.00	98.44	no	Neutral	Neutral	yes	no
01231	4	3	98.44	101.00	106.84	no	Disagree	Strongly Disagree	yes	no
01231	5	3	98.44	97.50	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01231	6	3	98.44	110.00	113.96	no	Strongly Disagree	Strongly Agree	no	–
01231	7	3	98.44	90.00	70.64	no	Strongly Agree	Neutral	no	–
01231	8	3	98.44	96.00	95.13	no	Neutral	Neutral	yes	no
01247	2	1	75.87	74.00	75.97	no	Disagree	Strongly Agree	no	–
01247	3	2	75.97	100.00	98.44	no	Agree	Neutral	no	–
01247	4	3	98.44	90.00	106.84	no	Agree	Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01247	5	3	98.44	130.00	118.03	yes	Disagree	Neutral	yes	no
01247	6	3	98.44	120.00	113.96	no	Disagree	Disagree	yes	no
01247	7	3	98.44	66.00	70.64	no	Neutral	Neutral	yes	no
01247	8	3	98.44	98.00	95.13	no	Disagree	Neutral	yes	no
01295	2	1	75.87	75.50	75.97	no	Disagree	Disagree	yes	no
01295	3	2	75.97	82.00	98.44	no	Agree	Strongly Agree	yes	yes
01295	4	3	98.44	105.00	106.84	no	Neutral	Neutral	yes	no
01295	5	3	98.44	89.00	118.03	yes	Agree	Agree	yes	yes
01295	6	3	98.44	130.00	113.96	no	Agree	Neutral	no	–
01295	7	3	98.44	69.00	70.64	no	Disagree	Disagree	yes	no
01295	8	3	98.44	98.00	95.13	no	Neutral	Disagree	yes	no
01317	2	1	75.87	70.00	75.97	no	Disagree	Strongly Agree	no	–
01317	3	2	75.97	79.00	98.44	no	Strongly Agree	Strongly Agree	yes	yes
01317	4	3	98.44	105.40	106.84	no	Agree	Strongly Agree	yes	yes
01317	5	3	98.44	70.00	118.03	yes	Agree	Strongly Agree	yes	yes
01317	6	3	98.44	132.00	113.96	no	Disagree	Disagree	yes	no
01317	7	3	98.44	65.60	70.64	no	Agree	Neutral	no	–
01317	8	3	98.44	91.40	95.13	no	Neutral	Neutral	yes	no
01335	2	1	75.87	70.00	75.97	no	Agree	Strongly Agree	yes	yes
01335	3	2	75.97	87.20	98.44	no	Disagree	Neutral	yes	no

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01335	4	3	98.44	118.40	106.84	no	Neutral	Disagree	yes	no
01335	5	3	98.44	78.40	118.03	yes	Agree	Strongly Agree	yes	yes
01335	6	3	98.44	112.00	113.96	no	Agree	Agree	yes	yes
01335	7	3	98.44	80.00	70.64	no	Disagree	Strongly Disagree	yes	no
01335	8	3	98.44	98.60	95.13	no	Neutral	Neutral	yes	no
01353	2	1	75.87	65.00	75.97	no	Neutral	Neutral	yes	no
01353	3	2	75.97	100.00	98.44	no	Agree	Neutral	no	–
01353	4	3	98.44	110.00	106.84	no	Disagree	Disagree	yes	no
01353	5	3	98.44	85.00	118.03	yes	Agree	Agree	yes	yes
01353	6	3	98.44	110.00	113.96	no	Neutral	Neutral	yes	no
01353	7	3	98.44	85.00	70.64	no	Neutral	Agree	no	–
01353	8	3	98.44	80.00	95.13	no	Neutral	Disagree	yes	no
01391	2	1	75.87	75.90	75.97	no	Agree	Neutral	no	–
01391	3	2	75.97	101.30	98.44	no	Disagree	Neutral	yes	no
01391	4	3	98.44	85.00	106.84	no	Agree	Agree	yes	yes
01391	5	3	98.44	75.00	118.03	yes	Strongly Agree	Agree	yes	yes
01391	6	3	98.44	139.00	113.96	no	Agree	Neutral	no	–
01391	7	3	98.44	65.00	70.64	no	Disagree	Disagree	yes	no
01391	8	3	98.44	98.00	95.13	no	Neutral	Neutral	yes	no
01488	2	1	75.87	76.90	75.97	no	Agree	Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01488	3	2	75.97	80.00	98.44	no	Neutral	Agree	no	–
01488	4	3	98.44	85.00	106.84	no	Neutral	Neutral	yes	no
01488	5	3	98.44	70.00	118.03	yes	Neutral	Agree	no	–
01488	6	3	98.44	120.00	113.96	no	Neutral	Strongly Disagree	yes	no
01488	7	3	98.44	65.00	70.64	no	Neutral	Strongly Disagree	yes	no
01488	8	3	98.44	99.00	95.13	no	Disagree	Disagree	yes	no
01596	2	1	75.87	70.00	75.97	no	Strongly Agree	Strongly Agree	yes	yes
01596	3	2	75.97	100.00	98.44	no	Neutral	Neutral	yes	no
01596	4	3	98.44	110.00	106.84	no	Neutral	Neutral	yes	no
01596	5	3	98.44	80.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01596	6	3	98.44	125.00	113.96	no	Agree	Neutral	no	–
01596	7	3	98.44	100.00	70.64	no	Agree	Neutral	no	–
01596	8	3	98.44	96.00	95.13	no	Neutral	Neutral	yes	no
01757	2	1	75.87	71.00	75.97	no	Agree	Strongly Agree	yes	yes
01757	3	2	75.97	78.00	98.44	no	Agree	Strongly Agree	yes	yes
01757	4	3	98.44	105.00	106.84	no	Neutral	Disagree	yes	no
01757	5	3	98.44	75.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01757	6	3	98.44	110.00	113.96	no	Agree	Neutral	no	–
01757	7	3	98.44	75.00	70.64	no	Neutral	Neutral	yes	no
01757	8	3	98.44	103.00	95.13	no	Agree	Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01785	2	1	75.87	76.20	75.97	no	Agree	Strongly Agree	yes	yes
01785	3	2	75.97	90.00	98.44	no	Neutral	Disagree	yes	no
01785	4	3	98.44	99.00	106.84	no	Strongly Agree	Strongly Agree	yes	yes
01785	5	3	98.44	68.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01785	6	3	98.44	100.00	113.96	no	Disagree	Neutral	yes	no
01785	7	3	98.44	62.00	70.64	no	Neutral	Neutral	yes	no
01785	8	3	98.44	93.00	95.13	no	Disagree	Agree	no	–
01791	2	1	75.87	75.50	75.97	no	Neutral	Neutral	yes	no
01791	3	2	75.97	88.00	98.44	no	Neutral	Neutral	yes	no
01791	4	3	98.44	98.40	106.84	no	Strongly Agree	Strongly Agree	yes	yes
01791	5	3	98.44	88.00	118.03	yes	Agree	Strongly Agree	yes	yes
01791	6	3	98.44	110.00	113.96	no	Neutral	Neutral	yes	no
01791	7	3	98.44	65.00	70.64	no	Neutral	Disagree	yes	no
01791	8	3	98.44	98.50	95.13	no	Strongly Agree	Strongly Agree	yes	yes
01794	2	1	75.87	74.00	75.97	no	Agree	Agree	yes	yes
01794	3	2	75.97	84.00	98.44	no	Agree	Agree	yes	yes
01794	4	3	98.44	110.00	106.84	no	Agree	Agree	yes	yes
01794	5	3	98.44	70.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01794	6	3	98.44	113.00	113.96	no	Agree	Agree	yes	yes
01794	7	3	98.44	60.00	70.64	no	Disagree	Neutral	yes	no

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01794	8	3	98.44	99.00	95.13	no	Disagree	Neutral	yes	no
01797	2	1	75.87	75.80	75.97	no	Agree	Strongly Agree	yes	yes
01797	3	2	75.97	88.00	98.44	no	Neutral	Neutral	yes	no
01797	4	3	98.44	105.00	106.84	no	Agree	Neutral	no	–
01797	5	3	98.44	65.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01797	6	3	98.44	110.00	113.96	no	Neutral	Neutral	yes	no
01797	7	3	98.44	70.00	70.64	no	Neutral	Disagree	yes	no
01797	8	3	98.44	96.00	95.13	no	Agree	Agree	yes	yes
01820	2	1	75.87	72.00	75.97	no	Strongly Agree	Neutral	no	–
01820	3	2	75.97	80.00	98.44	no	Strongly Agree	Agree	yes	yes
01820	4	3	98.44	100.00	106.84	no	Neutral	Neutral	yes	no
01820	5	3	98.44	94.00	118.03	yes	Agree	Agree	yes	yes
01820	6	3	98.44	86.00	113.96	no	Agree	Agree	yes	yes
01820	7	3	98.44	70.00	70.64	no	Agree	Strongly Agree	yes	yes
01820	8	3	98.44	94.00	95.13	no	Strongly Disagree	Agree	no	–
01858	2	1	75.87	71.90	75.97	no	Strongly Agree	Strongly Agree	yes	yes
01858	3	2	75.97	82.00	98.44	no	Disagree	Neutral	yes	no
01858	4	3	98.44	93.00	106.84	no	Strongly Agree	Strongly Agree	yes	yes
01858	5	3	98.44	81.00	118.03	yes	Neutral	Neutral	yes	no
01858	6	3	98.44	73.00	113.96	no	Strongly Agree	Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01858	7	3	98.44	68.00	70.64	no	Agree	Agree	yes	yes
01858	8	3	98.44	96.00	95.13	no	Neutral	Neutral	yes	no
01893	2	1	75.87	75.90	75.97	no	Strongly Agree	Disagree	no	–
01893	3	2	75.97	84.00	98.44	no	Agree	Agree	yes	yes
01893	4	3	98.44	90.00	106.84	no	Agree	Agree	yes	yes
01893	5	3	98.44	80.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01893	6	3	98.44	118.00	113.96	no	Agree	Agree	yes	yes
01893	7	3	98.44	75.00	70.64	no	Agree	Agree	yes	yes
01893	8	3	98.44	98.00	95.13	no	Strongly Agree	Strongly Agree	yes	yes
01905	2	1	75.87	75.50	75.97	no	Disagree	Neutral	yes	no
01905	3	2	75.97	87.00	98.44	no	Agree	Agree	yes	yes
01905	4	3	98.44	100.00	106.84	no	Disagree	Neutral	yes	no
01905	5	3	98.44	80.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01905	6	3	98.44	110.00	113.96	no	Agree	Neutral	no	–
01905	7	3	98.44	82.00	70.64	no	Agree	Agree	yes	yes
01905	8	3	98.44	99.00	95.13	no	Agree	Agree	yes	yes
01957	2	1	75.87	73.00	75.97	no	Neutral	Neutral	yes	no
01957	3	2	75.97	80.00	98.44	no	Agree	Agree	yes	yes
01957	4	3	98.44	106.00	106.84	no	Neutral	Disagree	yes	no
01957	5	3	98.44	112.00	118.03	yes	Agree	Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
01957	6	3	98.44	110.00	113.96	no	Neutral	Neutral	yes	no
01957	7	3	98.44	92.00	70.64	no	Neutral	Neutral	yes	no
01957	8	3	98.44	92.00	95.13	no	Strongly Disagree	Disagree	yes	no
01988	2	1	75.87	78.00	75.97	no	Agree	Agree	yes	yes
01988	3	2	75.97	100.00	98.44	no	Agree	Strongly Agree	yes	yes
01988	4	3	98.44	115.00	106.84	no	Agree	Neutral	no	–
01988	5	3	98.44	85.00	118.03	yes	Strongly Agree	Strongly Agree	yes	yes
01988	6	3	98.44	115.00	113.96	no	Neutral	Disagree	yes	no
01988	7	3	98.44	85.00	70.64	no	Neutral	Agree	no	–
01988	8	3	98.44	96.00	95.13	no	Disagree	Disagree	yes	no
11018	2	1	75.87	61.60	53.02	no	Disagree	Agree	no	–
11018	3	1	75.87	104.00	122.99	no	Agree	Strongly Agree	yes	yes
11018	4	1	75.87	59.00	79.86	no	Agree	Agree	yes	yes
11018	5	1	75.87	45.00	105.32	yes	Strongly Agree	Strongly Agree	yes	yes
11018	6	1	75.87	67.00	80.94	yes	Agree	Neutral	no	–
11018	7	1	75.87	117.00	126.47	yes	Agree	Agree	yes	yes
11018	8	1	75.87	73.00	88.38	yes	Disagree	Neutral	yes	no
11065	2	1	75.87	45.00	53.02	no	Agree	Agree	yes	yes
11065	3	2	53.02	100.00	122.99	no	Neutral	Strongly Agree	no	–
11065	4	1	75.87	90.00	79.86	no	Strongly Agree	Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11065	5	<5	–	78.00	105.32	yes	Strongly Agree	Strongly Agree	yes	yes
11065	6	3	122.99	82.00	80.94	yes	Agree	Strongly Agree	yes	yes
11065	7	6	80.94	136.00	126.47	no	Neutral	Neutral	yes	no
11065	8	5	105.32	79.00	88.38	no	Agree	Neutral	no	–
11070	2	1	75.87	72.00	53.02	no	Strongly Agree	Strongly Agree	yes	yes
11070	3	2	53.02	110.00	122.99	no	Neutral	Neutral	yes	no
11070	4	3	122.99	85.00	79.86	no	Disagree	Disagree	yes	no
11070	5	3	122.99	100.00	105.32	yes	Neutral	Strongly Disagree	yes	no
11070	6	5	105.32	110.00	80.94	no	Agree	Disagree	no	–
11070	7	6	80.94	86.00	126.47	no	Agree	Neutral	no	–
11070	8	4	79.86	65.00	88.38	yes	Neutral	Neutral	yes	no
11094	2	1	75.87	50.60	53.02	no	Disagree	Strongly Agree	no	–
11094	3	2	53.02	143.00	122.99	no	Agree	Agree	yes	yes
11094	4	<4	–	84.00	79.86	no	Disagree	Agree	no	–
11094	5	3	122.99	69.00	105.32	yes	Agree	Strongly Agree	yes	yes
11094	6	2,5	–	72.00	80.94	–	Agree	Strongly Agree	yes	yes
11094	7	3	122.99	87.00	126.47	yes	Agree	Neutral	no	–
11094	8	6	80.94	57.00	88.38	no	Agree	Strongly Agree	yes	yes
11097	2	1	75.87	65.00	53.02	no	Disagree	Neutral	yes	no
11097	3	1	75.87	90.00	122.99	no	Strongly Agree	Strongly Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11097	4	3	122.99	110.00	79.86	no	Disagree	Disagree	yes	no
11097	5	3	122.99	90.00	105.32	yes	Neutral	Neutral	yes	no
11097	6	4	79.86	70.00	80.94	yes	Neutral	Agree	no	–
11097	7	1	75.87	95.00	126.47	yes	Agree	Strongly Agree	yes	yes
11097	8	–	–	90.00	88.38	–	Agree	Agree	yes	yes
11103	2	1	75.87	50.00	53.02	no	Neutral	Neutral	yes	no
11103	3	2	53.02	120.00	122.99	no	Neutral	Neutral	yes	no
11103	4	3	122.99	81.00	79.86	no	Disagree	Neutral	yes	no
11103	5	4	79.86	53.00	105.32	yes	Strongly Agree	Strongly Agree	yes	yes
11103	6	5	105.32	70.00	80.94	no	Disagree	Disagree	yes	no
11103	7	6	80.94	106.00	126.47	no	Neutral	Neutral	yes	no
11103	8	6	80.94	97.00	88.38	no	Neutral	Disagree	yes	no
11112	2	1	75.87	50.60	53.02	no	Disagree	Disagree	yes	no
11112	3	1	75.87	115.00	122.99	no	Neutral	Strongly Disagree	yes	no
11112	4	1	75.87	120.00	79.86	no	Strongly Agree	Strongly Agree	yes	yes
11112	5	3	122.99	192.00	105.32	yes	Agree	Agree	yes	yes
11112	6	–	–	90.00	80.94	–	Neutral	Strongly Disagree	yes	no
11112	7	–	–	90.00	126.47	–	Neutral	Strongly Disagree	yes	no
11112	8	–	–	90.00	88.38	–	Disagree	Disagree	yes	no
11123	2	1	75.87	70.00	53.02	no	Agree	Strongly Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11123	3	1,2	–	83.00	122.99	no	Strongly Agree	Strongly Agree	yes	yes
11123	4	3	122.99	101.00	79.86	no	Agree	Strongly Agree	yes	yes
11123	5	1	75.87	70.00	105.32	yes	Strongly Agree	Strongly Agree	yes	yes
11123	6	5	105.32	82.10	80.94	no	Agree	Strongly Agree	yes	yes
11123	7	–	–	51.40	126.47	–	Strongly Agree	Strongly Agree	yes	yes
11123	8	–	–	125.90	88.38	–	Strongly Agree	Strongly Agree	yes	yes
11130	2	1	75.87	62.00	53.02	no	Strongly Agree	Strongly Agree	yes	yes
11130	3	1	75.87	90.00	122.99	no	Agree	Strongly Agree	yes	yes
11130	4	1	75.87	84.00	79.86	no	Agree	Neutral	no	–
11130	5	1	75.87	84.00	105.32	yes	Strongly Agree	Strongly Agree	yes	yes
11130	6	1	75.87	60.00	80.94	yes	Agree	Neutral	no	–
11130	7	1	75.87	86.00	126.47	yes	Agree	Strongly Agree	yes	yes
11130	8	1	75.87	60.00	88.38	yes	Strongly Agree	Strongly Agree	yes	yes
11138	2	1	75.87	43.60	53.02	no	Disagree	Disagree	yes	no
11138	3	1	75.87	108.80	122.99	no	Agree	Strongly Disagree	no	–
11138	4	3	122.99	95.20	79.86	no	Agree	Agree	yes	yes
11138	5	4	79.86	65.00	105.32	yes	Strongly Agree	Strongly Agree	yes	yes
11138	6	5	105.32	60.00	80.94	no	Agree	Disagree	no	–
11138	7	6	80.94	86.00	126.47	no	Agree	Neutral	no	–
11138	8	6	80.94	80.00	88.38	no	Disagree	Neutral	yes	no

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11174	2	1	75.87	60.00	53.02	no	Disagree	Disagree	yes	no
11174	3	1	75.87	90.00	122.99	no	Agree	Strongly Agree	yes	yes
11174	4	3	122.99	110.00	79.86	no	Agree	Strongly Agree	yes	yes
11174	5	1	75.87	85.00	105.32	yes	Agree	Strongly Agree	yes	yes
11174	6	1	75.87	70.00	80.94	yes	Neutral	Neutral	yes	no
11174	7	1	75.87	85.00	126.47	yes	Neutral	Disagree	yes	no
11174	8	6,7	–	100.00	88.38	no	Neutral	Neutral	yes	no
11199	2	1	75.87	50.00	53.02	no	Neutral	Neutral	yes	no
11199	3	1	75.87	104.00	122.99	no	Neutral	Disagree	yes	no
11199	4	1	75.87	75.00	79.86	no	Agree	Strongly Agree	yes	yes
11199	5	4	79.86	90.00	105.32	yes	Agree	Strongly Agree	yes	yes
11199	6	4	79.86	70.00	80.94	yes	Agree	Strongly Agree	yes	yes
11199	7	1	75.87	95.00	126.47	yes	Strongly Agree	Strongly Agree	yes	yes
11199	8	4	79.86	50.00	88.38	yes	Agree	Agree	yes	yes
11207	2	1	75.87	66.00	53.02	no	Neutral	Neutral	yes	no
11207	3	1	–	78.00	122.99	no	Agree	Agree	yes	yes
11207	4	1	–	72.00	79.86	no	Disagree	Strongly Disagree	yes	no
11207	5	1	–	66.00	105.32	no	Strongly Agree	Neutral	no	–
11207	6	1	–	72.00	80.94	no	Disagree	Disagree	yes	no
11207	7	1	–	72.00	126.47	no	Strongly Disagree	Disagree	yes	no

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11207	8	1	–	72.00	88.38	no	Neutral	Disagree	yes	no
11257	2	1	75.87	75.00	53.02	no	Agree	Strongly Agree	yes	yes
11257	3	<3	–	85.00	122.99	no	Agree	Strongly Agree	yes	yes
11257	4	3	122.99	90.00	79.86	no	Neutral	Neutral	yes	no
11257	5	3	122.99	105.00	105.32	yes	Agree	Strongly Agree	yes	yes
11257	6	5	105.32	90.00	80.94	no	Disagree	Strongly Disagree	yes	no
11257	7	6	80.94	115.00	126.47	no	Neutral	Disagree	yes	no
11257	8	7	126.46	105.00	88.38	no	Agree	Neutral	no	–
11294	2	1	75.87	50.00	53.02	no	Neutral	Disagree	yes	no
11294	3	<3	–	70.00	122.99	no	Strongly Agree	Strongly Agree	yes	yes
11294	4	<4	–	77.00	79.86	no	Neutral	Neutral	yes	no
11294	5	3	122.99	80.00	105.32	yes	Agree	Strongly Agree	yes	yes
11294	6	5	105.32	63.00	80.94	no	Strongly Agree	Strongly Agree	yes	yes
11294	7	3	122.99	60.00	126.47	yes	Strongly Agree	Strongly Agree	yes	yes
11294	8	–	–	92.00	88.38	–	Neutral	Neutral	yes	no
11363	2	1	75.87	62.00	53.02	no	Neutral	Disagree	yes	no
11363	3	2	53.02	130.00	122.99	no	Neutral	Neutral	yes	no
11363	4	3	122.99	77.00	79.86	no	Disagree	Disagree	yes	no
11363	5	3	122.99	60.00	105.32	yes	Strongly Agree	Strongly Agree	yes	yes
11363	6	5	105.32	70.00	80.94	no	Agree	Neutral	no	–

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11363	7	3	122.99	110.00	126.47	yes	Neutral	Neutral	yes	no
11363	8	7	126.47	85.00	88.38	no	Neutral	Disagree	yes	no
11499	2	1	75.87	50.10	53.02	no	Neutral	Disagree	yes	no
11499	3	1	75.87	85.00	122.99	no	Neutral	Disagree	yes	no
11499	4	3	122.99	81.20	79.86	no	Disagree	Neutral	yes	no
11499	5	1	75.87	80.00	105.32	yes	Neutral	Strongly Disagree	yes	no
11499	6	5	105.32	69.30	80.94	no	Neutral	Disagree	yes	no
11499	7	1	75.87	100.00	126.47	yes	Neutral	Disagree	yes	no
11499	8	7	126.47	83.20	88.38	no	Neutral	Disagree	yes	no
11539	2	1	75.87	65.00	53.02	no	Disagree	Neutral	yes	no
11539	3	<3	–	82.00	122.99	no	Strongly Agree	Strongly Agree	yes	yes
11539	4	3	122.99	90.00	79.86	no	Disagree	Disagree	yes	no
11539	5	3	122.99	95.00	105.32	yes	Neutral	Neutral	yes	no
11539	6	5	105.32	85.00	80.94	no	Neutral	Disagree	yes	no
11539	7	3	122.99	100.00	126.47	yes	Neutral	Agree	no	–
11539	8	2	53.02	65.00	88.38	yes	Neutral	Neutral	yes	no
11572	2	1	75.87	50.00	53.02	no	Neutral	Neutral	yes	no
11572	3	1	75.87	88.00	122.99	no	Neutral	Neutral	yes	no
11572	4	1	75.87	70.00	79.86	no	Agree	Agree	yes	yes
11572	5	3	122.99	115.00	105.32	yes	Neutral	Neutral	yes	no

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11572	6	5	105.32	95.00	80.94	no	Neutral	Disagree	yes	no
11572	7	6	80.94	85.00	126.47	no	Agree	Strongly Agree	yes	yes
11572	8	7	126.47	115.00	88.38	no	Agree	Strongly Agree	yes	yes
11576	2	1	75.87	60.00	53.02	no	Agree	Strongly Agree	yes	yes
11576	3	1	75.87	80.00	122.99	no	Strongly Agree	Strongly Agree	yes	yes
11576	4	3	122.99	95.00	79.86	no	Neutral	Neutral	yes	no
11576	5	3	122.99	100.00	105.32	yes	Disagree	Disagree	yes	no
11576	6	5	105.32	75.00	80.94	no	Neutral	Neutral	yes	no
11576	7	5	105.32	95.00	126.47	no	Agree	Disagree	no	–
11576	8	7	126.47	75.00	88.38	no	Neutral	Disagree	yes	no
11588	2	1	75.87	60.70	53.02	no	Agree	Neutral	no	–
11588	3	1	75.87	126.00	122.99	no	Agree	Agree	yes	yes
11588	4	3	122.99	82.00	79.86	no	Disagree	Neutral	yes	no
11588	5	3	122.99	109.00	105.32	yes	Neutral	Disagree	yes	no
11588	6	5	105.32	70.00	80.94	no	Agree	Neutral	no	–
11588	7	3	122.99	82.00	126.47	yes	Strongly Agree	Strongly Agree	yes	yes
11588	8	7	126.47	84.00	88.38	no	Disagree	Strongly Disagree	yes	no
11711	2	1	75.87	70.00	53.02	no	Strongly Agree	Strongly Agree	yes	yes
11711	3	1	75.87	105.00	122.99	no	Neutral	Neutral	yes	no
11711	4	1	75.87	125.00	79.86	no	Strongly Agree	Strongly Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11711	5	1	75.87	105.00	105.32	yes	Neutral	Agree	no	–
11711	6	1	75.87	92.00	80.94	yes	Disagree	Strongly Disagree	yes	no
11711	7	1	75.87	85.00	126.47	yes	Agree	Neutral	no	–
11711	8	1	75.87	110.00	88.38	yes	Disagree	Disagree	yes	no
11881	2	1	75.87	55.00	53.02	no	Agree	Disagree	no	–
11881	3	1	75.87	85.00	122.99	no	Agree	Neutral	no	–
11881	4	1	75.87	87.00	79.86	no	Agree	Disagree	no	–
11881	5	3	122.99	74.00	105.32	yes	Neutral	Strongly Agree	no	–
11881	6	4	79.86	75.00	80.94	yes	Agree	Strongly Agree	yes	yes
11881	7	5	105.32	85.00	126.47	no	Agree	Agree	yes	yes
11881	8	7	126.47	90.00	88.38	no	Disagree	Neutral	yes	no
11900	2	1	75.87	58.00	53.02	no	Disagree	Neutral	yes	no
11900	3	<3	–	125.00	122.99	no	Neutral	Neutral	yes	no
11900	4	3	122.99	85.00	79.86	no	Disagree	Disagree	yes	no
11900	5	3	122.99	97.00	105.32	yes	Disagree	Disagree	yes	no
11900	6	5	105.32	75.00	80.94	no	Neutral	Neutral	yes	no
11900	7	1	75.87	85.00	126.47	no	Agree	Strongly Agree	yes	yes
11900	8	7	126.47	88.00	88.38	no	Neutral	Agree	no	–
11901	2	1	75.87	48.00	53.02	no	Neutral	Disagree	yes	no
11901	3	1	75.87	105.00	122.99	no	Neutral	Strongly Agree	no	–

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11901	4	3	122.99	85.00	79.86	no	Agree	Agree	yes	yes
11901	5	1,3	–	102.00	105.32	yes	Neutral	Neutral	yes	no
11901	6	2	53.02	75.00	80.94	yes	Agree	Strongly Agree	yes	yes
11901	7	1	75.87	89.00	126.47	yes	Strongly Agree	Strongly Agree	yes	yes
11901	8	7	126.47	85.00	88.38	no	Strongly Disagree	Neutral	yes	no
11972	2	1	75.87	75.40	53.02	no	Neutral	Neutral	yes	no
11972	3	<3	–	98.00	122.99	no	Neutral	Neutral	yes	no
11972	4	<4	–	100.00	79.86	no	Neutral	Neutral	yes	no
11972	5	<5	–	100.00	105.32	yes	Neutral	Agree	no	–
11972	6	–	–	90.00	80.94	–	Agree	Disagree	no	–
11972	7	–	–	95.00	126.47	–	Agree	Strongly Agree	yes	yes
11972	8	–	–	100.00	88.38	–	Neutral	Strongly Disagree	yes	no
11975	2	1	75.87	70.00	53.02	no	Agree	Strongly Agree	yes	yes
11975	3	1	75.87	75.00	122.99	no	Agree	Agree	yes	yes
11975	4	2	53.02	60.00	79.86	no	Neutral	Neutral	yes	no
11975	5	1	75.87	70.00	105.32	yes	Disagree	Neutral	yes	no
11975	6	1	75.87	70.00	80.94	yes	Neutral	Disagree	yes	no
11975	7	1	75.87	85.00	126.47	yes	Disagree	Strongly Agree	no	–
11975	8	1	75.87	70.00	88.38	yes	Neutral	Disagree	yes	no
11979	2	1	75.87	70.30	53.02	no	Agree	Strongly Agree	yes	yes

Table C.3: Raw Data from Distance Predictions and Surprise Ratings (continued)

Subject	Trial	Reference	Reference	Predicted	Simulation	Anomaly	Surprise Ratings			
ID	Number	Trial(s)	Distance	Distance	Distance	Elicited?	Rater A	Rater B	Agree?	Surprised?
11979	3	2	53.02	65.00	122.99	no	Strongly Agree	Strongly Agree	yes	yes
11979	4	3	122.99	107.00	79.86	no	Strongly Agree	Strongly Agree	yes	yes
11979	5	4	79.86	90.00	105.32	yes	Strongly Disagree	Neutral	yes	no
11979	6	<5	–	87.00	80.94	yes	Agree	Agree	yes	yes
11979	7	–	–	130.00	126.47	–	Agree	Strongly Agree	yes	yes
11979	8	–	–	102.00	88.38	–	Neutral	Strongly Agree	no	–

Table C.4: Raw Data from Debriefing Questioning

(a) aOFAT

Subject ID	How was the flaw identified? Unprompted	Prompted	It wasn't
01039		✓	
01042		✓	
01083	✓		
01104	✓		
01153		✓	
01194		✓	
01225	✓		
01231	✓		
01247			✓
01295	✓		
01317	✓		
01335	✓		
01353		✓	
01391	✓		
01488		✓	
01596		✓	
01757	✓		
01785		✓	
01791		✓	
01794	✓		
01797	✓		
01820			✓
01858		✓	
01893	✓		
01905	✓		
01957			✓
01988	✓		
Total	14	10	3

(b) PB-L_8

Subject ID	How was the flaw identified? Unprompted	Prompted	It wasn't
11018			✓
11065			✓
11070			✓
11094		✓	
11097			✓
11112			✓
11123			✓
11130		✓	
11138			✓
11174			✓
11199			✓
11207			✓
11257			✓
11294			✓
11363	✓		
11499			✓
11539			✓
11572			✓
11576			✓
11588			✓
11711			✓
11881			✓
11900			✓
11901			✓
11972			✓
11975		✓	
11979			✓
Total	1	3	23

This page intentionally left blank.

Human Detection of Computer Simulation Mistakes
in Engineering Experiments

Designed and typeset by the author using the LaTeX typesetting system.

Composed in *Minion Pro*, the OpenType version of the serif typeface *Minion* designed by Robert Slimbach for Adobe Systems in 1989, and *Myriad Pro*, the OpenType version of the sans-serif typeface *Myriad* designed by Robert Slimbach and Carol Twombly for Adobe Systems in 1992.

All figures without axes were created by the author using PGF/Ti*k*Z, a graphics system for use within LaTeX created and maintained by Till Tantau of the University of Lübeck.

All figures with axes were created by the author using PGFPlots, a plotting package for use within LaTeX created and maintained by Christian Feuersänger of the University of Bonn.