ASSESSING COGNITIVE WORKLOAD FROM MULTIPLE PHYSIOLOGICAL MEASURES USING WAVELETS AND MACHINE LEARNING

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

with a

Major in Neuroscience

in the

College of Graduate Studies

University of Idaho

by

Roger Lew

December 2013

Major Professor: Brian P. Dyre, Ph.D.

Authorization to Submit Dissertation

This dissertation of Roger Lew, submitted for the degree of Doctor of Philosophy (Ph.D.) with a major in Neuroscience and titled "ASSESSING COGNITIVE WORKLOAD FROM MULTIPLE PHYSIOLOGICAL MEASURES USING WAVELETS AND MACHINE LEARNING," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor _____Date_____
Dr. Brian P. Dyre

Committee Members _____Date_____
Dr. Richard B. Wells

_____Date_____
Dr. Terence Soule

_____Date_____
Dr. Steffen Werner

Department Administrator _____Date_____
Dr. Terence Soule

Final Approval and Acceptance by the College of Graduate Studies

_____Date_____
Dr. Jie Chen

Abstract

The overall safety and reliability of critical systems may be improved if interfaces can be tailored to the current cognitive states of their operators. For this to be realized, online measures of cognitive workload need to be developed. This dissertation proposes that cognitive measures based on physiological indicators provides the most potential in real world environments where task performance is difficult to quantify and operators may not be able to periodically self-report their workload. Here, the primary aim is the development and evaluation of algorithms for identifying cognitive workload from multiple relatively unobtrusive physiological measures using wavelet decomposition and machine learning. To support his primary aim, a tracking task was developed that allowed workload difficulty to be subtly and continuously manipulated in a systematic fashion. This manipulation was validated against subjective ratings of workload as well as with a secondary random number generation task. After establishing a means of controlling workload difficulty pupil diameter, skin conductance, heart rate, and heart rate variability were recorded and used in conjunction with machine learning to build classifiers of workload difficulty. Applying discrete wavelet decomposition to physiological measures and training classifiers to specific individuals yielded algorithms that could classify workload difficulty and derivative workload difficulty with enough accuracy to support practical applications (> 90% accuracies).

Acknowledgements

I would like to begin by gratefully and sincerely thank my committee members without whom this line of research would have never materialized. First I would like to thank my dear friend and mentor Brian Dyre for his continued support and unwavering idealism. Next, I would like to thank Steffen Werner and his critical eye and innate ability to keep an eye on the *big picture.* Next, I would like to thank Terry Soule for introducing me to the dark arts of artificial intelligence and genetic programming. Lastly, I would like to thank Rick Wells for his sage wisdom and rigorous tutelage.

Dedication

To my loving and patient wife, Tara Lew.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AC	Alternating Current
ALPS-SS	Age-Layered Population – Steady State
ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
BCI	Brain Computer Interface
CTT	Critical Tracking Task
CV	Coefficient of Variation
CWT	Continous Wavelet Transform
DARPA	Defense Advanced Research Projects Agency
dB	Decibel
DC	Direct Current
DFT	Discrete Fourier Transform
DTFT	Discrete Time Fourier Transform
DURESS	DUal REservoir System Simulation
DWT	Discrete Wavelet Transform
ECoG	Electrocorticography
EEG	Electroencephalography
ERP	Event Related Potential
FFT	Fast Fourier Transform
fMRI	Functional Magnetic Resonance Imaging
fNIR	Functional Near Infrared Imaging
GP	Genetic Programming
GSR	Galvanic Skin Resistance
HR	Heart Rate
HRV	Heart Rate Variability
Hz	Hertz
ICA	Index of Cognitive Activity
IPP	Implantable Prosthetic Processors
LDA	Linear Discriminate Analysis
MRT	Multiple Resource Theory
NASA-TLX	National Aeronautics and Space Administration - Task Load Index
NHTSA	National Highway Traffic Safety Administration
OFC	Orbital Frontal Cortex
PCA	Principal Component Analysis
PSD	Power Spectral Density
PD	Pupil Diameter
RMS	Root Mean Squared
RMSE	Root Mean Squared Error
RNG	Random Number Generation
SC	Skin Conductance
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SD	Standard Deviation
SNR	Signal-Noise-Ratio
STFT	Short Time Fourier Transform
SWAT	Subjective Workload Assessment Technique
SVM	Support Vector Machines
TCD	Transcranial Doppler Sonography
VMPFC	Ventromedial Prefrontal Cortex
WWI	World War I

Chapter 1: Introduction and Background

When human decisions affect the operation of critical systems, such as nuclear reactors, a small but non-trivial potential for disaster exists. Such disasters are often termed *low-probability high-consequence events* (Ellingwood & Wen, 2005). In our modern society the potential benefits of critical systems are arguably deemed to outweigh the associated risks (Komiyama & Kraines, 2008). Nuclear power increases our autonomy and reduces our carbon-footprint, space exploration expands our knowledge. However, tragic incidents like the collapse of the Tacoma Narrows Bridge, the 2010 British Petroleum oil spill, the Fukushima Daiichi nuclear incident, and the explosion of the space shuttle Challenger demonstrate that extreme caution must be taken at every step in the design, operation, and maintenance of critical systems. The repercussions of such failures have immediate and enduring consequences.

The recent Fukushima Daiichi nuclear incident in Japan is on par with the Chernobyl incident of 1986 that littered 6.7 metric tons of radioactive material over 200,000 square kilometers. Twenty-five years after the incident there are still over 3,500 known radioactive hotspots. The remediation is expected to continue for at least another 50 years (Peplow, 2011). Unfortunately the problem is complicated by the fact that radioactive material can be moved by wind and precipitation. While diffusion can help, weather patterns often result in dangerously high levels of radiation in isolated regions relatively far from the sites of origin. Detecting contamination is complicated by the fact that the material is invisible to the naked eye and odorless. Material that finds its way into the soil must also be treated quickly to avoid ground water contamination. These contaminated areas surrounding Chernobyl have documented a ten-fold increase in thyroidal cancer rates (Peplow, 2011). Such nuclear incidents also have long-lasting psychological effects. To this day managing exposure to radiation is a normal affair for Belarusian families. Children are socialized at an early age to the concept of radiation and their exposure is carefully monitored in a

manner similar to how one would keep track of a child's height and weight. As outsiders, we can only imagine the added stress living in such an environment would present.

Unfortunately, current global energy consumption begets increasingly complex technologies needed to obtain and use natural resources. For instance, if oil consumption continues unimpeded oil reserves will almost certainly be depleted by the end of the 21st century. By 2050 the global population is expected to reach 9 billion. Because of this population growth and the combined effect of globalization, fossil fuel consumption is as high as ever (Komiyama & Kraines, 2008). Forty years ago finding and obtaining oil was relatively easy. As these ancient reserves are depleted the logistics of finding and safely obtaining the oil become much more difficult. Deep sea drilling is one consequence of this fact. Incidents like the Deepwater Horizon oil spill, which has been attributed to lax adherence to safety standards on the part of British Petroleum, Halliburton, and Transocean, show that complacency is not an option (Drilling, 2011).

Some may wonder if the benefits from technologies like deep sea drilling and nuclear power are worth the risks. From the dawn of civilization humans have used technology, and from that time technological innovation has been a constant unyielding force. Human innovation is derived from our intelligence and is as essential to being human as walking on two feet and possessing opposable thumbs. Yet the context of our current situation leads one to wonder about human life at the turn of the 22nd century. Komiyama and Kraines (2008, p. VIII) put it thusly:

> It cannot be denied that the twin titans of science and technology have given human beings the potential to destroy ourselves. But if we develop science and technology wisely, we can use them to create a sustainable environment supporting a comfortable lifestyle in a clean and beautiful planet that humanity can enjoy for generations to come. Therefore, we need to make the correct choices concerning the direction of technology, and these choices can be made and implemented only through the consensus of society. There has never been a time when a good relationship between society and technology has been more important.

Perhaps the key insight Komiyama and Kraines are making is that technology should be viewed with ambivalence. Technology in and of itself is not *good* or *bad*. A second lesson to be learned from our history of technology is that human and machine must function synergistically.

Often, when one hears of *complex technology* the visceral reaction is negative. Every participant of the digital era knows how infuriating technology can be. Norman (2011) argues that it is not complexity that causes problems, but rather the frustration and complications caused by technology. Norman points out that in many regards humans have a natural attraction for observing and understanding complexity. Science, visual art, music and sports may all be products of this drive. In Norman's framework simplicity is not diametrically opposed to complexity. Complexity describes the state of the world; simplicity describes a state of the mind. Complexity reflects the intricacies between the operations of a system. Confusion and frustration result when humans are not able to form a mental model of a system. Confusion and frustration should be viewed as *the enemy,* not complexity. Increasing the complexity of a system may even serve to alleviate confusion and frustration at the cost of making it more difficult to engineer.

The law of conservation of complexity, or Tesler's law (Tesler & Saffer., 2007), states that "Every application must have an inherent amount of irreducible complexity. The only question is who will have to deal with it." For example early motorcycles required not only familiarity with operation of an internal combustion engine, but with the particular machine as well. To start the engine one must first make sure the motorcycle is in neutral, set the throttle position, set the choke, then manually retard the timing. Next the rider must use the kickstarter to position the piston at the beginning of its intake stroke by listening for the tell-tale sound of air escaping from the exhaust valve. Finally the rider could attempt to kickstart the engine and carefully close the choke and advance the timing as the machine warmed up to operating temperature. To make matters even more confusing the appropriate settings for the throttle, choke, and timing depended on the ambient temperature as well as the operating temperature of the machine. Poor selection would result in the machine not starting with the possibility of the engine backfiring and thrusting the kickstarter directly into the operator's shin. Crawford (2009) describes these early motorcycles as "more convenient than a horse, but surely not by much." Really early machines even required using

a hand operated pump to intermittently lubricate the inner workings of the crankcase. Neglecting to do so would destroy the engine. In contrast modern motorcycles and automobiles have automated starting systems that check to make sure the vehicle is out of gear (in park, or at least the clutch is disengaged), adjust the amount of fuel and air entering the engine, precisely control the ignition of the air and fuel, and even turn over the engine.

Today, talk of such systems actually seems mudane, and this is the point. Vehicles today are much more complex than vehicles of the past. These new vehicles all come with vacuum systems, electronic ignition systems, electronic fuel injection systems, mechanical pumps, solenoids, compressors, electric motors, computers and the multitude of sensors that go along with them. A modern automobile for instance, may contain over 30 computers requiring as many as 100 million lines of code (Motavalli, 2010). Despite their increased complexity modern vehicles are safer and easier to operate from the perspective of drivers. First adopters of motorcycles with mechanical fuel pumps were likely hesitant to relinquish control to one of the most important aspects of the machines operation. At first they probably paid close attention to every nuance of the machine in an effort to evaluate the competency of the innovative device. To assuage rider anxieties oil pressure gauges would be installed to allow riders to validate the systems performance at a glance. Overtime riders would slowly begin to trust and even prefer mechanical oil pumps to the manual hand operated pumps of the past.

While additional system complexity makes systems as a whole become much more complex, appropriately implemented abstraction and hierarchical layering of technology can actually reduce the complexity that any single human has to manage. For example, a word processor end-user need not know how to develop applications, an application developer need not know the instruction set of the underlying hardware, an OS developer need not know how circuit gates function, and so on. Ideally, at each level of the hierarchy the complexity that any particular engineer, or team of engineers, need to deal with should be manageable even though the systems as a whole are much

more complex. A layer of abstraction may be dependent on the reliability of lower levels of abstraction, but theory and experience have shown that nearly defect-free technological systems can be implemented (Gertman & Blackman, 1994). Because technology is so reliable the largest source of unreliability is more often than not the human component.

Incorporating the human component into the design of a system is by far the greatest challenge. In 1988 the newly developed Airbus A320 was heralded as a revolution in aircraft automation and safety. The plane featured advanced cockpit instrumentation, fly-by-wire controls and a *flight-protection* system that modulated control inputs based on the aircrafts current speed and altitude as well as the aircraft's aerodynamic properties to prevent the plane from crashing or overstressing the airframe (Casey, 1998).

On June 26, 1988 a new A320, under control of Captain Michel Asseline, was scheduled to make a low-speed low-altitude flyover at an airshow in Mulhouse Habsheim, France. Shortly after takeoff the *flight-protection*, *autothrottle* and *alpha floor* systems were all intentionally disengaged, if they were left running the plane would never let the crew conduct the relatively risky maneuver. On the approach the crew had difficulty visually identifying the Mulhouse Habshiem airfield and consequently had to descend at a rather steep glideslope. To accommodate the descent the A320 engines were set to idle. As the aircraft loomed over the airfield it began to lose speed without the flight crew's recognition. When this fateful fact finally came to their attention the plane was 100 feet above the ground with an airspeed of only 132 knots. The captain quickly pitched the plane upward and engaged the throttle but the idle speed of the turbines delayed the onset of thrust. By the time the engines began to wind-up the plane was 30 feet above the ground and the massive twin turbines only served to engorge large amounts of debris before the plane burst into flames. All told, three passengers died, and 50 were injured. In the aftermath, the lack or planning, late identification of the airfield, pomp and circumstance of the event, as well as the presence of attractive female guests on the flight deck were deemed contributing factors to the final outcome.

The primary fault of the incident was placed primarily on the Captain's overconfidence in the flight performance of the A320. Casey (1998, pp. 104-105) puts it thusly:

> Instead of treating the flight protection system like the safety guardrail on a winding mountain road, he employed it to define the limits of aircraft flight. It was as if he used the guardrail to negotiate the curve, not treating it as a protective barrier placed there in the event that he lost control. He had turned the flight protection system off for the maneuver and thereby had to fly by the standard rules of flight – something for which he may not have been entirely prepared.

In hindsight, it is easy to identify the series of mistakes made by the flight crew in the final moments before the crash. However, during those moments it is also important to remember that the crew was overtasked and likely impaired by stress.

The AirBus incident demonstrates that developing human and machine systems entails more than just developing robust technological systems. The humans that operate those systems have their own set of constraints. Compared to computers, humans are prone to errors of logic, perceptual illusions, vigilance decrements, cognitive tunneling, along with dozens of other impairments. Despite our shortcomings humans are still vital to many aspects of operating critical systems due to their ability recognize patterns, diagnose faults, and quickly generate and implement remediations (Boring & Kelly, 2008). In critical systems human error is the most significant contributor to the reliability of a system (Gertman & Blackman, 1994). Critical systems can be made safer by understanding human error and by designing systems to better accommodate their human operators. The emerging field of augmented cognition proposes that human machine interaction can be improved by making the machines able to recognize the internal states of the operator and adapt the presentation of information or the internal processes of the system to better suit the needs of the operator as well as external demands or system limitations. Augmented cognition systems are still in early stages of development. To use information pertaining to the mental states of operators those mental states must first be assessed. The efforts of this dissertation

focus on how cognitive workload can be quantified from physiological measures. Once, mental states can be reliably and accurately measured they must be used effectively by the critical systems.

With complex and safety critical systems the human operators are paradoxically essential to operation, yet simultaneously the largest contributor to the overall reliability of the system (Boring & Kelly, 2008; Gertman & Blackman, 1994). By understanding human cognition and utilizing on-line metrics of cognitive workload to tailor in real-time the human machine interfaces, potential exists to increase the overall reliability and safety of processes where there is a potential for high-consequence low-probability events.

Humans sensory data processing is not only *bottom-up,* but also *top-down* as the processes depend on expectations set by prior experiences and mental representations formed through prior experience (Kahneman, 1973). These top-down mechanisms make humans highly adaptable (compared to their current silicon counterparts) and excel at recognizing complex patterns. Humans also have unique disadvantages. Humans are easily overburdened with too much information, are poor at multitasking, prone to lapses of logic, and can be severally affected by stress and fatigue. A more in-depth account of these short-comings can be found in Chapter 2 of this dissertation.

By incorporating cognition into machine human interfaces systems can maximize human advantages will minimizing human weaknesses. If systems can identify when there operators are overburdened they can adapt to compensate by reducing the workload of the operator. For example, an augmented cognition system for air traffic controllers could automatically hand over planes to other controllers when cognitive workload indicators suggest a controller is overburdened. Or a system may adapt the presentation of information to the specific circumstances. For example, in process control a single failure may produce a cascade of alarms. Operators must maintain awareness of critical operational states while attempting to diagnose the root cause. In situations where operators are overburdened, increasing the saliency of the most

probable causes of failure may aid operators. While this mitigation may help operators diagnose the critical problem, it may also lead to loss of overall situational awareness if operators neglect critical non-salient indicators. Thus, in many circumstances the best way for a machine to interact with the operator is dependent on not only the state of the system, but the state of the operator. It follows that in order to most effectively interact with operators a system needs to be aware of the cognitive workload of its operators.

Here I propose that physiological indicators of cognitive workload may be the best solution. There are several ways of measuring cognitive workload, but many are ill-suited to real-time complex tasks where performance criteria is hard to quantify and humans are unable to report or unaware of their cognitive state. Next I discuss physiological measurement technologies and why the focus here is on simple and inexpensive physiological technologies like pupil diameter, skin conductance, and heart rate variability. To process the physiological signals a variety of approaches have been attempted. Here I am interested in using wavelet decomposition and machine learning to increase the overall efficacy and reliability of physiologically based workload measures.

Previous research has established that autonomic nervous system activity elicits physiological responses in pupil diameter, skin conductance, and heart rate variability (see section 2.2.4 for a full review and citations). In my first attempt (Experiment 5.1) at assessing mental workload from physiological measures I was most concerned with maintaining ecological validity. To this end, I used a simple process control simulator (DURESS) where participants manipulated the flow of water through a network of pumps, valves, aquifers, and heaters. I attempted to manipulate task difficulty through component failures at specified times. Offline I examined short-time Fourier transforms (STFT) of skin conductance (SC) and pupil diameter (PD). This provided some indications that the spectra of physiological measures are correlated with workload (see Chapter 5.1). Unfortunately, the failures in the control plant often went unnoticed by participants which would consequentially have a null effect on workload.

In the second experiment (Experiment 5.2) I decided to forgo ecological validity and maximize the salience of workload changes by using a continuous dual axes pursuit tracking task. To manipulate difficulty the control mappings reversed at periodic intervals. To objectively compare the Fourier components principle component analysis (PCA) was employed. Principle component analysis is a method for reducing the dimensionality of data. PCA is a method of determining a new orthogonal basis for a set of data such that the variance is captured maximally by a reduced set of components (see Section 3.1.9). Out of the six participants only one showed statistically reliable differences between the normal and reversed control mapping conditions. Part of this null effect can be attributed to the unexpected ability of some of the participants to immediately adapt to the reversed mapping. The classification analysis was also constrained by the inherent fixed time frequency resolution of STFT and assumptions of linearity in the classification.

The third experiment (Experiment 5.3) followed in the direction of Marshall (2000; 2002; 2007) and used wavelet analysis to decompose the physiological measures in both time and frequency. In contrast to STFT, wavelet decomposition provides optimized temporal localization across the frequencies under examination (see Section 3.3.4 & 3.4). With DFT and short time DFT the frequency bins are linearly spaced. In contrast the frequency bands are logarithmically spaced with discrete wavelet decomposition. The logarithmic spacing is often a more natural mapping to phenomena. My effort differs from Marshall in how I classify the resulting wavelet components. Here I use a machine learning technique known as Genetic Programming (GP) to identify models which can relate the physiological signals to performance error and task difficulty.

In computer science genetic programming has been shown to be a versatile tool for complex problems with large multidimensional parameter spaces. In the augmented cognition domain it has seen surprisingly little use. Genetic Programming is an iterative machine learning technique that is well suited for optimizing non-linear problem spaces with high dimensionality (see Machine Learning). It accomplishes this feat by first generating a population of random solutions. Better

solutions are selected and recombined in the hope of generating even better solutions. At each iteration, the worst solutions are thrown away to maintain a constant population size from generation to generation. After hundreds and sometimes thousands of generations the solutions improve. This process of selecting and combining fit solutions essentially reduces and focuses the problem space the algorithm must search through. Because the search is not comprehensive there is no guarantee GP will find always find the optimal solution. However, in many of the domains which employ GP performing a comprehensive sweep of the problem space is not feasible and GP is capable of producing solutions comparable to with human experts and rivaling other machine learning techniques (Eiben & Smith, 2003).

We believe GP is well suited for augmented cognition because of its extensibility. GP can easily be applied to different physiological signals, multiple physiological signals, or non-physiological signals. GP does not make any intrinsic assumptions between how the physiological parameters relate to the underlying constructs. In sum, there are three important differences between the Preliminary Experiment 3 and previous work. First, this study uses both raw time series and wavelet analysis of multiple physiological measures (SC and PD) to predict performance and task workload, which I expect should be more predictive than PD signals alone because the redundant measures may better differentiate signal from noise. It is also likely that PD and SC carry non-redundant information that is related to workload as correlations between physiological measures are often low (Kahneman, 1973). Second, the wavelet components were estimated as continuous variables rather than processed into binary variables as described by Marshall. Third, this study compared two modeling approaches based on the raw time series data and wavelet-estimated power spectrum: a traditional linear regression/discriminate analysis approach and a genetic programming (GP) approach. The results from Experiment 3 suggests that wavelet decomposition is an effective means of extracting time frequency information from physiological signals and that non-linear classification algorithms are essential to utilize the time frequency

information inherent in these physiological signals. I have also shown that GP is capable of integrating information from pupil diameter (PD) and skin conductance (SC) into an assessment of workload during a discrete manual tracking task.

The incorporation of a verbal random number generation task provides evidence that both discrete and continuous tracking involve central executive processes (Experiments 5.4-5.7). This suggests that despite the lack of face validity to process control tasks, manual tracking tasks load some of the same cognitive resources involved with process control.

For a physiological measure to be useful to process control operations it must be predictive. A system that only has the ability to identify workload *after* task performance completely degrades is of limited practical utility. Previous experiments have attempted to classify workload after gross changes in task difficulty (Experiments 5.1-5.3). Abruptly changing task difficulty makes it easier to identify physiological changes but makes it difficult to form a predictive measure of workload. Experiment 5.4 switched to a compensatory (closed-loop) tracking task based off of the critical instability task (McDonnell & Jex, 1967; McRuer & Graham, 1965). This allowed for gradual changes in task difficulty. Task performance was found to be monotonically linked to workload. Results suggest that predictive measures of workload and subsequently task performance can be formed from physiological measures. Random forests were able to classify both the magnitude and derivative of a workload input signal with significant accuracy (> 90% in many cases).

Contrary to initial hypotheses the incorporation of additional physiological measures (heart rate variability, respiration) did not significantly improve the efficacy of workload estimates obtained via machine learning techniques. Previous evidence has observed physiological measures often have low correlations with one another (Kahneman, 1973). This suggests that they may provide non-redundant information pertaining to mental workload. Similar to previous studies, experiment 5.8 found only moderate (0.2-0.5) and statistically reliably correlations between HRV,

PD and SC. Despite the absence of strong correlations, suggesting linearly non-redundant information, including more variables did not reliably improve the efficacy of workload estimates.

Taken together, this body of evidence suggests that pupil diameter and skin conductance signals carry information relevant to cognitive workload. Information relevant to cognitive workload can be extracted in a timely fashion and could be potentially in mitigating the potential of low-probability high-consequence events.

Chapter 2: Neuropsychology of Cognitive Workload

2.1 Theoretical Foundations

The key to developing synergistic human and machine interaction is in understanding human cognitive processes and limitations, and then designing machine systems to work synergistically with those strengths and weaknesses. However, this is easier said than done because despite decades of research a unified construct of *cognitive workload* is still elusive (Moray, 1988). On the surface the concept seems relatively straight forward. Cognitive workload should simply convey the cognitive effort a person devotes to a task or tasks. Below the surface is where complexities arise. With physical tasks the problem of calculating physical workload has been solved long ago. One just needs to calculate the force required to move an object with a known mass over a given distance from point *A* to point *B*. In the psychological domain quantifying the cognitive equivalent has been more problematic. If we examine the myth of Sisyphus it should be clear that the torturous element is not the physical strain of moving a large boulder up a hill, but rather the monotony and frustration caused by watching it roll back down and having to push it up again for all eternity. How can one to quantify such anguish, and how can one design tasks antithetical to the task imposed on Sisyphus? These are the challenges presented to cognitive scientists and human factors practitioners. The field of Augmented Cognition proposes that if cognitive states can be identified, then systems can adapt to better suit an operator by off-loading tasks or altering how information is presented. The crucial component becomes defining and identifying cognitive workload.

To be practically applicable, a measure of cognitive workload should ideally be highly diagnostic, sensitive, and reliable. However, it should be noted, that in many circumstances even crude measures can be viable. For example, several automotive manufacturers have developed systems that monitor driver drowsiness and trigger auditory and verbal alerts when drivers become fatigued (Barr, Howarth, Popkin, & Carroll, 2005; Barr, Popkin, & Howarth, 2009).

Identifying whether a person is awake or nearly asleep is an extremely insensitive measure of alertness, yet such systems could substantially reduce the over 50,000 annual crashes due to drowsiness/fatigue (NHTSA, 1997). While performing mentally engaging tasks our autonomic nervous system is constantly regulating our generalized arousal. Evolutionary psychologists Cosmides and Tooby (1997, p. 85) argue that "our modern skulls house a stone age mind" and our actions are still highly influenced by *instinctive* tendencies. Even when we face modern problems our innate responses are formed by rather primitive autonomic midbrain activity. Furthermore, most of the processing which occurs in the mind is hidden from our explicit awareness. This is why gross changes, like the fact that we might be falling asleep, may go unnoticed. Human and machine interfaces can be in improved if cognitive workload can be reliably assessed.

There are three basic approaches to assessing cognitive workload. Workload can be measured subjectively through self-reports, but in practical settings, having workers periodically report their workload is potentially distracting and therefore ill-suited. The second approach is to measure a worker's performance. In practical settings this is problematic because performance is often multidimensional and difficult to define. Secondly, performance may be monotonically related to difficulty, but may not be sufficiently sensitive if operators are able to maintain high performance until they are completely overwhelmed. Lastly, workload can be assessed using physiological signals. Stress, fatigue, drowsiness, and cognitive processing all have underlying physiological mechanisms. These mechanisms leave their signatures in easily monitored physiological signals like skin conductance and pupil diameter. The problem is that those signatures are embedded in layers of noise from other physiological mechanisms that may or may not be related to cognitive workload. Developing measures of cognitive workload based on physiological indicators requires developing a theoretical understanding of underlying components (frustration, time pressure, fear, anguish, etc.) as well as a theoretical understanding of the mechanisms that implement cognitive workload. Traditionally the study of cognitive science is a study of functionalism. Functionalism, as

a philosophy of mind, is not overly concerned with explaining psychological phenomena mechanistically. Functionalism describes a phenomena in terms of abstract psychological constructs like memory and intelligence instead of describing the cause and effect interaction of physical components. The approach segregates the functions of the mind from the physical implementation of the functions. Prior to the cognitive revolution using such abstract psychological constructs was heavily discouraged. Chomsky's 1959 work is often cited as delivering the death blow to behavioralism (Chomsky, 1959). He argued that incorporating constructs could explain phenomena of verbal behavior that could not be easily explained otherwise. For example, a theory of language based purely on behavioralism cannot adequately explain how young children can understand and produce sentences they have never encountered. The use of constructs has undoubtedly increased the power and sophistication of psychological theories but it is important to not become entrenched in functionalism.

Functionalism can describe how a phenomenon operates but cannot address what produces the phenomenon. It is only in the last 20 years or so that neuroscience has *freed* cognitive science from functionalism (Parasuraman & Rizzo, 2008). The emerging field of cognitive neuroscience emphasizes both mechanistic and functional descriptions. It is not only important to functionally describe psychological phenomena, but to identify the underlying neural substrates responsible for those functions. On one side of the equation we have cognitive models with tentative constructs and tentative interactions amongst constructs. On the other side of the equation we would like to have a fine-grained neurophysiologic map of what is occurring in conjunction with every thought and action, but in reality we only have the *reflections* and *shadows* of neural activity provided by the various brain imaging technologies. Within neuroergonomics/Augmented Cognition the current approach to bridging the gap is to divide the problem based on the various technologies. Methods using EEG (electroencephalogram) fall into one sub-domain, while methods using transcranial Doppler sonography fall into another. As

discussed below EGG has good temporal resolution but poor spatial resolution. The signals recorded by EGG represent the activity of large groups of neurons. On the other hand, transcranial Doppler sonography measures blood flow velocity in vessels feeding the brain. It might be more apt to say that brain imaging technologies record artifacts of neural activity. Because all current technologies are all rather crude in at least one regard (see review) using a single measure is akin to reconciling the structure of a three dimensional object from a single perspective through an aberrated lens.

This problem is essentially epistemological. How can we know our representation of an object, the human brain, reflects its true nature, or is it even possible to know the true nature of an object based solely on the structural and physiological artifacts? Philosophy of science would suggest that we can't know the objects true nature but through experimentation we can at least come up with a model that approximates the pattern of empirically observed relationships. At best we can find links between tentative constructs and artifacts of neural activity. When both sides of the equation are treated as containing unknowns it is difficult to find anything worthwhile. Even though the construct of cognitive workload and models of cognitive processing are by no means definitive or well established the functional theories of workload are more advanced than the mechanistic. **For this reason it makes more sense to treat existing constructs as axiomatic within the experimental framework. This allows cognitive workload to be operationally defined by established tasks and validated using established behavioral measures and secondary tasks.** Of course, as any disciple of science knows, the absence of evidence is not the evidence of absence. Failing to find links between cognitive workload and artifacts of neural activity could mean the physiological measures were invalid, the constructs were incorrect, the techniques used to establish the links were not powerful enough, the effect size was too small to find reliable differences, or a combination of these factors. At this time, the best we can do is to find physiological correlates to psychological constructs.

Neuroscience has a decent understanding of how a few neurons communicate with one another, but the sheer interconnectedness of the 100 billion or so neurons in a typical human brain has left many of its intracies obscured. The current approach is slightly more sophisticated than using a multimeter to naively probe a broken transistor radio. With a bit of patience even someone with only a bit of experience with electronics might be able grossly segregate the components of the radio into functional groups: power supply, radio receiver, amplifier, etcetera. After these functional divisions have been identified one could begin diagnostically testing their operational model and perhaps fix the radio given the right tools, spare components, lots of patience, and a bit of luck.

Manipulating a radio's controls while monitoring the multimeter provide empirical evidence related to the mechanistic operation of the radio. The task of identifying cognitive states from physiological measures is not too different from this analogy. The tasks required that participants and the obtained physiological measurements relate to the mechanistic operation of human cognition. One only needs to define the possible inputs to the model and how the output of the model relates to the performance (fitness) of the model. Genetic programming simultaneously compares a population of hundreds of thousands of competing models selecting the ones that best fit the data. Genetic programming implemented in this manner may reveal functional homologs to internal processes. Genetic programming is discussed in detail in Chapter 4. This chapter continues by elaborating on the theoretical construct of cognitive workload (Section 2.2), the practical application domains of cognitive workload measures (Section 2.3), and state of physiological measurement technologies (Section 2.4).

2.2 Cognitive Workload Theory and Application

2.2.1 ***Defining Cognitive Workload.*** (Kahneman, 1973) describes workload as demand imposed on an operator and draws an analogy with an electrical circuit. In this analogy the operator is a power supply with finite capacity. The task is analogous to an electrical load presented to the power supply. With electrical circuits and human operators overload occurs when the load exceeds the capacity of the supply. In human terms overload means the task demands more cognitive resources than the operator can supply. Gopher and Donchin (1986) extend this analogy by pointing out that the current flowing through the load is an interaction between the supplied voltage and the impedance of the load as described by Ohm's law. One cannot determine the current flowing through the circuit without knowing both. In an analogous manner workload reflects an interaction between the operator and the task. Some operators may have the equivalent of higher supply voltage reducing the relative amount of effort required to perform a given task.

We can extend Kahneman's electrical analogy even further by introducing the concept of impedance. Impedance describes how resistance varies with respect to frequency in alternating current (AC) circuits. In AC circuits electrical loads are not purely resistive because they may present parallel capacitance and inductance to the supply. Capacitive loads present falling impedance with increasing frequency, and inductive loads present falling impedance with decreasing frequency. The resulting load is time varying and is described by the impedance curve of the load. In an analogous manner the load presented by a task cannot be fully described by a single scalar value. Tasks present varying demands on separate cognitive resources.

Wicken's multiple resource theory (MRT) expands upon work by Kantowitz and Knight (1976) and Navon and Gopher (1979). It conceptualizes cognitive resources along three dimensions. The first dimension divides resources into perception, processing, and response stages. The second divides resources into visual, auditory, tactile, and olfactory modalities. The third dimension divides resources into spatial or verbal codes (Wickens, 2002). Humans are capable of

simultaneously performing some tasks presented to separate modalities with virtually no decrements in performance compared to performing the tasks alone. This suggests that they are effectively utilizing independent resources. However, in other scenarios tasks presented to independent modalities do show reduced performance.

Driving an automobile requires visual perception, spatial processing, and motor responses, while talking on a cell phone requires auditory perception, verbal processing, and verbal responses. A theory based on multiple resource theory would predict that these two tasks should not interfere with one another. Contrary to MRT a growing corpus of research has shown that cell phone usage significantly degrades driving performance (Strayer & Johnston, 2001; Strayer, Drews, & Johnston, 2003; McCarley, Vais, Pringle, Kamer, Irwin, & Strayer, 2004; Just, Keller, & Cynkar, 2008). Just and Carpenter (1992; 1993) have developed a capacity theory of mental resources where psychologically defined concepts of executive functioning, visual processing and so on are linked to cortical regions. The measured neural activation in these regions signifies how the mental resources are utilized. Simplistically it can be conceptualized by assigning cortical regions to the mental resources outlined by MRT. Just, Keller, and Cynkar (2008) recorded fMRI contrasts comparing activation between performing a simulated driving task and simultaneously performing the simulated driving task along with an auditory comprehension task (listen to verbal statements and non-verbally report whether they are true or false). As one would expect language regions show increased activation when both tasks were performed. Surprisingly, fMRI revealed no differences in the amount of executive activation and showed decreased activation in spatial and visual regions when the participants were performing both tasks.

This suggests that even though the tasks theoretically load different resources a *central executive* might limit how these resources can be utilized when both tasks are presented. The central executive is a construct first proposed by Baddeley and Hitch (1974) and builds on work by Broadbent (1958) . The central executive is one of three subcomponents of a model of working

memory. The other two components are the phonological loop and the visuo-spatial sketchpad. The central executive serves to direct attention to salient sensations, interpret perceptual information from the various modalities and to coordinate actions. For the first 20 years research efforts focused primarily on understanding the phonological loop and the visuo-spatial sketchpad. The central executive was viewed as a homunculus of sorts. It was an abstraction that served a variety of functions but was not well understood (Baddeley, 1996). Baddeley contends that regardless of whether the central executive is a single coordinated system or a collection of largely autonomous processes the concept aids in developing a functional understanding of working memory. Furthermore, once all the functions of the homunculus are understood it is no longer a homunculus. The central executive is more of an umbrella to classify a set of problems. Many executive functions are carried out or dependent on the frontal lobes (Shallice, 1982; 1988) but the functional to anatomic mappings are not always straight-forward. A single dysfunction may be the result of damage to multiple anatomical regions. To add further perplexity, patients may have damage to the same anatomical region but show different functional deficients (Baddely & Wilson, 1988).

Unlike controlled laboratory experiments, real world work environments are almost always multimodal, have extraneous sources of noise, changing performance criteria as situations evolve, and shifting task concurrency. In the power supply metaphor these factors influence the load impedance presented to a power supply. Whether the supply can meet the demand depends on the load impedance as well as the source impedance of the supply. Ideally the source impedance should be much lower than the load impedance. If the load is held constant the power transfer becomes less and less efficient as the source impedance increases. The central executive could be regarded as a collection of functions that dynamically modulates the source impedance dependence on the output impedance and internal states. The underlying mechanisms of how the central executive accomplishes this are not entirely understood, but evidence suggests that multiple resource theory alone is insufficient to explain workload and performance with all concurrent tasks.

Now that the construct of cognitive workload has been introduced, it is appropriate to review how it has been assessed.

2.2.2 ***Behavioral measures of workload.*** Behavioral measures of workload refer to measures that utilize task performance. The previous method of quantifying driving errors while simultaneously talking on a cellphone is an example of a behavioral measure of workload. Behavioral measures can be subdivided into primary and secondary task measures. Primary task performance could be system errors, data entry speed, driving deviations, etc. They have the benefit of being straight forward and in many situations having high face validity. The downside to primary task measures is that in the "underload" region they may exhibit little variability (insensitive). In other words performance may be very good until the performer is overloaded at which point performance "falls off a cliff" (Kahneman, 1973).

Secondary task measures were designed to address the underload shortcoming of using primary task performance. When an operator performs two tasks simultaneously performance on the second task can be used to measure the residual capacity not used by the first (Kahneman, 1973; Navon & Gopher, 1979). One of the intriguing phenomena leading to multiple resource theory is that some concurrently presented tasks can share available resources with little or no interference (Wickens, 2002). However, even in controlled laboratory settings the use of secondary tasks can be complicated with how the performance goals are specified or with the particular tasks chosen (Moray, 1988). In many practical applications tasks are not explicitly defined, or many tasks may be performed simultaneously which makes using behavioral measures difficult if not impossible. For example, consider a grocery store cashier. One may assess performance by the number of times per minute the cashier can scan, or the number of transactions per hour a cashier can perform, but these measures may be inversely correlated to how pleasant a cashier is to the customers they serve.

2.2.3 ***Subjective measures of workload.*** Subjective measures most closely follow workload theory and regard workload as being multidimensional. They assess workload by presenting operators with self-report questionnaires. The National Aeronautics and Space Administration - Task Load Index (NASA-TLX) is a workload assessment tool that breaks workload into six subscales identified by Hart and Staveland (1988). These six dimensions are: mental demands, physical demands, temporal demands, own performance, effort, and frustration. These six dimensions are designed to account for individual differences. Some may not find a task loading if it does not have a physical component. Some individuals become frustrated easier than others (Moray, 1988). Likewise, the Subjective Workload Assessment Technique (SWAT) breaks workload into similar subscales of time load, mental effort, and stress. Subjective measures have the benefits of being easy to derive and good sensitivity to changes in workload.

Unlike behavioral and physiological measures subjective measures can be subject to bias and require the rater to report honestly. Other disadvantages of subjective measures are that "online" measures intrude on the operator's primary task and "offline" measures only provide workload information after the task has already been completed. An additional problem with subjective measures is they rely on introspection. Humans may be aware of what they are doing, but may not be aware of how they are doing it. Milner and Goodale (2006) proposed the two-stream hypothesis for visual processing. According to this hypothesis the ventral stream is responsible for "what" while the dorsal pathway is responsible for "how." Actions often run autonomously without cognitive intervention. The reader might have had the experience of looking through their rear view mirror to wonder if they just ran a red light, or perhaps found themselves parked in their driveway instead of at the grocery store. Because we are not aware of how we are doing something, we may not be able to accurately report workload.

2.2.4 ***Physiological measures of workload.*** Another factor to consider is how generalized arousal modulates the availability of cognitive resources. The Yerkes-Dodson Law (Yerkes &

Dodson, 1908) describes an inverted-U relationship between arousal and performance. Peak performance occurs at moderate levels of arousal. Low levels of arousal lead to boredom and frustration. High levels lead to anxiety and exhaustion. Optimum performance occurs in between these extremes. The optimal amount of arousal may differ from individual to individual. These stressors cause the autonomic nervous system to increase generalized arousal. In highly stressful situations an autonomic nervous system response physically and psychologically prepares one for action. This is colloquially known as the "fight-or-flight" response. The fight-or-flight response is elicited by the autonomic nervous system (ANS).

The ANS has two antagonistic sub-systems that work in concert with one another to unconsciously control a variety of physiological functions. The sympathetic sub-system activates blood flow to the muscles and heart. Simultaneously, the parasympathetic sub-system inhibits blood flow to internal organs and the gastrointestinal system. While this psychobiological response might be helpful for tasks of a physical nature, like running from a large predator, it can have negative repercussions when threats are non-physical (Lundberg, 1993). Previous research has shown that army recruits under high levels of stress showed decreased problem solving performance as well as decreased working memory (Capretta & Berkun, 1962; Berkun, 1964). Work by Porcelli and Delgado (2009) suggests that even acute stress induced risk reflection in financial decision making. Risk reflection is a phenomenon where choice criterions become more conservative when choosing between potential gains and more risky when choosing between potential losses. Decision making research has also shown that participants consider far less information and use strategies that more heavily weight negative information when under time pressure (Svenson & Maule, 1993). Several strategies can be used to combat poor decision making under stress.

Extensive training and/or experience can be helpful in developing automaticity for complex procedures. Experts are also better at identifying the most critical details to a problem and

generating working solutions (Klein, Orasanu, Calderwood, & Zsambok, 1993). However, experts are also described as overconfident in their abilities (Ericsson, Charness, Feltovich, & Hoffman, 2006). A second short coming of training and experience is that even the most extensive training is limited to a miniscule fraction of a multitude of possible faults that can occur within a complex system. Real world problems are not restricted to the limited set of training scenarios presented to human operators. To quote Murphy "what can go wrong, will go wrong." To overcome this shortcoming a second, albeit more complicated approach, is to develop technologies which can assess an operator's cognitive state and use that additional information to support operator decisions, optimize human performance, and reduce human error even during novel situations.

2.2.5 ***Conclusions.*** One might initially assume that behavioral, subjective and physiological measures presented here all reflect the same underlying construct of workload, but these measures often only weakly correlate with one another. For a skilled operator behavioral measures may change very little despite the operator's subjective and physiological workload ratings increasing. While behavioral, subjective, and physiological workload measures may all be monotonic the relationships between them are not necessarily linear (Kahneman, 1973). The most appropriate type of measure depends heavily on one's goals and contextual constraints.

In real workplace settings task goals are often not explicitly defined as they might change with situational circumstances. For example, in the early stages of a building fire the appropriate course of action might be to extinguish the fire. If the fire grows the appropriate course of action becomes evacuating occupants and to not wasting resources on containing the blaze. In dynamic real world settings defining the task, let alone task error, is not possible this makes using behavioral measures of workload infeasible. Subjective measures of workload would require operators be cognizant of their current workload being too great and hampering their decision making in order to make the decision to reduce their workload. This is problematic for the obvious reason. This leaves physiological measures of workload. For these reasons Augmented Cognition has placed a

strong emphasis on developing reliable instantaneous measures of human workload based on physiological measures (O'Neill, 2006). As a broad generalization, the most successful techniques thus far have employed measures which directly monitor brain activity, like EEG and fNIR (functional near-infrared imaging). Other simpler, less expensive, and less obtrusive physiological measures such as pupil diameter (PD), galvanic skin conductance (SC), and heart rate variability (HRV) have shown promise but have not yet shown the same reliability as EEG and fNIR. The intent of this study is to examine the efficacy of using multiple physiological measures (SC, PD, and HRV) to measure mental workload for augmented cognition systems. Within the field of augmented cognition finding reliable, sensitive, instantaneous, and diagnostic measures of mental workload is the primary hurdle.

A quality of measure of workload needs to satisfy several requirements. It should be reliable, or exhibit relatively low variability relative to its range. The measure should ideally be highly sensitive, or able to distinguish small differences in workload. The measure should be instantaneous, or able to identify changes in mental workload shortly after they change. Lastly, measures should be diagnostic, or capable of identifying the utilization of several underlying psychological constructs (attention, memory, etc.).

The ideal physiological indicator should be able to resolve not only gross changes in workload (rest vs. active), but fine changes as well. Physiological workload measures should also generalize between different types of tasks (physically demanding vs. mentally demanding), individuals, and sub-populations (novice vs. expert).

Developing cognitive workload estimates based on physiological measures has two primary technological challenges. The first is to develop technology that can record physiological data. The ideal device should have excellent spatial and temporal resolution, require minimal power, be robust to noise, unobtrusive (does not interfere with the task), mobile, and cost next to nothing. Of course the ideal device has yet to be realized. The most successful techniques have utilized

measures which directly monitor brain activity like continuous electroencephalography (EEG), event related potentials (ERPs) from EEG, and functional near-infrared imaging (fNIR) (St. John, Kobus, Morrison, & Schmorrow, 2004). Despite initial shortcomings research into less intrusive, less costly measures like pupil diameter and skin conductance continue. The second technological challenge is devising signal processing techniques to extract the requisite information from the raw signals. Advances in signal processing techniques are likely to increase the efficacy of pupil diameter and skin conductance.

2.3 Practical Uses of Physiological Signals in Human Machine Interfaces

While humans are not the only creatures that utilize tools, we are among a small set of creatures that do. Many argue that tool use is intrinsic to what it means to be human. From the dawn of human tool usage to the near present, what we conceptualize as tools are static and mechanical. A tool is a hammer, or a drill press, or a computer. The human interacts with the environment through the tool, but the relationship between the user and the tool is typically not viewed as reciprocal. If a machinist catches their glove in a drill press, the drill press is unaware and uncaring of the soft flesh that lies within. It is only recently that technology has begun developing systems and interfaces that can monitor users and utilize that information to control how the systems functions and interact with their users. For example, the SleepCycle alarm clock uses a phone's accelerometer to gauge a user's sleep patterns (typically ~90 minute periodic). The phone wakes them up when it senses they are in the lightest sleep phase (Gavon, 2010). Nest thermostat can learn the schedule of its occupants and adapt its programming to keep them comfortable, while saving energy when a space is unoccupied (Pogue, 2011). SawStop table saws can differentiate between wood and flesh and prevent serious injury by stopping their blades within 10 ms (Newsome, 2007). A cabinet manufacturer remarks "The accidents are usually caused by human

error, but this saw grants you forgiveness." These examples are simple, yet the value is clear. Accommodating technological systems can offer benefits in comfort, efficiency, and safety.

For years human factors engineering has touted that "a machine should be fit to the human and not human to machine." Traditionally this has meant designing machines so that they are easy to use. This is an essential first step, but untapped potential lies in taking this tenet of human factors more literally. Machines that can act as if they are aware of our behaviors, intentions, and emotional states will likely serve us better than technological automatons, acting on their own will.

In order for our technological counterparts to act as if they are aware of our cognitive states they must have data to base their awareness on. This is where physiological measures come into play. The mind and body are two sides of the same coin. While we don't fully understand the mind it is mechanistically linked to our physiology. Therefore, changes of the mind and physiological changes are one and the same.

At present, the implications of such technological vigilance are uncertain. How willing are humans to participate in being constantly monitored? Transhumanists think that in the future, humans may control machines by thought alone and machines may to present information to users via direct neural links (Hughes, 2004). Invasive brain computer interfaces (BCI) have already been demonstrated that can enable locked-in individuals to communicate (Birbaumer, 2006). The question is whether we *could*, but whether we *should*. This could be a dissertation unto itself and will not be fully explored. In the same breathe, not elaborating at all maybe altogether ethically irresponsible.

Philosopher Phillip Kitcher suggests that contrary to traditional notions, science is not value free. Many view science as free and pure form of inquiry completely independent of moral, social, political and religious values. Kitcher suggests this idealistic view of science is a myth. Kitcher has a fairly detailed and elaborate argument, but the gist is that science rarely falls into nice clean categories distinguishing "basic" research from "applied" research from "engineering." The

boundaries are messy and a consequence of this is that pure motivations for Truth are difficult to substantiate. Real motivations may be also be financially motivated (personal or otherwise) or motivated by a desire for recognition, societal influences, or other tainted sources. Scientists are humans and as such cannot fully disentangle their scientific pursuits from their daily lives. Science values empiricism and disconfirmation. But when sufficient evidence is lacking, scientists have a tendency to fail back on subjective judgments for drawing interpretations (Kitcher, 2003).

If a science is not value free, what is the role of science in a democratic society? What are the responsibilities of practicing scientists? In later chapters, Kitcher delves into these questions. One of his suggestions is that part of a scientist's duty is to evaluate the potential ramifications of their research and that those ramifications should even influence whether a particular line of research should be pursued. Arguable, no one understands those implications better than the scientist. If they act as if they are displaced from that responsibility such questions may never be posed, let alone answered. The good news is our future as humans or cyborgs is being discussed. The bad news is it is being discussed on the periphery. If we hold strong to our traditional notion of value free science and leave the discussion of values to *bio-ethicists* we may find their conclusions to not be representative of our own. By that time, the tail may have wagged the dog. Innovations that seem beneficial and innocuous now may lead to capabilities we may not have envisioned.

Before ending this discussion, I suggest that it may be unwise to summarily dismiss the potential future where brain computer interfaces are ubiquitous and our consciousness has transcended its present form. Such notions seem so far-fetched, that they are out of the realm of possibility. But, maybe that is a false intuition. Computational innovation tends to follow exponential trends. Thirty years ago how many among us would have predicted the computational power and capabilities of our smartphones, let alone the rate of smartphone adoption. Estimates place smartphone adoption at ten times the pace of early personal computer adoption. In China, year-over-year adoption exceeds 400% (Miot, 2012). This argument could be taken further by

quantitatively assessing trends, systematically examining the technological hurdles, and forming predictive models. Even without such elaboration, the point can be made, that even if we can't know for certain what *will* happen in 30 years, we should not be passive to what *could* happen in 3 years, and what *might* happen in 30 years.

In the preface to this dissertation, I suggested that technology is not *good* or *bad.* Likewise, potential technology is should be viewed with ambivalence. With this in mind, we should return to introducing the domains simultaneously tackling the practical uses of physiological measures. These domains are intradisciplinary to the cognitive science, but have slightly different approaches, foundational bases, and user populations. The demarcations between these fields of Augmented Cognition, Neuroengineering, Adaptive Automation, and Human Reliability are often blurred.

Wittgenstein recognized that categorical classifications based on a strict adherence to a set of common features often fails. (Wittgenstein, 1953/2001) argued (posthumously) that in natural language categorical descriptions are more akin to family resemblances. If a father has a double chin it is necessary for his son to also have one in order for them to "look alike." Here the reader should think of the distinctions between these fields as one of family resemblance and not of adherence to a strict feature set. As these fields and technologies develop it is likely that the distinctions between them will become even more blurred. The analysis given here is intended to give a broad overview and provide context for how the computational approaches examined here fit into the big picture.

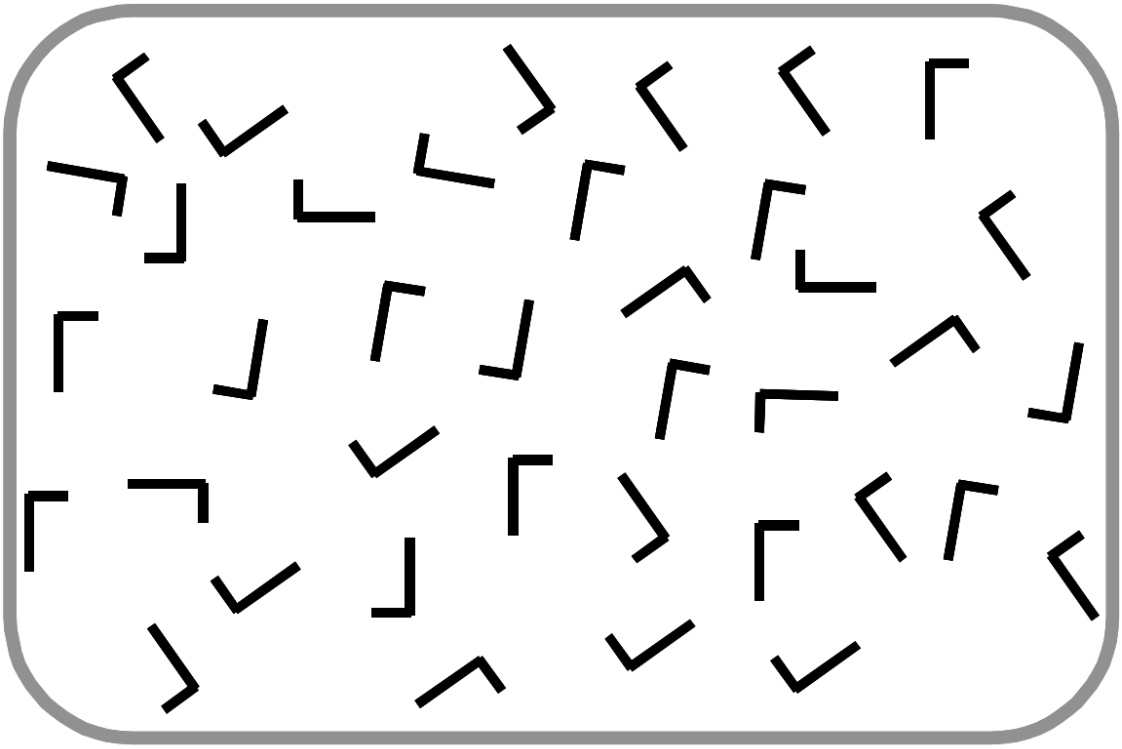
2.3.1 ***Neuroengineering.*** (Parasuraman & Rizzo, 2008) describe neuroengineering as developing Brain Computer Interfaces (BCIs) or other less-invasive channels for monitoring brain signals for human interaction with both the natural and the human-made environment. Information can be used to control a machine. For example signals from the pre-motor and motor cortex can be used to control a computer cursor and allow a person who is "locked-in" to communicate. Similarly BCIs can be used to control a prosthetic limb. BCI also applies to cases

where signals are sent to the brain to alter one's perceptions. Cochlear implants can restore hearing by directly stimulating the auditory nerve. Technologies which might be able to restore vision by directly stimulating the primary visual cortex are in their early phases. The devices listed above are often referred to as "therapy" technologies because they exist to restore missing function.

Other technologies generally known as "enhancements" propose giving humans abilities like direct knowledge and skill acquisition, displaced consciousness, and superhuman intelligence. The time scale of these enhancements is highly speculative along with the ethical implications such technologies. Current neuroergonomics research primarily focuses on clinical therapy technologies. Neuroergonomics relies strongly on the physiological mechanisms which drive human perception and cognition. In clinical applications the risks presented by invasive physiological monitoring devices is outweighed by the potential benefits they can provide to the patients. As these technologies become more reliable and safer they will likely see use in non-clinical applications like augmented cognition.

2.3.2 ***Augmented Cognition.*** Augmented cognition refers to technologies that improve upon human perceptual limitations or enhance cognitive capabilities. Human cognitive capabilities differ from individual to individual and are dynamic as they are influenced by stress and fatigue. Augmented cognition proposes that human machine interfaces should be adaptable to the changing capabilities and limitations of the user (St. John, et al., 2004). Traditional interface design focuses on designing interfaces to match human capabilities. Interaction design is usually specific to a particular task or problem, and the solutions obtained through task analysis, prototyping, and optimization may not generalize to other tasks. In contrast to traditional interface design, augmented cognition focuses on technologies that will improve upon specific cognitive limitations (Pavel, Wang, & Li, 2002). For example, searching for the letter "L" amongst a field with rotated and mirrored distracters requires a serial search (Figure 2.3.1). Pavel, Wang, and Li (2002) demonstrated that reducing the contrast of areas that have been fixated upon can improve the time

Figure 2.3.1 *Example of a search task. The goal is to find the letter L. Reducing the contrast of areas that have been fixated upon can improve the time required to find the target.*



required to find a target when many distracters were present (30 distractor condition). Reducing the contrast of objects helped participants keep track of the objects that had already been examined. A second approach is to use real-time cognitive state detectors to adapt the interface of a system to better suit a user (Young, Clegg, & Smith, 2004). One of the challenges is with devising the cognitive state detectors, but figuring out how to implement the cognitive state information also presents challenges. Open-loop tasks may become unstable when feedback is added. Qualitatively, this is especially true when things update too quickly. In the long-term some envision "human-computer dyads" as achieving magnitude order increases in cognitive abilities (Schmorrow & McBride, 2004).

2.3.3 ***Adaptive Automation.*** Automation refers to technologies which displace human decision making and control with a computational and/or mechanical system. Many automated systems effectively reduce human workload and require limited supervision or intervention. For example, most drivers take it for granted that their car is capable of shifting itself. However, when automation fails or when it is difficult to control the results can be devastating. Degani (2004) attributes the June 30, 1994 crash of an Airbus A330 during flight testing in large part to a failure of the autopilot system to fully and appropriately engage. Shortly after takeoff the autopilot system failed to take into account the fact that one engine was intentionally disabled to simulated an engine failure. The plane continued to climb while losing airspeed and eventually stalled killing the three crew and four passengers on board. This illustrates that in certain cases automation can ironically increase risk and human workload. To be effective the user and the system must function synergistically.

Adaptive automation refers to such systems in which both the user and the system can initiate changes in the level of automation. The challenge of adaptive automation is identifying the appropriate level of automation for given circumstances and how to change the level of automation to facilitate operator performance, maximize system safety, and maximize system productivity.

Traditionally adaptive automation has had a strong basis on system theory, theoretical models of human performance, and usability techniques. In some context identifying when an automated system should take over is relatively straight forward. For example, if a pilot passes out due to extreme g-forces having the plane stabilize itself and not crash into the ground is a perfectly reasonable course of action (Scerbo, 1996). Often times, deciding what an automated system should do is not so clear. In recent years adaptive automation begun looking at psychophysiological measures to trigger changes in the state of automation. Here adaptive automation clearly overlaps with augmented cognition. The primary distinction is that augmented cognition technologies most always utilize physiological measures. Adaptive automation may be based on other diagnostic measures such as critical environmental events, operator performance measurements, operator modeling, or hybrid methods (Parasuraman, Bahri, Deaton, Morrison, & Barnes, 1992).

2.3.4 ***Human Reliability.*** Catastrophic failures can often be attributed to both technological/environmental and human error. Human reliability analysis is the science of qualitatively and quantitatively identifying the human contribution to risk in human machine systems and reducing overall risk (Boring & Kelly, 2008). In recent years the concept of *human error* has been evolving. The traditional view held that complex systems were assumed to be inherently reliable and safe. Mishap arises when humans undermine systems due to their incompetence in understanding the system. The new view recognizes that when mishap occurs, the human assessments and actions might have made sense in the moment.

The logical fallacy of hindsight bias may cause us to think that the outcome was more predictable in the moment than it actually was. In the new view, human errors are a sign of deeper trouble. The goal should be to identify the circumstances surrounding the decisions and correct the *organization* to support safety (Dekker, 2006). Technological systems must take into account that people do not communicate perfectly, are prone to distraction, illogical decision making, and complacency. Despite their shortcomings, humans are still an essential component to safely

operating complex systems. Automated systems work well when they encounter situations that they are familiar with, but they can be wholly unreliable when they encounter unfamiliar situations. In such situations human supervision is not only preferred, but necessary. Even if automated systems can correctly identify the root cause of a malfunction, human operators are still imperative to decide the best solutions given the constraints of the scenario. In the old view technology was considered sufficient if it provided all the relevant information to operators. The new view recognizes that the information's format and presentation matters.

All fields discussed in this section are important to the theory, implementation, and actual use of human machine interfaces using physiological signals. The following section will review the current state of technology used to record physiological signals.

2.4 Current and Future Physiological Measurement Technologies

The domains discussed in the previous section rely on neuroimaging and physiological monitoring technologies to provide information relevant to neuroanatomy and neurophysiology. These technologies serve two critical roles. First, they are tools which aid in increasing our empirical knowledge of the brain and for testing theoretical models simulating the brain. Secondly, and more obviously, these measures provide the signals for which human machine interfaces use to interact with the natural and human-made environment. These technologies have a wide array of strengths and shortcomings when it comes to their spatial resolution, temporal resolution, signal-to-noise characteristics, invasiveness, obtrusiveness, and cost (Parasuraman & Rizzo, 2008). To date an ideal device for making physiological measurements does not exist. For clinical applications such as brain computer interfaces (BCIs) used to restore the function of limbs, control prosthetics, or restore communication the risks and costs associated with invasive and costly devices such as electrocortigraphy (ECoG) may be justified (Kotchetkov, Hwang, Appelboom, Kellner, & Connolly, 2010). Within other domains, like augmented cognition, less invasive and less costly technologies

like pupil diameter, skin conductance, and HRV are better suited. The following section describes the relative strengths and weaknesses of existing measurement technologies.

2.4.1 ***Functional Magnetic Resonance Imaging (fMRI).*** fMRI has high (good) spatial resolution of less than one millimeter. fMRI has low (poor) temporal resolution on the order of .5 to 2 seconds. By scanning a localized area at a higher rate the temporal resolution can be improved to 100 ms (Sabatinelli, Lang, Bradley, Costa, & Keil, 2009) but even this improved scanning rate is still one to two orders of magnitude slower than ECoG and EEG. fMRI devices are also not mobile, costly to obtain, and costly to run. The primary advantage of fMRI is that it provides both three-dimensional structural and functional information. Structure and function can be obtained for both cortical and sub-cortical regions.

An fMRI machine consists of a large superconducting magnet. A superconducting magnet is similar to an electromagnet with the primary exception being they are cooled with liquid helium to a temperature of 4.2 Kelvin. At this temperature the niobium-titanium coil has no electrical resistance which means no energy is lost to heat and large amounts of current can be applied to the magnet. Conventional electromagnets would have to be much larger and the electricity required to power them would make fMRIs prohibitively expensive The superconducting magnet is typically between 1.5 to 3 Tesla. For comparison a typical refrigerator magnet is about 1/1000 of a Tesla. A second coil called a gradient coil creates a well controlled magnetic field within the imaging plane. When tissue is placed within this plane polarized molecules stop spinning and align with the magnetic field. When the tissue is pulsed with an RF coil these molecules quickly resonate before re-orienting with the magnetic field. Different types of molecules (water, fats, etc.) will resonant in ways which makes it possible to obtain structural information (Hornak, 2010).

Functional information is obtained analyzing the resonances of oxygen carrying hemoglobin molecules. Oxygenated-hemoglobin shows a slightly different resonance after it becomes deoxygenated. Increases in neural activity are correlated with increased metabolic activity. For

reasons that are not completely known the amount of oxygenated blood supplied to active areas actually exceeds demand. This increased ratio of oxygenated hemoglobin to deoxygenated hemoglobin is detectable and referred to as the blood oxygenation level dependent (BOLD) effect (Parasuraman & Rizzo, 2008). The unique structural and functional imaging capabilities of fMRI imaging make it an invaluable tool for scientific research. However the high costs and sheer size make it poorly suited for implementing human machine interfaces.

2.4.2 ***Electroencephalography (EEG).*** EEG is popular for both research and development of BCIs as well as for assessing mental workload because it has very good temporal resolution (0.1 ms is not uncommon), non-invasiveness, established, and affordability. The primary drawbacks of EEG are its susceptible to noise and its poor spatial resolution. The use of high density arrays with multidimensional analysis have greatly improved the spatial resolution of EEG (Weis, Romer, Haardt, Jannek, & Husar, 2009; Sajda, Muller, & Shenoy, 2008), but noise is still problematic.

EEG uses electrodes placed on the scalp to detect small changes in electrical potential caused by the summation of synchronous activity of large neuronal populations. Before EEG can record these signals they must first pass through the skull which attenuates the signal. EEG must also contend with noise from eye movements, muscles, electrodes, line power, and nearby equipment which further degrades the signal-to-noise ratio (SNR; Parasuraman & Rizzo, 2008).

In theoretical research and clinical applications noise can be compensated by multiple stimuli exposures and averaging multiple recordings to uncover event-related potentials (ERPs) in the time-domain (Gazzaniga, Ivry, & Mangun, 2002). In real-world applications high levels of noise and insufficient information regarding stimulus onsets make using ERPs impractical (Huang, T.-P., & Makeig, 2007). At present EEG research requires EEG experts to sort measurements from artifacts. Most efforts to use EEG for human machine interfaces use time-frequency based methods such as short time Fourier transform (see section 2.3) or wavelets (see section 2.5). With these

methods workload has been associated with increased frontal midline theta activity in conjunction with increased parietal midline alpha activity (Parasuraman & Rizzo, 2008).

EEG measurement must also contend with what has come to be known as the inverse problem. The electrical currents which are measured on the scalp originate from within the cranium. The inverse problem describes identifying the source of these currents from scalp measurements. Helmholtz in the late 19th century solved the *forward solution:* the location of a single electrical event or locations of multiple electrical events within a homogenous conducting medium will reliably produce the same pattern on the surface of a sphere. However, given a surface pattern an infinite number of inverse solutions exist. The approach to the inverse problem is to attempt to optimize solutions based on parsimony by first modeling single electrical events. If a single event cannot be found to emulate the observed pattern then two electrical points are modeled, and so on. As the reader is probably aware, at any given time multiple brain locations are likely to be active (Gazzaniga, Ivry, & Mangun, 2002). For human machine interfaces machine learning algorithms can work around the inverse problem. Algorithms need only to identify patterns from the measured signals. Sourcing the activity to particular locations may reduce the complexity of the problem, but is not strictly necessary. In other research settings, such as identifying epileptogenic zones, the inverse problem becomes extremely important. Electrocorticography or ECoG is similar to EEG except that electrodes are placed within the cranium. This does not eliminate the inverse problem but does greatly reduce localization errors (Zhang, van Drongelen, Kohrman, & He, 2008). ECoG is discussed in more detail in the following section.

2.4.3 ***Electrocorticography (ECoG).*** ECoG uses electrode arrays placed directly on the neocortex or on the dura mater. The best performing BCIs use dense ECoG arrays of more than 100 electrodes. With monkeys this type of setup has reached an information rate of 6.5 bits per second. This information rate is roughly equivalent to typing 15 words per minute (Santhanam, Ryu, Yu,

Afshar, & Shenoy, 2006). Contemporary technologies using EEG can only achieve 5-10 characters per minute. With ECoG rates of 10 bits per second are expected in the near future (Linderman, et al., 2008). Because of the limitations associated with EEG the future of reliable and autonomous BCIs for clinical applications rest with ECoG. For adaptive systems utilizing measures of mental workload the information rate requirements are much smaller.

BCIs with EEG will undoubtedly show further progress but are unlikely to match the performance of ECoG systems (Lotto, Congedo, Lecuyer, Lamarche, & Arnaldi, 2007). The primary shortcoming of ECoG is its invasiveness. Current ECoG systems require transcutaneous connectors to power and transfer data from electrode arrays. Besides being highly susceptible to infection this also limits the functional lifetime of an ECoG implant to one year (Linderman, et al., 2008). The solution is to develop implantable prosthetic processors (IPP) that can be powered through induction in the same manner current cochlear implants are powered.

The primary challenge with creating an IPP is to keep the power dissipation to within 80 mW/cm2 to reduce the risk of damaging brain tissue. The IPP will have specialized circuitry for sensing, digitizing, spike sorting, decoding, and transmitting data. The rationale for performing the signal processing within the IPP is to reduce amount of information the IPP must transfer wirelessly to reduce power dissipation. The factors affecting signal quality are fairly well understood. The amounts of power required by each stage of the IPP are within feasible tolerances. The neuroengineers proposing these IPP suggest that the hardware challenges are actually less problematic than the software challenges and larger performance gains can be made on the software side (Linderman, et al., 2008). These IPPs may eventually see use in non-clinical applications related to augmented cognition, adaptive automation, and human reliability but not in the near future. Besides the hardware obstacles there are still many software obstacles to overcome.

Current BCIs are only operated for short sessions. Before every session the BCIs must be trained. Some of the training parameters are only stable for a couple of hours. Software need to become more robust by being capable of continuously adapt to changing neuronal patterns. Performance gains can also be made by making the software context dependent. For example, prosthetic limbs should behave differently when the patient is asleep versus when they are awake. Research into context sensitive BCIs is just getting started (Linderman, et al., 2008).

2.4.4 ***Functional Near-Infrared Imaging (fNIR).*** fNIR uses an array of sensors to detect light scattering caused by oxygenated hemoglobin supplying blood vessels. fNIR has much better spatial resolution, about 1 cm, compared to EEG but has poorer temporal resolution on the order of 10ms. Another important distinction is that fNIR can only monitor localized cortical changes where EEG arrays can cover the entire scalp. Compared to existing technologies fNIR is still in its infancy (Parasuraman & Rizzo, 2008). However within the domain of Augmented Cognition fNIR has seen more success. St. John et al. (2004) conducted a review of DARPA funded augmented cognition projects and identified fNIR as one of the more promising and reliable techniques of assessing cognitive workload.

2.4.5 ***Transcranial Doppler Sonography (TCD).*** TCD is a non-invasive relatively inexpensive technology which uses ultrasound to monitor the velocity of blood flowing through cerebral arteries. Ultrasonic frequencies do not readily penetrate the cranium which makes some individuals with cranial anomalies difficult to measure. TCD was initially developed to diagnosis subarachnoid hemorrhages. Its use has been adapted to human machine interfaces by monitoring blood flow velocities through the main stem intracranial arteries. Faster velocities over baseline are indicative of greater metabolic activity in to the corresponding cortical hemisphere. Davies and Parasuraman (1982) found that TCD could detect increased workload elicited by a vigilance task. Shaw, Guagliardo, de Visser, and Parasuraman (2010) found TCD was not especially sensitive or

instantaneous.

2.4.6 ***Pupil Diameter (PD).*** The pupil is innervated by both the sympathetic and parasympathetic nervous systems and is influenced by mental concentration, anxiety, attention, motivation, emotional excitement, lighting, respiratory, and blood pressure fluctuations (Just, Carpenter, & Miyake, 2003). PD has been linked to differences in difficulty of tasks including sentence processing, mental calculations and user interface evaluation (Murata & Iwase, 2000; Just & Carpenter, 1993; Nakayama & Shimizu, 2004; Nakayama & Katsukura, 2007). While these measures have shown signs of promise, they generally have smaller effect sizes and are less reliable between individuals than EEG and fNIR (St. John et al., 2004). Despite these shortcomings PD and SC should not be completely dismissed. Compared to EEG and fNIR they are generally less costly, more portable, and less obtrusive.

Pupil diameter measurements focus a camera on the eye. Because the cornea is highly reflective to infrared light most systems use an infrared illuminator to increase the contrast of the pupil relative to the iris. Besides developing the actual equipment to record physiological signals a second technological challenge is to devise algorithms to process these raw sensor signals and integrate it into automated systems (St. John et al., 2004). Previous studies have applied spectral analysis to PD to predict workload with some success. Nakayama and Shimizu (2002; 2004) found the power spectrum density of PD between 0.1- 0.5 Hz (Hertz, cycles per second) and 1.6 -3.5 Hz to increase with difficult mental arithmetic. Murata and Iwase (2000) only examined spectral density under 0.6 Hz with the Sternberg memory search task and found that power in this band decreased with the number of elements in the search task. While the presence of this spectral signature is intriguing, Nakayama and Shimizu did not provide an instantaneous measure of workload.

Marshall's Index of Cognitive Activity, or ICA, (2000; 2002; 2007) uses wavelet decomposition to estimate the PD power spectrum for measuring real-time cognitive activity. This approach uses thresholds to convert the real valued wavelet components to a bit depth of 1.

Converting the wavelets to 1-bit resolution reduces noise but is also likely to remove information relevant to cognitive workload. Workload classification is then based on either summing the processed wavelet components contained in the binary vectors, or applying linear discriminant analysis or neural network estimation to all or part of the vectors of binary wavelet components.

2.4.7 ***Skin Conductance (SC, also known as galvanic skin resistance, GSR).*** As scientifically minded readers you may already know that electrical resistance is the reciprocal of electrical conductance. Regardless of what the parameter is called the changes result from the psychogalvanic reflex response in which the palmar and plantar eccrine sweat glands fill with a solution primarily comprised of water and NaCl in response to stress (Harrison, Boyce, Loughnan, Dargaville, Storm, & Johnston, 2006; Jacobs, et al., 2001). GSR is perhaps most noted for its role in a lie detector (polygraph). In an experimental study conducted during real police interrogations GSR was significantly better than chance with a fairly standard card test procedure. In the card test the suspect was shown a card with a number between 1 and 6, they were then instructed to answer "no" to all of the following questions and asked "did you choose card number {1, 2, 3, 4, 5, 6}." GSR alone was able to detect the correct card in 35 of 62 suspects, which works out to be more than three times better than chance (Kugelmass, Lieblich, Ben-ishai, Opatowski, & Kaplan, 1968). In the context of criminal procedures this is certainly not beyond *a reasonable doubt*, but out of context implies at least some validity for the ability of GSR to actually measure stress.

In an observational study examining preoperative stress, skin conductance was highly correlated with blood pressure, epinephrine and norepinephrine levels (Storm, Myre, Rostrup, Stokland, Lien, & Ræder, 2002). Eccrine sweat glands are also capable of reabsorbing NaCl to reduce salt loss. The psychogalvanic reflex is mediated by the sympathetic division of the autonomic nervous system. When the eccrine glands fill electrical conductivity on the skin increases, and when eccrine glands reabsorb NaCl there is a corresponding decrease in conductivity. Skin Conductance or GSR measurement devices usually have two electrodes which are

placed a short distance apart either on the palm or on the index and middle finger. A small amount of DC voltage is applied across the electrodes and the current between them is measured. Skin conductance can then be calculated by simply dividing the measured current by the applied voltage (Ohm's law defines the relationship between current, voltage, and resistance; conductance is the inverse of resistance).

Skin conductance signals can be analyzed in a variety of ways. The simplest is to simply calculate mean conductance over a given epoch. This method is known as the tonic response or the skin conductance level (SCL; tonic is more commonly found in the older literature). A second method examines the phasic response (or skin conductance response, SCR), a transient increase which quickly returns to a baseline state. Harding and Punzo (1971) used a stimulus response task where the number of distracters was manipulated to increase the uncertainty of responding. Some trials required motor responses to targets while other trials required only paying attention to the presentation. Harding and Punzo found that both tonic and phasic responses were higher when trials required motor responses. They also found that phasic responses were highly affected by the uncertainty manipulation.

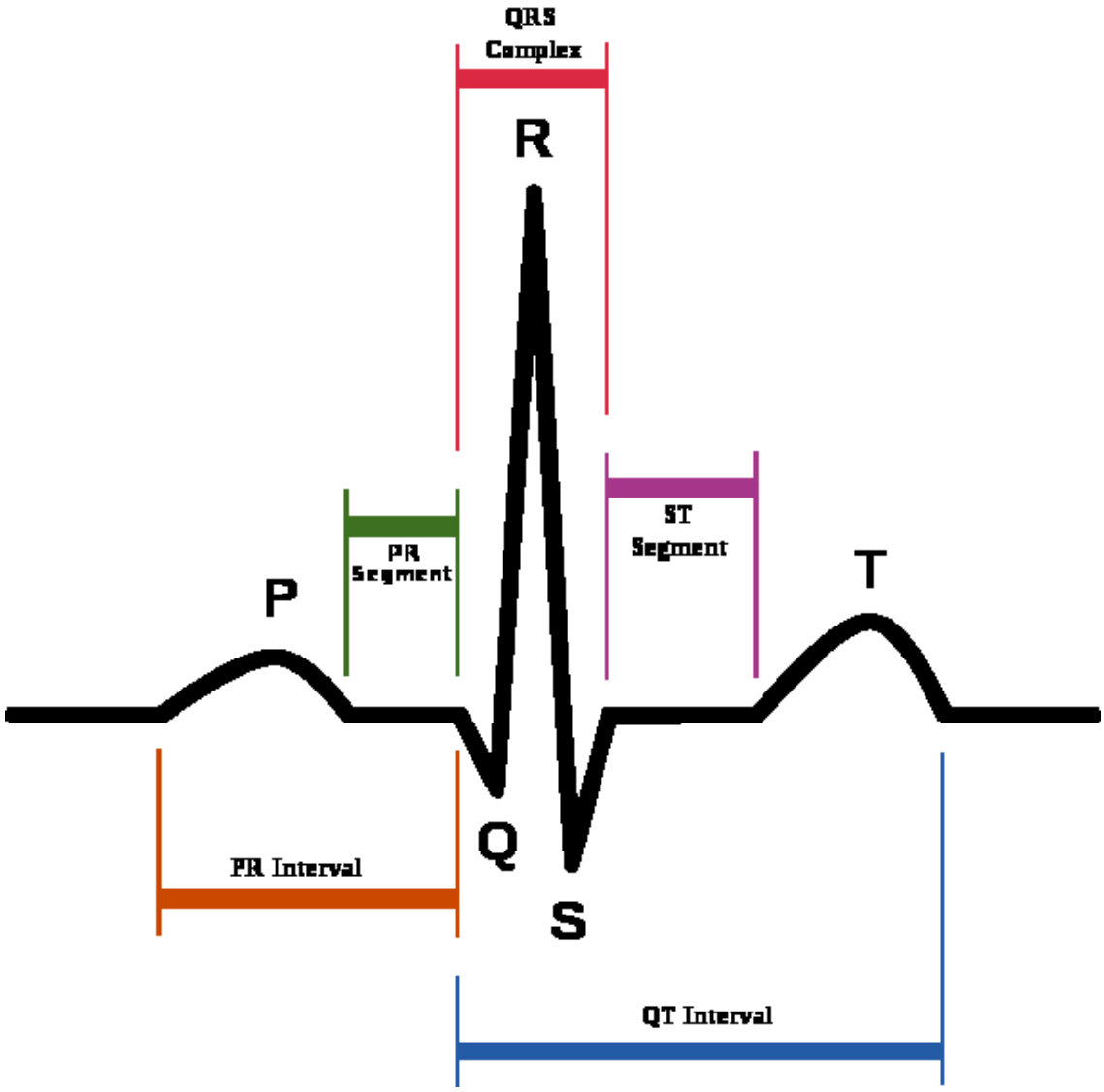
More current research suggests that tonic and phasic responses dissociate between activity in the orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (VMPFC) and activity in striate and extrastriate cortices, anterior cingulate and insular cortices, thalamus, hypothalamus and lateral regions of prefrontal cortex (Nagai, Critchley, Featherstone, Trimble, & Dolan, 2004). This suggests that tonic response is associated with general arousal where the phasic response is associated with task dependent emotional processing and resource overload. Figner and Murphy (2010) suggest that phasic responses are often times anticipatory, although they may be interoceptively (relating to stimuli within the participant) or exteroceptively (relating to stimuli outside the participant) rooted. Positive emotions may also cause phasic responses but not necessitate concern for operator intervention. These sorts of issues make distinguishing whether a

response genuinely reflects stress difficult. Also, measuring tonic responses is complicated by the fact that eccrine glands will also secrete as a thermoregulation mechanism regardless of whether a stress inducing stimuli is present (Grimnes, 1982). Skin conductance measures also show large individual variability caused by physiological differences in the density of eccrine sweat glands, gender differences, differences between extroverts and introvert (Orme-Johnson, 1979).

2.4.8 ***Heart Rate Variability (HRV)*** As the name implies HRV reflects the standard deviation about the mean heart rate. HRV is measured using an electrocardiogram which detects electrical fluctuations reflecting heart muscle activity. During each heart beat the muscle cells build up an electrical charge followed by a depolarizing response and corresponding contraction. In the electrocardiogram this rhythm is elucidated by the P, Q, R, S, and T phases. The R phase shows a sharp transient response reflecting the depolarization and is most easily detectable. The *R-R interval* is used to calculate heart rate (HR) in beats per minute (bpm; see **Figure 2.4.1**). Heart rate is heavily influenced by breathing. Heart rate increases during inspiration, and decreases during expiration. Although the mechanisms are not understood in their entirety HRV also reflects cardiac autonomic activity (Billman, 2011). Then HRV can be calculated as the variability about the HR (Murai, Hayashi, Nagata, & Inokuchi, 2003).

In the frequency domain heart rate variability can be grossly divided into low frequency and high frequency components. Low frequency activity (< 0.1 Hz) may or may not reflect sympathetic nervous system activity. Some studies have shown increases in low frequency activity while other studies have shown no change or even decreases in conjunction with sympathetic activity. Houle and Billman (1999) theorize that low frequency activity reflects a complex interaction of sympathetic and parasympathetic activity. High frequency (> 0.2 Hz) content is viewed as a reliable indicator of parasympathetic efferent activity (Houle & Billman, 1999; Guger, Leeb, Pfurtcheller, Antley, Garau, & al., 2004; Wiederhold, Davis, & Wiederhold, 1998).

Figure 2.4.1 *Depiction of a normal sinus rhythm ECG trace.The fluctuations in the trace represent sequences in the heart beat cycle. HRV is calculated on the timing between R peaks (the RR interval).*



Backs (1995) found that mean heart rate as well as HRV shows reliable correlations between gross changes in difficulty although it did not provide enough resolution to distinguish between fine changes in task difficulty. Fournier, Wilson, and Swain (1999) found HR and HRV to be more sensitive to task difficulty. They found significant differences between single and multiple-task conditions as well as between multiple-task low and high conditions, as well as between medium and high conditions. In their study HRV lacked the sensitivity to distinguish between low and medium multiple-task conditions. HRV is complicated by noise from respiration causing fluctuations with each and every breath (respiratory sinus arrhythmia). HRV is also susceptible to heart rate fluctuations caused by changes in environmental conditions and physical exertion. Individual differences also contribute to HRV. Increased low frequency HRV correlates with cardiovascular disease (Guger, et al. 2004). The measure remains attractive because of its low cost, unobtrusiveness, and robustness to external noise.

2.5 Conclusions

Here I am particularly interested in using pupil diameter, skin conductance, and heart rate variability for assessing mental workload. These measures are all relatively inexpensive, relatively unobtrusive, and can be used simultaneously without causing interference with one another. Most importantly these measures are all currently at our disposal. As previously stated the primary aim of this study is to examine whether specialized algorithms designed to make use of multiple measures and compensate for individual differences can provide greater reliability and sensitivity. The pace of technological advancement is ferocious and future technology will undoubtedly offer faster, smaller, cheaper, less noisy physiological measurement devices. The computational techniques I use here are designed to evolve and make the most of the information available to them to process, analyze, and interpret physiological signals. As such they are extensible to other

measures. The measure I have chosen to use here are convenient, but I believe enough precedent has been established to use them as a valid test bed for our primary aim.

Machine learning refers to a corpus of techniques which allow computer systems to recognize patterns in data. Recognizing patterns from physiological signals can be troublesome for a variety of reasons. The number of input signals can quickly grow, and the relationship between the input variables and the desired output are likely non-linear because the processes which generate them are non-linear. If we recall our discussion on stress individuals may show large differences in their reactions and performance when presented with identical scenarios. The BCI research has shown that physiological responses of individuals vary from day to day. To combat these problems I employ a machine learning technique known as genetic programming. To increase the efficacy of genetic programming I first apply wavelet transformation to the physiological variables. As previously discussed frequency content to sympathetic and parasympathetic activity are often localized to different frequency bands. Wavelet transformation can discriminate between these bands while maintaining the temporal resolution needed for making instantaneous measure of mental workload.

The following chapters provide introductions to wavelet transformation and genetic programming. Then three preliminary experiments are presented. The first two are far from definitive but they provide a record of our rationale and progress. Experiment 3 shows some efficacy of wavelet decomposition in conjunction with genetic programming to function as a *lagging* indicator of performance. In real world situations, indicators of mental workload need to *lead* or predict changes in performance. An experiment is proposed which tests the ability of wavelet transformation and genetic programming to function as a leading indicator of performance.

Chapter 3: Spectral Analysis

3.1 Preface

This document aims to provide an accessible introduction to the "nuts and bolts" of spectral analysis and wavelets. Unlike most introductions to the topic, it is intended for an interdisciplinary audience. The introduction here attempts to build the reader's mathematical intuition (conceptual) instead of focusing on mathematical formalism and rigor (mechanics). This approach may differ from what readers are familiar with.

To understand the current state of math education one must look to the post WWI (World War I) Bourbaki mathematicians. The Bourbaki were a young group of French mathematicians who published numerous works under the pen name "Nicholas Bourbaki." The Bourbaki can be credited for providing the mathematical style we are accustomed to today. The Bourbaki stressed a formal "definition-proof-theorem" format where mathematical proofs which begin by declaring definitions followed by several lemmas (intermediary proofs) until the solution is reached. The definitions and lemmas must all correspond to a set of axioms or self-evident statements. For instance, in Euclidean geometry one axiom is that two parallel lines on a plane will never intersect. However, non-Euclidean geometries may not include this axiom. If every step is well established based on the chosen axioms, definitions, and lemmas the proof is rigorous (Wells, 2007). Most of us probably take it for granted that the theorems contained within our mathematical texts contain this level of rigor but historically this has not always been true (Munson, 2010). Prior to the Bourbaki, many mathematics followed in the intuitional tradition embraced by many including Henri Poincaré. Instead of simply working solutions forward from mathematical definitions and lemmas, Poincaré would also work backwards from his intuitions. Poincaré describes his creative process in *Foundations of Science* (1913/2012, p. 387):

> For fifteen days I strove to prove that there could not be any functions like those I have since called Fuchsian functions. I was then very ignorant; every day I seated myself at my

> work table, stayed an hour or two, tried a great number of combinations and reached no results. One evening, contrary to my custom, I drank black coffee and could not sleep. Ideas rose in crowds; I felt them collide until pairs interlocked, so to speak., making a stable combination. By the next morning I had established the existence of a class of Fuchsian functions, those which came from the hypergeometric series; I had only to write out the results, which took but a few hours.
>
> Then I wanted to represent these functions by the quotient of two series; this idea was perfectly conscious and deliberate, the analogy with elliptic functions guided me. I asked myself what properties these series must have if they existed, and I succeeded without difficulty in forming the series I have called theta-Fuchsian.
>
> Just at this time I left Caen, where I was then living, to go on a geologic excursion under the auspices of the school of mines. The changes of travel made me forget my mathematical work. Having reached Coutances, we entered an omnibus to go some place or other. At the moment when I put my foot on the step the idea came to me, without anything in my former thoughts seeming to have paved the way for it, that the transformations I had used to define the Fuchsian functions were identical with those of non-Euclidean geometry. I did not verify the idea; I should not have had time, as upon taking my seat in the omnibus, I went on with a conversation already commenced, but I felt a perfect certainty. On my return to Caen, for conscience' sake, I verified the result at my leisure. (Poincaré, 1908).

Poincaré's work is often criticized for lacking rigor, and rightly so. Many of his papers are filled with sloppy, incomplete, or non-existent proofs. Over the years some of these proofs have been explicated, but many have turned out to be plain wrong. Poincaré recognized the necessity of rigor in mathematics generally, but was not overly concerned with rigor with his own work. In Poincaré's work the detail of his intermediary proofs reflect the amount of confidence he had in those particular intuitions. Once Poincaré was intrinsically satisfied he had solved a problem he put it aside and went looking for the next problem. In Poincaré's view too much rigor and formalism just results in restraining intuition (McLarty, 1997).

The Bourbaki mathematicians detested Poincaré's use of intuition and sloppy style. In some regards, modern mathematics should be grateful to the Bourbaki. They established mathematical structures based on sets for defining collections of objects (subsets, sets of subsets), as well as operations and relations between objects. These *mother* structures of algebraic, topological, and order are irreducible and independent of the chosen system of axioms (Wells, 2007). However, in

hindsight it seems only rationale to question whether abandoning mathematical intuition altogether is wise. Poincaré argued that if all of mathematics was required to be rigorous and formal then no math existed before the 1820s. Perhaps then, abandoning mathematical intuition is a case of throwing the baby out with the bathwater. Rigor and formalism are necessary for establishing theorems, but not necessarily for discovering them, and most certainly not for presenting them.

A repercussion of the Bourbaki is that mathematical documents have a tendency to make perfect sense if you already know what they mean to begin with, but almost no sense if you are unacquainted with the topics. There are several reasons for this. First off mathematics is about abstraction and generalization. The power of mathematics is in its ability to be applied to different datasets and different problems. The mathematics does not care whether they are looking at seismological data, cosmological data, or physiological data. The downside to this abstraction is that it can be difficult for non-mathematicians to get a firm grasp of what the mathematics mean when no concrete examples are provided. Another reason mathematics is inaccessible is because mathematics is a very precise language. Math is about finding invariance – things that remain unchanged in the midst of change – and the conditions which guarantee they don't change. The consequence is that mathematical theorems read like a laundry lists of conditions, lemmas, and properties.

In contrast, here we try to acquaint the reader without drowning them in mathematical formalism. Here I may describe a concept approximately before trying to describe it precisely. We also make use of footnotes to point out notation the reader might not be familiar with, or to explain things that might not be obvious. We can also make use of visual examples and illustrations wherever possible and attempt to use the language of math to support and analyze the visual examples. Where proofs are included we try to make them as intuitive and easy to follow as possible. Some proofs are disregarded all together but referenced by name for those who wish to

look into them, and some longer proofs, which are deemed to have substantial importance, are included but mostly for the sake of completeness. Where possible we try to point out insights which tie together mathematical topics which are usually treated separately.

Those of us in applied fields of study want to solve problems not necessarily discover theorems. Carl Sagan said “If you want to make an apple pie from scratch, you must first create the universe.” So it is with analytical techniques. When we use an analysis of variance to test whether conditions are statistically different from one another or apply a Fast Fourier transform algorithm to a time varying signal we are relying on a *universe* of axioms, lemmas, and theorems we may or may not be aware of. In many instances, some ignorance is understandable; we can’t be expected to understand *the universe* before running a t-test, but perhaps one of the consequences is relying on analytical techniques as opposed to the mathematics underlying the techniques. To illustrate in joke form (Quintopia, 2007):

> A mathematician, physicist, and an engineer are asked to find the volume of a red sphere. The mathematician looks at the sphere and then measures the circumference and then divides by 2 to calculate the radius and then performs a triple integral to find the volume. The physicist looks at the sphere and then gets a graduated cylinder and fills it part way with water. She then submerges the sphere in the water and states that the volume of the sphere is equal to the displacement of the water. The engineer takes the sphere and starts look at it all over, making sure that he sees the whole surface of the sphere. Then he pulls out a book of tables and starts flipping through it. He starts to get nervous as he nears the end of the book. When he finishes with the book, he pulls out another book of tables and keeps flipping through it. When he finishes with his second book he asks "Hey, do any of you have tables for red spheres?" In frustration he decides "Aw, never mind, I've got a red cube table, I'll just approximate it to that."

The point of the joke is not to poke fun at engineers. After all this is in many aspects an applied endeavor as much as anything else. The point is to illustrate a trend towards taking technique *X* applying it to measure *Y* and publishing the results of the endeavor, then taking technique X’ and applying to Y and publishing the endeavor, ad nauseam. One only needs to look to the title of this very document. Topically, this strategy does seem to yield progress, but what are we really doing? Engineer A. R. Dykes (1946) description of structural engineering might provide some insight:

> Structural engineering is the art of modeling materials we do not wholly understand into shapes we cannot precisely analyze so as to withstand forces we cannot properly assess in such a way that the public at large has no reason to suspect the extent of our ignorance.

The description alludes to the fact that in applied settings we often need to deal with large amounts of uncertainty. Different fields have different techniques for dealing with uncertainty. Physicists try and reduce uncertainty by controlling the environment. Mathematicians have the luxury of abstraction. In engineering we need tools that are flexible, but we should not fall into the trap of over relying on the tools while forgetting the fundamentals. Available computational tools *automate* aspects of data pre-processing, reduction, and analysis. This automation can be misguided. Exploiting their powers requires only a superficial understanding of their processes. By understanding the mathematics behind these tools we may be able to accomplish our goals more simply or perhaps more effectively.

3.2 Introduction

Psychology provides theoretical considerations for defining what mental workload is. Our review of human workload suggests that physiological measures capture relevant information pertaining to workload, although it is still unclear exactly how those measures predict workload. By tackling the mathematical foundations beyond traditional techniques of collecting time-frequency information from physiological we can likely devise simpler, more robust, more effective techniques tailored to the data under evaluation. Machine learning algorithms are still needed to fill the gap between mapping signal features to mental workload. If the quality of the input data is improved the accuracy and reliability of machine learning algorithms, like genetic programming, is likely to increase.

So far we have reviewed the necessity of extracting workload measures from physiological signals. We have also reviewed physiological signals of interest. As previously discussed our physiology utilizes complementary feedback mechanisms with oscillatory characteristics. Here we

turn our attention to how we can extract the spectral information contained within physiological measures. The majority of spectral analysis can be traced to Joseph Fourier's 1807 discovery of Fourier Series – a means of representing periodic signals as infinite series of cosine and sine functions (Weisstein, Fourier Series, 2011).

Much of the mathematical advancements in the subsequent 200 years have been built on Fourier's original ideas. Fourier transforms allow us to represent time varying signals as frequency varying signals. In the real world many signals are non-ergodic (non-stationary) meaning their frequency characteristics are not stable over time. With such signals it is useful to be able to characterize signals in both time and frequency. This is possible with short-time Fourier transforms (STFT, moving window Fourier transforms) is a suboptimal as we will soon see. Moving window Fourier transforms (STFT) provide fixed time and frequency resolution. The size of the window dictates both the time and the frequency resolution. The uncertainty principle prohibits localizing signals in both time and frequency. Longer windows yield better frequency resolution but poorer time resolution. Shorter windows yield better time resolution but poorer frequency resolution. Secondly, discrete Fourier transforms are also suboptimal because they linearly divide frequency when logarithmic divisions are often a more natural fit for real signals.

It is only in the past 30 years or so that wavelet transforms have offered an optimized approach to extracting both time and frequency information from signals by scaling kernel functions to the frequencies they are examining. This provides an optimal compromise between time and frequency resolution. Discrete wavelet transforms provide orthogonal frequency bands with logarithmic spacing. Understanding conceptually what wavelet decomposition does is not overly difficult, but an understanding of how wavelet decomposition works and why it works requires some dedication and some mathematical formalization.

3.3 Vector Spaces

Before we introduce vector spaces let's begin with something we should all have some familiarity with: the real number line. The real number line extends from $-\infty$ to ∞.[1] It contains integers (numbers: …, -2, -1, 0, 1, 2, …), rational numbers (numbers that can be defined as fractions: $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, …), and irrational numbers (numbers that cannot be expressed by fractions: pi, *e*, $\sqrt{3}$). In mathematical notation real numbers are often denoted as $\mathbb{R}$. In practical terms the significance of real numbers is that we can use them to measure physical quantities. Measurement is possible because $\mathbb{R}$ is a basic algebraic structure known as a *field*.

The field of real numbers is mathematically *complete* and *ordered*. Completeness essentially means that $\mathbb{R}$ has infinite precision. If we only had whole numbers we would not be able to take the average of 1, 4, and 5 because $10/3$ is not a whole number. The whole number line does not provide enough precision to capture 3-1/3. In a similar manner the set of rational numbers does not have enough precision to capture $\sqrt{2}$. Completeness is described by saying that $\mathbb{R}$ contains all limits. Irrational numbers, like $\sqrt{2}$, can be describes as limits of Cauchy Sequences. The second property of order means that the elements can be arranged in magnitude. The consequences of these properties mean that for scalars *a*, *b*, and *c* in $\mathbb{R}$ we can use these familiar operations:

$a + b = b + a$	*Community*
$a + (b + c) = (a + b) + c$	*Associativity*
$a + (-a) = (-a) + a = 0$	*Inversivity*
$a(b + c) = ab + ac$	*Distributivity*
$a + 0 = a, a \times 1 = a$	*Identity*

Vector spaces extend these properties and operations of real numbers to multiple dimensions.

---

[1] The concept of Infinity (∞) is one that is often misunderstood. ∞ is not a number but a *process* that is formally unbound. The concept was originally proposed by Cantor (1887). The "∞" symbol is used in two ways 1) to represent a magnitude which either increases above all limits or decreases to an arbitrary smallness but always remains finite (e.g. the sequence: $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, …), or 2) as a limit point such that any point lying "in finite distance" from it has the same behavior as the limit point.

3.3.1 ***Definition of a Vector Space.*** While the real numbers work great for single dimensional concepts like length and mass, the complexity of most problems often require representing them with more than a single dimension. For example, the Earth is a three dimensional spherical object but the scale is so large that it appears flat from our terrestrial perspective. We could apply a three dimensional Cartesian coordinate system to reference the location of objects on the surface using three parameters, but we instead we often project the sphere onto a plane which allows us to represent the spatial relationships between physical objects on planar surface. When we project small areas of the globe onto a Euclidean plane the distortion caused by the projection is minimal relative to the represented area. This mathematical deconstruction is known as a Riemannian manifold and allows one to treat the curvature of the Earth as a Euclidean plane.

The practical implication of a map is that we can look at a map and figure out how to get from point A to point B based on the spatial arrangement of landmarks on the map, we can use scale to estimate the distance between objects because we know the shortest distance between two points is a straight line. We can measure the distance between two points using the Pythagorean theorem. We might take these properties for granted, but here we wish to point out that these affordances are they are endowed from the fact that Euclidean space is generalized from the real number line to higher dimensions. A point on a plane can be referenced by two scalar coordinates x (left to right) and a y (up and down). These scalars can be expressed as a vector $\mathbf{p} = (x, y)$. A vector is basically a list of coordinates. In the Euclidean plane vectors will always have two dimensions. In three-dimensional space a vector will have three coordinates. For example, $\mathbf{p} = (x, y, z)$. Like scalars, vectors can be together added together. For vectors: $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$:

$$\mathbf{x} + \mathbf{y} = (x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$$

They can also be multiplied by scalars:

$$a \cdot \mathbf{x} = a \cdot (x_1, x_2, \dots, x_n) = (ax_1, ax_2, \dots, ax_n)$$

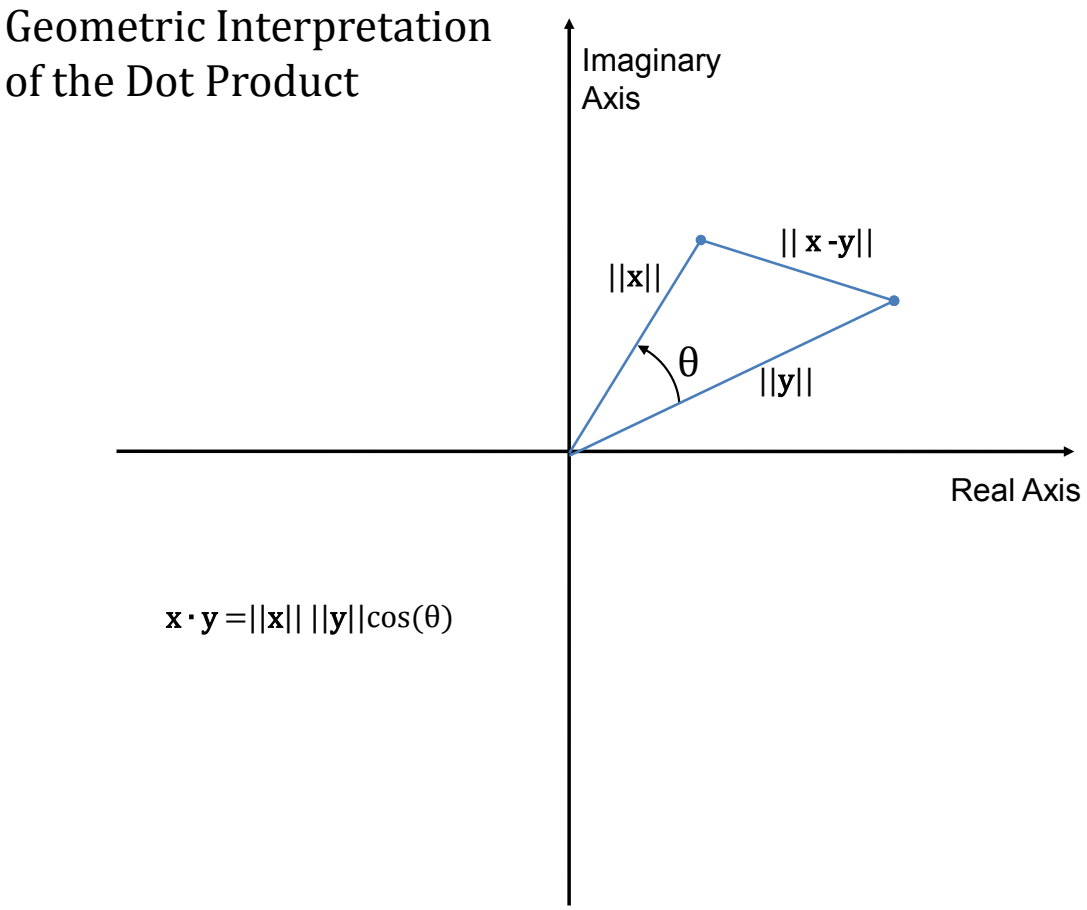
where · denotes scalar multiplication. With vector addition and scalar multiplication we can define a vector space as a set of vectors that abide by the following rules of community, associativity, inversivity, distributivity, and identity. For vectors **x**, **y**, and **z** and scalars *a*, *b*, and *c*:

$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$	*Community*
$\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$	*Additive Associativity*
$(ab) \cdot \mathbf{x} = a(b \cdot \mathbf{x})$	*Multiplicative Associativity*
$\mathbf{x} + (-\mathbf{x}) = (-\mathbf{x}) + \mathbf{x} = 0$	*Inversivity*
$a \cdot (\mathbf{x} + \mathbf{y}) = a \cdot \mathbf{x} + a \cdot \mathbf{y}$	*Scalar Distributivity*
$(a + b) \cdot \mathbf{x} = a \cdot \mathbf{x} + b \cdot \mathbf{x}$	*Vector Distributivity*
$\mathbf{x} + 0 = \mathbf{x}$	*Additive Identity*
$1 \cdot \mathbf{x} = \mathbf{x}$	*Multiplicative Identity*

Other vector operations exist but are not necessary to include in the definition of a vector space; we will hold off discussing them until they become relevant.

Vector spaces can also extend to higher dimensions. For instance a Euclidean $\mathbb{R}^3$vector space could be defined with the basis $\mathbf{i} = (1,0,0)$, $\mathbf{j} = (0,1,0)$, and $\mathbf{k} = (0,0,1)$. Vector spaces may also have non-geometric analogues. For example, suppose we wanted to track the progress of 100 runners over a 10 kilometer course. We could construct a $\mathbb{R}^{100}$ vector space where each index separately tracks the distance each runner has traveled. Since the location of any given runner does not depend on any of the other runners the 100 dimensions are linearly independent of one another. This vector space would be called a 100-tuple vector space. Vectors may also be continuous.

Figure 3.3.1 *Geometric Interpretation of the Dot Product. The dot product between two vectors x and y relates to the angle between them through the law of cosines.*



3.3.2 ***Basis Vectors.*** A two-dimensional vector space of real numbers may be denoted as $\mathbb{R}^2$. Within this system we can define two vectors. The first is $\mathbf{i} = (1,0)$ and begins at the origin and extends along the positive x-axis one unit. The second is $\mathbf{j} = (0,1)$ and also begins at the origin, but extends one unit along the positive y-axis. These vectors **i** and **j** are called unit vectors because they have a magnitude (length) of 1. The set of vectors $\{\mathbf{i}, \mathbf{j}\}$ can be called a basis for $\mathbb{R}^2$ since they *span* $\mathbb{R}^2$ and are *linearly independent* to one another.[2]

The spanning property means that any point in $\mathbb{R}^2$ can be specified as a linear combination of **i** and **j**. For example a vector $\mathbf{p} = a\mathbf{i} + b\mathbf{j}$. Furthermore, any choice of scalars *a* and *b* will result in a vector that is contained in $\mathbb{R}^2$. The linear independence property means that scalars **a** and **b** are independent of one another. With the basis vectors defined parallel to the Cartesian axes linear independence means a point can move parallel to the x-axis without changing its distance from the y-axis, and a point can move parallel to the y-axis without changing its distance from the x-axis .

The set of vectors $\{\mathbf{i}, \mathbf{j}\}$ can be called orthonormal vectors because they are basis vectors which are also unit vectors. Within $\mathbb{R}^2$an infinite number of alternative bases can be defined. For instance we can rotate the vectors **i** and **j** by 45° in the clockwise direction. Our new vectors denoted as $\mathbf{i_1} = \left(\sqrt{2}/2\,, \sqrt{2}/2\right)$ and $\mathbf{j_1} = \left(\sqrt{2}/2\,, -\sqrt{2}/2\right)$ would also be a valid basis for $\mathbb{R}^2$. This basis would also be considered orthonormal.

3.1.3 The dot product. We previously mentioned that basis vectors are linearly independent if they are perpendicular to one another. We can also test the linear independence of vectors by looking at their dot product. The dot product of two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is defined as,

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots, + x_n y_n. \qquad .$$

[2] The brackets are used to refer to a mathematical collection of objects called a set. The objects within the set are called elements. The *cardinality* of a set measures the total number of elements in a set.

The concept of a dot product is generalized to continuous complex functions in the following sections and becomes essential to spectral analysis. Here we present these concepts in 2 and 3 dimensional Euclidean space which makes it easier to explain and illustrate. The reader should keep in mind that although the content presented here may seem like a distraction in the larger context it is rather crucial.

The length, magnitude, or norm of a vector $\mathbf{x}$ is denoted as $\|\mathbf{x}\|$ and defined as,

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

[Technically not all vector spaces calculate length in this manner. This formula for length holds in Euclidean space and other Lebesgue spaces (L^p spaces) where $p = 2$. For this reason the definition shown above is also known as the Euclidean norm. Understanding this technicality is not important to the concepts at hand but is mentioned for completeness.]

In Euclidean geometry the dot product of $\mathbf{x}$ and $\mathbf{y}$ is related to the angle between the vectors θ and the magnitudes of the vectors through,

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos(\theta) \quad .$$

Figure 3.3.1 graphically depicts the geometric relationship between $\mathbf{x}$, $\mathbf{y}$ and θ. To prove this we can use the law of Cosines. The law of Cosines $c^2 = a^2 + b^2 + 2abcos(\theta)$ relates the lengths of the sides of a triangle a, b, c to the angle θ opposite side c.

The origin, point $\mathbf{x}$, and point $\mathbf{y}$ specify a triangle. The angle at the origin's corner is given by θ. The length of the side opposite θ is $\|\mathbf{x} - \mathbf{y}\|$ (see Section 2.2.8.1-2 on metric and normed spaces). When we apply the law of cosines:

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos(\theta)$$

$$(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) = \mathbf{x} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y} - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos(\theta)$$

$$\mathbf{x} \cdot \mathbf{x} - \mathbf{x} \cdot \mathbf{y} - \mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} = \mathbf{x} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y} - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos(\theta)$$

$$-\mathbf{2}\, \mathbf{x} \cdot \mathbf{y} = -2\|\mathbf{x}\|\|\mathbf{y}\| \cos(\theta)$$

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos(\theta)$$

With some algebraic manipulation,

$$\theta = \cos^{-1}\left(\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right) \quad .$$

The Cauchy Schwartz inequality proves that,

$$|\mathbf{x}\cdot\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$$

and guarantees that the argument will always be bounded between the interval $[-1, 1]$.[3]

When we calculate the angle between our orthonormal basis vectors **i** and **j** we can see that the argument for the inverse cosine function becomes 0 because $\mathbf{i}\cdot\mathbf{j}$ is 0 and both **i** and **j** are unit vectors. The inverse cosine of 0 is $\pi/4$ or 90° which agrees with our geometric interpretation of the basis vectors being perpendicular. We can interpret this result by stating that if the dot product of two vectors is zero those vectors are linearly independent from one another. When we check our rotated basis vectors $\mathbf{i_1} = \left(\sqrt{2}/2\,, \sqrt{2}/2\right)$ and $\mathbf{j_1} = \left(\sqrt{2}/2\,, -\sqrt{2}/2\right)$ we see that:

$$\mathbf{i_1}\cdot\mathbf{j_1} = \frac{\sqrt{2}}{2}\left(\frac{\sqrt{2}}{2}\right) + \frac{\sqrt{2}}{2}\left(-\frac{\sqrt{2}}{2}\right) = \frac{1}{2} - \frac{1}{2} = 0, \therefore \mathbf{i_1} \text{ and } \mathbf{j_1} \text{are linearly independent.}$$

When vectors have more than two dimensions the geometric interpretation of the dot product still applies. In higher dimensions two vectors will always define a plane and θ will relate to that plane.

3.3.3 ***Vector Subspaces.*** Another interesting and often useful property of vector spaces is that they may contain subspaces. A subspace is a set of vectors that is closed under addition and scalar multiplication. The subset *S* is considered a subspace of $\mathbb{R}^n$ if the following hold:

(1) The zero vector **0** is an element of *S*

(2) If vectors **u** and **v** are elements of *S*, then **u + v** is an element of *S*

(3) If vector $\mathbf{v} \in S$ and *c* is a scalar, then *c***v** is an element of *S*

For example, $\mathbb{R}^3$contains subspaces defined by basis {0,0,0}, all lines that run through the origin, all planes that run through the origin, and the set $\mathbb{R}^3$is a subspace of itself. Now that we have presented a basic overview of vector spaces let's return to the intertwined spirals problem.

[3] This is important because $\cos^{-1}$ is only defined between $[-1, 1]$.

3.3.4 ***The Complex Plane.*** In section 2.1 we presented the intertwined spirals problem and discussed how switching from Cartesian coordinates to Polar coordinates significantly reduces the difficulty of the problem for machine learning algorithms. When we use Polar coordinates instead of Cartesian coordinates we are really using an alternative basis in the $\mathbb{R}^2$ vector space. The polar coordinate system uses two parameters. The parameter r specifies the radial distance from the origin, and the second parameter θ specifies the angle in the counterclockwise direction from the x-axis. The polar basis $\{\mathbf{e}_r, \mathbf{e}_\theta\}$ can be defined in terms of the Cartesian unit vectors **i** and **j** as follows:

$$\mathbf{e}_r = \cos(\theta)\mathbf{i} + \sin(\theta)\mathbf{j}$$

$$\mathbf{e}_\theta = -\sin(\theta)\mathbf{i} + \cos(\theta)\mathbf{j}$$

The vector $\mathbf{e}_r$ is referred to as the radial vector and the vector $\mathbf{e}_\theta$ is referred to as the traverse or angular vector. A second advantage of the polar coordinate representation is that along with complex number theory it lends itself to describing polar coordinates. To begin this discussion let us first introduce complex numbers. A complex number is a number with a real and imaginary part. Complex numbers are usually expressed in the form $a + jb$, where a and b are real numbers and the j term represents $\sqrt{-1}$ such that $j^2 = -1$, $j^3 = -j$, and $j^4 = 1$. If we recall from trigonometry, these complex numbers represent angles on the complex plane containing the unit circle. In many ways the complex plane is isomorphic to $\mathbb{R}^2$ vector space but it is not identical to $\mathbb{R}^2$. To understand how they differ let us first define a $\mathbb{R}^4$ vector space M_{22} with basis matrices of:

$$E_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad E_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$E_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \text{ and } \qquad E_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The basis vectors shown here meet the linear independence and spanning properties for a vector space. Now we can introduce a two-dimensional subspace $\mathbb{C}$ within M_{22} defined by the basis matrices

$$\text{Re} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \qquad \text{Im} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Here we see that the complex plane is a two-dimensional subspace of the four-dimensional vector space M_{22}.

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = a\text{Re} + b\text{Im}$$

In our more familiar complex notation:

$$a\text{Re} + b\text{Im} = a + jb.$$

We can see that that $j^2 = -1$ identity still holds by applying product multiplication to Im:

$$\text{Im} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = -\text{Re}.$$

We can move on to show the multiplication of two complex numbers $a\text{Re} + b\text{Im}$ and $c\text{Re} + d\text{Im}$

$$(a\text{Re} + b\text{Im})(c\text{Re} + d\text{Im}) = ac\text{Re}^2 + bc\text{ImRe} + ad\text{ReIm} + bd\text{Im}^2$$

$$= (ac - bd)\text{Re} + (\text{ad} + bc)\text{Im}.$$

Because the product of two complex numbers can always be expressed as proportions of Re and Im we can say that the subspace $\mathbb{C}$ is closed under matrix multiplication. Geometrically this is analogous to saying that the product of two complex numbers will always remain on the complex two-dimensional plane within the 4-dimensional M_{22} vector space. Thorugh similar exercises $\mathbb{C}$ can be shown to be closed to divison, addition, and subtraction. In addition $\mathbb{C}$ also meets the additive identity property and the multiplicative identity property. Because $\mathbb{C}$ has these properties $\mathbb{C}$ is also an algebraic field. The closed topology of the field $\mathbb{C}$ guarantees that any polynomial equation of degree $n \geq 1$ has a sum total of n real and imaginary roots. This is known as the fundamental theorem of algebra. In this context, we can see that the natural habitat for polynomial functions is $\mathbb{C}$.

The length of a complex number is found using the modulus or complex norm. For a complex number $\mathbf{z} = a + jb$ the modulus is denoted as $|\mathbf{z}|$ and defined as:

$|\mathbf{z}| \stackrel{\text{def}}{=} \sqrt{a^2 + b^2}$.

3.3.5 ***Polar Coordinates.*** Many problems such as calculating the distance between two points are well suited for Cartesian coordinates. Other problems may better lend themselves to a polar coordinate system. In machine learning there is well known binary classification benchmark known as the intertwined spirals problem. The goal of the intertwined spirals problem is to build a classifier which given a set of points can identify whether they belong to either spiral A or spiral B (Chalup & Wiklendt 2007; see Figure 3.3.2). When learning algorithms are trained using points defined by their Cartesian coordinates learning progresses much slower and is generally less accurate compared to when the same algorithms are trained using the same points but given polar coordinates. Complex numbers can be transformed into polar coordinates through the following identities:

$$a + jb = r[\cos(\alpha) + j\sin(\alpha)] \qquad \textbf{3.3.1}$$

where,

α is the angle

r is the magnitude

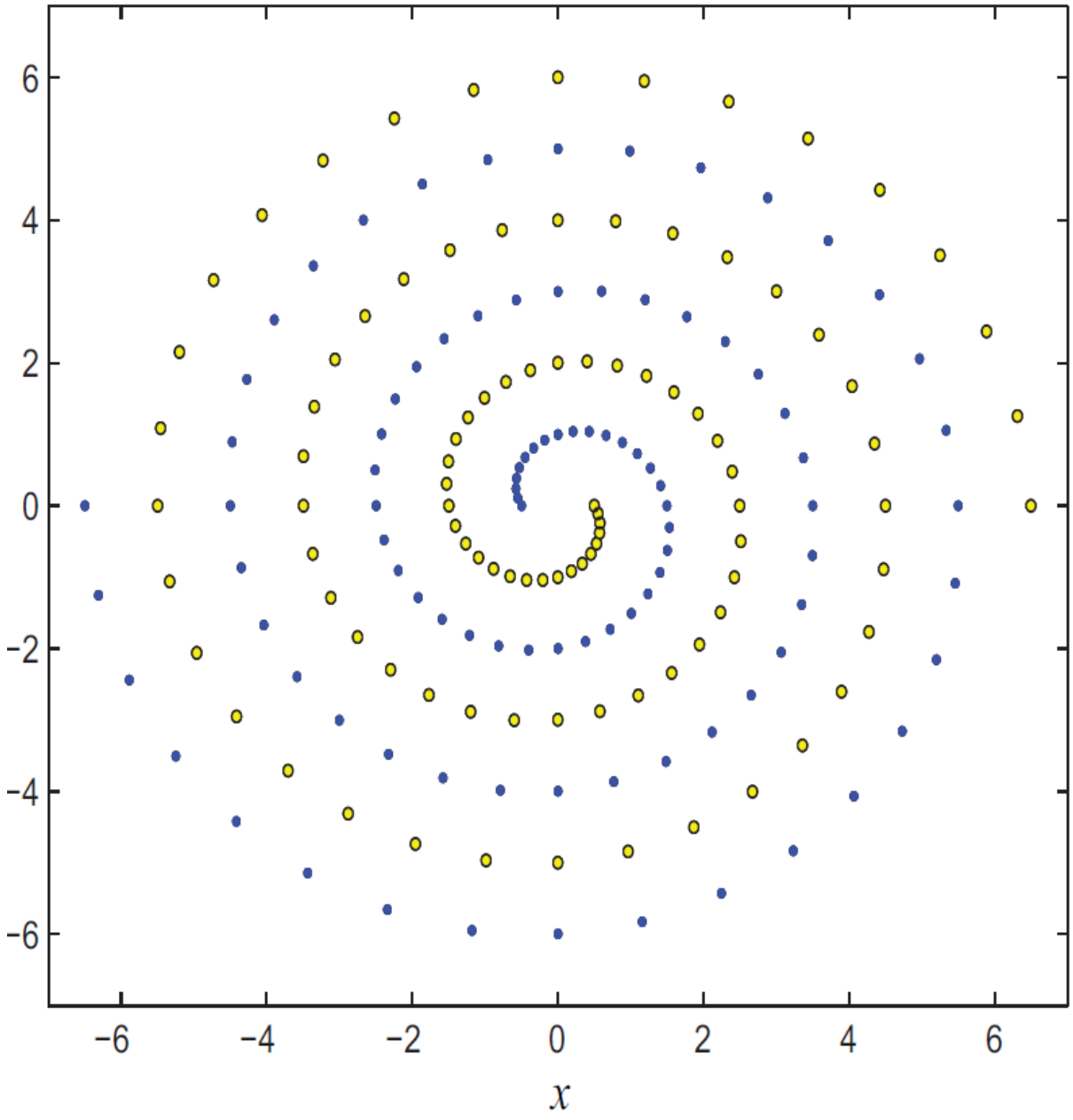
$a + jb \neq 0$

$$r = \sqrt{(a + jb)(a - jb)} = \sqrt{a^2 + b^2}$$

$$\tan(\alpha) = \frac{b}{a}, \sin(\alpha) = \frac{b}{r}, \cos(\alpha) = \frac{a}{r}$$

Using polar coordinates simplifies the multiplying and dividing of vectors. In polar form the product is derived by simply multiplying the magnitudes and summing the angles. Division accomplished by dividing the magnitudes and subtracting the denominator's angle from the numerator's angle. While some tasks are easier to accomplish with polar coordinates some operations, such as representing signals, transforming signals, and solving differential equations,

Figure 3.3.2 *Illustration of the intertwined spirals problem. A classifier to distinguish a point as belonging to spiral A (blue) or to spiral B (yellow).*



to the use of Euler's Formula will be presented in section 3.4.1.1

3.3.6 ***Function Spaces.*** So far our attention has been focused on finite dimensional vector spaces. For Newtonian physics three dimensional Euclidean space is a powerful tool. However, Euclidean space is ill-suited for working with time-varying signals. A key concept to understanding spectral analysis is that the basis for a vector space does not have to be a set of finite dimensional vectors. A basis can also be a set of functions. For example in Fourier analysis a basis will always consists of sine and cosine functions at different frequencies.

When a basis is a collection of functions the basis is used to define functions. In these cases a vector space is a function space (Weisstein, Function Space, 2011). The function space provides the same concepts of closure and spanning as their vector space equivalents. In spectral analysis we have an underlying assumption that signals possess periodic characteristics that can be described by variations of amplitude and phase across an infinite number of frequencies. To deal with the theoretical implications of infinite dimensional function spaces we need to have infinite bases with infinite number of elements. Discrete decompositions will always have bases with a finite number of elements.

3.3.7 ***Abstract Vector/Function Spaces.*** To work with these theoretical concepts we have a variety of abstract vector spaces. These vector spaces share a great deal of overlap, but the idea behind having so many classifications of vector spaces is that they provide slightly different utilities due to the fact that the rigidity of the classification requirements trades off with increasing mathematical structure. One of the most important abstract vector spaces for Fourier analysis and wavelet analysis is Hilbert space.

A rather dry mathematics joke with its origins traced to the halls of MIT goes something like this (Weisstein, Hilbert Space, 2011):

"Do you know Hilbert?"

"No"

"Then, what are you doing in his space?"

(*laughter should result*)

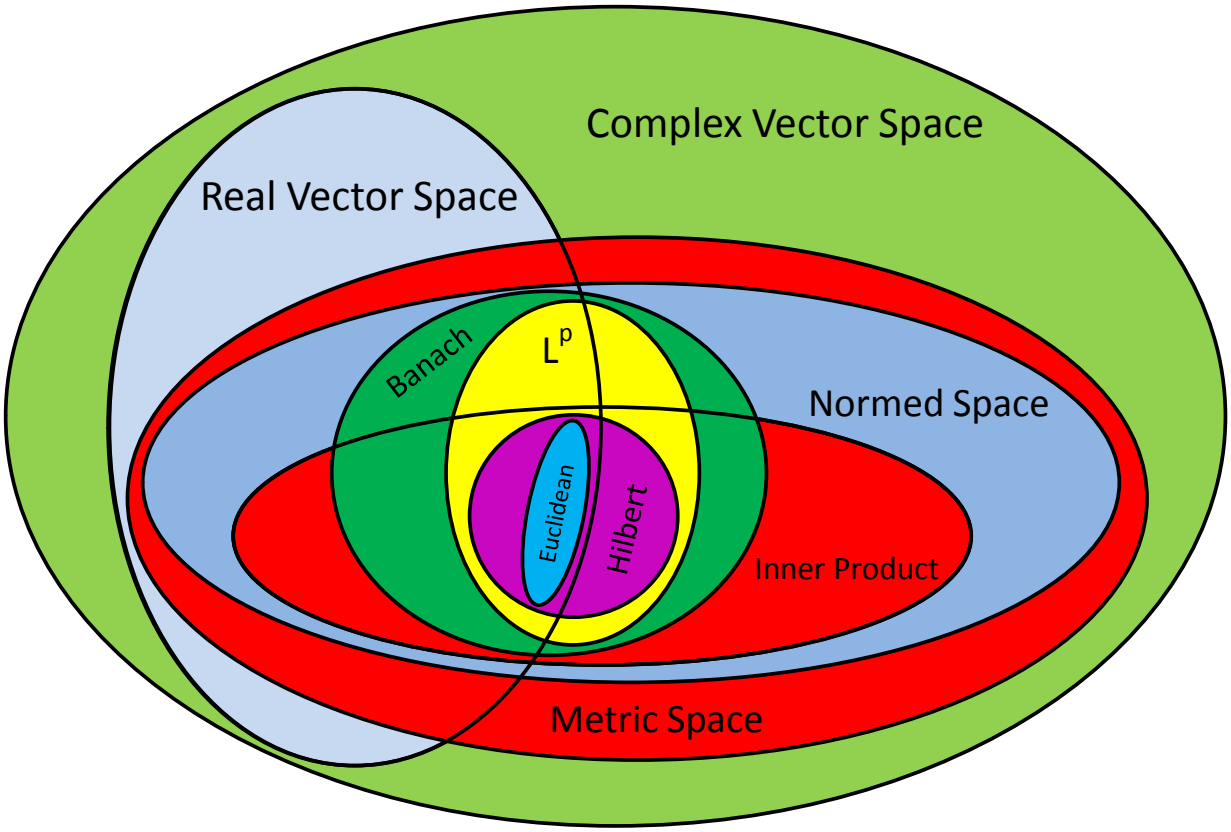
The "humor" of the joke is that Euclidean space is a Hilbert space. Figure 3.3.3 presents a Venn diagram illustrating the hierarchy of abstract vector spaces in relation to Euclidean space. The importance of Hilbert space is that it generalizes the dot product to higher dimension vector spaces including infinite vector spaces. The inner product can be real or complex. The utility of Hilbert spaces is the additional structure they provide to these more abstract *inner product* spaces (Weisstein, Inner Product Space, 2011). Hilbert spaces also generalize the Euclidean distance metric to abstract vector spaces, and Hilbert spaces are complete. They also come with orthonormal basis which are extremely useful when decomposing signals into their constituent frequency components. To begin a more technical presentation of Hilbert spaces we first introduce metric spaces and work our way up through the hierarchy presented in the Figure 3.3.3 Euler diagram.[4]

3.3.8 ***Metric space.*** We begin our discussion by looking at metric spaces. A metric space is a real or complex vector space with a distance function (Weisstein, Metric Space, 2011). The distance function $d(\mathbf{x}, \mathbf{y})$ measures the distance between elements **x** and **y.** The distance function must satisfy the following four properties for a vector space V and elements $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ in V:

$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$	*Symmetry*
$d(\mathbf{x}, \mathbf{y}) > 0, \quad \text{if } \mathbf{x} \neq \mathbf{y}$	*Non-negativity*
$d(\mathbf{x}, \mathbf{y}) = 0, \quad \text{if } \mathbf{x} = \mathbf{y}$	*Identity*
$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$	*The Triangle Inequality*

[4] Euler diagrams are similar to Venn diagrams in that they depict relationships between sets. Euler diagrams are less restrictive because empty sets do not need to be explicitly represented.

Figure 3.3.3 *Hierarchy of abstract vector spaces. This Euler diagram conveys the humor of the joke from section. It also attempts to illustrate the relationships between various vector spaces.*



For example a distance function satisfying these requirements could be defined as:

$$d(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } \mathbf{x} \neq \mathbf{y} \\ 0, & \text{if } \mathbf{x} = \mathbf{y} \end{cases}$$

In most situations this distance function would not be terribly useful, but it would satisfy the distance metric requirement of a metric space. Next, we introduce normed spaces. All normed spaces are metric spaces but not all metric spaces are normed.

3.3.8.1 Normed space. A normed space is a real or complex vector space with a norm function (Weisstein, Normed Space, 2011). For a vector space *V*, elements **x, y** in *V*, and scalar c in $\mathbb{R}$ or $\mathbb{C}$ the norm function $\|\mathbf{x}\|$ must satisfy the following properties:

$\|\mathbf{x}\| > 0$, if $x \neq 0$

$\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$

$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

All normed spaces are metric spaces because a distance function is associated with their norm. For elements **x, y** in *V* the distance function of a normed space is defined as:

$d(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \|\mathbf{x} - \mathbf{y}\|$.

Recall from section 2.2.3 the Euclidean norm was defined as:

$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$**.**

For two points $\mathbf{p_1} = (x_1, y_1)$ and $\mathbf{p_2} = (x_2, y_2)$ in the Euclidean plane we can see that the distance between them is:

$$\begin{aligned} d(\mathbf{p_1},\mathbf{p_2}) &= \|\mathbf{p_1} - \mathbf{p_2}\| \\ &= \|(x_1 - x_2, y_1 - y_2)\| \\ &= \sqrt{(x_1 - x_2, y_1 - y_2) \cdot (x_1 - x_2, y_1 - y_2)} \\ &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \end{aligned}$$

a distance formula that we should all be familiar with.

3.3.8.2 Banach space. A Banach space is a complete normed space (Moslehian, Rowland, &

Weisstein, 2011). Earlier, we discussed what completeness is in the context of real numbers. Here we elaborate on the concept by defining a complete space as one that is square-integratable or quadratically integratable. A continuous function $f(x)$ is square-integratable if:

$$\int_{-\infty}^{\infty} |f(x)|^2 \, \mathrm{d}x < \infty.$$

If a finite dimensional vector space has a norm then completeness can be proved through Cauchy sequences. So any finite dimensional normed vector space is also a Banach space. This brings us one abstract space away from getting to Hilbert space. Next up are Lesbesgue spaces.

3.3.8.3 Lesbesgue spaces ($\boldsymbol{L^p}$ spaces). Lesbesgue spaces are complete normed spaces where the norm is generalized from Euclidean space (Rowland, Lp-space, 2011). Recall that for finite vector $\mathbf{x}$ in Euclidean space the Euclidean norm is:

$$\|\mathbf{x}\| = \sqrt[2]{\mathbf{x} \cdot \mathbf{x}}\,.$$

Showing how the norm is generalized to Lesbesgue spaces is perhaps easier done by showing than explaining. Finite Lesbesgue spaces are denoted as ℓ^p and the norm belongs to the family of functions given by:

$$\|\mathbf{x}\|_p = \sqrt[p]{\mathbf{x} \cdot \mathbf{x}}, \quad \text{for } 0 < p < \infty.$$

When $p = 2$ the space can be called ℓ^2 and the norm is equivalent to the Euclidean norm.

It follows that the distance metric for finite vectors $\mathbf{x}$ and $\mathbf{y}$ is given by:

$$d_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p = \sqrt[p]{(\mathbf{x} - \mathbf{y})^{\mathrm{T}} \cdot (\mathbf{x} - \mathbf{y})}, \quad \text{for } 0 < p < \infty.$$

As with Banach spaces in the finite case completeness can be proved through Cauchy sequences.

For continuous cases, Lesbesgue spaces are denoted as L^p. Continuous Lesbesgue spaces generalize the square-integratable requirement presented in the Banach space section (on page 67) to ensure they converge:

$$\int_{-\infty}^{\infty} |f(x)|^p \, \mathrm{d}x < \infty.$$

Assuming the above is true the norm for an as L^p space is defined as:

$$\|f(x)\|_p = \left(\int_{-\infty}^{\infty} |f(x)|^p \, \mathrm{d}x \right)^{\frac{1}{p}}$$

The Riesz–Fischer theorem satisfies the completeness requirement for L^p spaces. In mathematics we may refer to the Euclidean plane as $L^2(\mathbb{R}^2)$ or refer to the complex plane as $L^2(\mathbb{C})$. Phew, we have finally made it to Hilbert spaces.

3.3.8.4 Inner product space and Hilbert space. Since we have gone through the trouble of reviewing abstract vector spaces we can now define a Hilbert space as a complete normed vector space with a norm induced by the inner product (Weisstein, Function Space; Normed Space, 2011). As previously mentioned inner product spaces generalize the dot product to higher dimensional finite vector spaces and continuous vector spaces. Inner product spaces are function maps which take two vectors as arguments whose output can be in $\mathbb{R}$ or $\mathbb{C}$. The $\mathbb{R}$ case is more straightforward and is in many ways isomorphic to the dot product presented earlier. The primary difference being the norm $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$ is not a required property of an inner product space. For a real vector space V and vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and scalar $c \in \mathbb{R}$ the inner product is defined as a real function $\langle \mathbf{x}, \mathbf{y} \rangle : V \times V \to \mathbb{R}$ such that:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$$

$$\langle c\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, c\mathbf{y} \rangle = c\langle \mathbf{x}, \mathbf{y} \rangle$$

$$\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{z}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$$

$$\langle \mathbf{x}, \mathbf{x} \rangle > 0, \quad \text{when } \mathbf{x} \neq 0$$

$$\langle \mathbf{x}, \mathbf{x} \rangle = 0, \text{ when } \mathbf{x} = 0 \quad .[5]$$

When $\mathbb{C}$ is a complex vector space and $\langle \mathbf{x}, \mathbf{y} \rangle : V \times V \to \mathbb{C}$ things get a little more complicated. Consider the following for $\mathbf{x} \in V$ when V is complex.:

[5] The symbol $\in$ means "element of." The notation $\langle \mathbf{x}, \mathbf{y} \rangle : V \times V \to \mathbb{R}$ means the inner product of $\mathbf{x}$ and $\mathbf{y}$ is given by $\langle \mathbf{x}, \mathbf{y} \rangle$ and the function is a map which takes two elements from real vector spaces and has a real output.

$$0 < \langle j\mathbf{x}, j\mathbf{x}\rangle = j\langle \mathbf{x}, j\mathbf{x}\rangle = j^2\langle \mathbf{x}, \mathbf{x}\rangle = -\langle \mathbf{x}, \mathbf{x}\rangle < 0$$

Obviously, we have a contradiction here. The inner product cannot be both less than AND greater than 0. To resolve the contradiction we can take the conjugate of the second argument before calculating the products. This is known as the Hermitian Inner Product (Rowland, Hermitian Inner Product, 2011). For finite complex vectors **z** and **w:**

$$\langle \mathbf{z}, \mathbf{w}\rangle \stackrel{\text{def}}{=} \sum_i \mathbf{z_i}\overline{\mathbf{w_i}}.$$

When V is a complex vector space for vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and scalar $c \in \mathbb{R}\vee\mathbb{C}$ the inner product is defined as a real function $\langle \mathbf{x}, \mathbf{y}\rangle: V \times V \to \mathbb{C}$ such that:

(1) $\langle \mathbf{x}, \mathbf{y}\rangle = \overline{\langle \mathbf{y}, \mathbf{x}\rangle}$

(2) $\langle c\mathbf{x}, \mathbf{y}\rangle = c\langle \mathbf{x}, \mathbf{y}\rangle$

(3) $\langle \mathbf{x}, c\mathbf{y}\rangle = \bar{c}\langle \mathbf{x}, \mathbf{y}\rangle$

(4) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z}\rangle = \langle \mathbf{z}, \mathbf{x} + \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle$

(5) $\langle \mathbf{x}, \mathbf{x}\rangle > 0$, when $\mathbf{x} \neq 0$

(6) $\langle \mathbf{x}, \mathbf{x}\rangle = 0$, when $\mathbf{x} = 0$.[6]

From the Property (2) we can see that the Hermitian inner product is linear in its first argument and from Property (3) we can see it is antilinear with its second argument. By definition this means the complex inner product is sesquilinear. Because the real inner product, shown above, is linear for both its first and second arguments it is by definition bilinear form. The increased complexity of the operation is a tradeoff for maintaining the Properties (5) and (6).

Using the Hermitian inner product defined above we can check to see if our contradiction is resolved:

$$0 < \langle j\mathbf{x}, j\mathbf{x}\rangle = j\langle \mathbf{x}, j\mathbf{x}\rangle = j\,\bar{j}\langle \mathbf{x}, \mathbf{x}\rangle = -j^2\langle \mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle$$

[6] Mathematicians and Engineers usually take the complex conjugate of the second argument. Physicists prefer to take the complex conjugate of the first argument. The properties presented above only apply when the second argument is conjugated. We will stick to this assumption throughout the text.

[The complex conjugate of j is $-j$. Since $j^2 = -1$ it follows that $-j^2 = 1$ which resolves the contradiction.]

The complex inner product properties presented above apply equally well regardless of whether the vectors are finite or continuous. The inner product of continuous complex functions $f(x)$ and $g(x)$ with infinite bounds is

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x)\overline{g(x)}\, dx.$$

The bounds on continuous inner product spaces are not required to be infinite. For example The inner product of continuous complex functions $f(x)$ and $g(x)$ could also have finite bounds

$$\langle f, g \rangle = \int_{0}^{1} f(x)\overline{g(x)}dx$$

or more generically defined as

$$\langle f, g \rangle \stackrel{\text{def}}{=} \int_{a}^{b} f(x)\overline{g(x)}dx.$$

The inner product will converge if $f(x)$ and $g(x)$ are square integratable due to the Cauchy-Schwartz inequality for integrals (Daubechies, 1992)

$$\left| \int_{a}^{b} f(x)\overline{g(x)}dx \right| \leq \left(\int_{a}^{b} |\, f(x)|dx \right)^{1/2} \left(\int_{a}^{b} |\, g(x)|dx \right)^{1/2}$$

Hilbert spaces do many things. For starters, they allow us to represent time varying signals as complex valued frequency varying signals. Hilbert spaces along with *frame theory* provide methods for arbitrarily reducing or increasing the dimensionality of data in useful ways. Frame theory also provides an interesting way of thinking about vector space basis.

3.3.9 ***Introduction to Frame Theory.*** Frames are *generalized bases* to sets of vectors in Hilbert space that span the vector space but are not necessarily linearly independent. This implies

that all bases are frames but not all frames are bases. Frames were originally defined by (Duffin & Schaefer, 1952) to decompose non-harmonic Fourier series. Daubechies, Grossman and Meyer (1986) related frames to wavelets, and Daubechies (1990; 1992) continued the formalization of frames. Previously when we presented bases we said that $\{(1,0), (0,1)\}$ is a basis for $L^2(\mathbb{R}^2)$. The vectors span $L^2(\mathbb{R}^2)$ and are linearly independent. The set of vectors $\{(1,0), (0,1), (\sqrt{2}/2, \sqrt{2}/2)\}$ would constitute a frame for $L^2(\mathbb{R}^2)$ but not a basis. The frame vectors spans $L^2(\mathbb{R}^2)$ but the vectors are not linearly independent from one another. The first and second elements are orthogonal to one another, but the inner product of the first and third is non-zero, and the inner product of the second and third is also non-zero. With the intertwined spirals problem we explained why it can be helpful to transform a vector from one basis to another. Frames provide a similar utility by allowing us to transform from one frame to another and providing conditions ensuring stable reconstruction from one frame to another.

Frames are extremely versatile in terms of application. To give a few examples, Support Vector Machines (SVMs) use frames to project data into higher dimensions. Data that may be non-linear in two dimensional *feature space* (a space is defined not necessarily by time or frequency but by characteristic features of the data) for instance may be linear in higher dimensional feature spaces. This is often referred to as the "Kernel Trick." A second application of frames is Principle Component Analysis (PCA). Principle component analysis is a method for reducing the dimensionality of data. PCA is a method of determining a new orthogonal basis for a set of data such that the variance is captured by the components (basis vectors) in an ordered fashion. With PCA the first component will predict the most variability and the following components will capture descending amounts of variability. When given high dimensionality data sets with large amounts of noise PCA may be able to capture a large portion of the variability with only a handful of the original components. With spectral analysis frame theory allows us to interpolate the spectra between frequency bins with discrete time Fourier transformation. The continuous short-time

Fourier transform and the continuous wavelet transform use frames to approximate the spectra over bounded intervals. Discrete wavelet transforms use frames to ensure stable reconstruction. Keep in mind these examples are just provided to highlight the utility of frames. At this juncture the reader isn't expected to understand the examples in great detail; they are merely to give the reader an idea of where we are headed.

3.3.9.1 Precisely what is a frame? To explain frame theory it is helpful to have some familiarity with mathematical transformations and experience working with inner product spaces. If the reader is unfamiliar with these concepts they are encouraged to read through sections 2.3 – 2.3.3. Recall that the Fourier series allows us to represent a time varying signal $f(x)$ in a Hilbert space $\mathcal{H}$ as a sum of complex sinusoidal basis functions and scalar coefficients (assuming we define the basis to match the domain of the function and the function is square-integratable). The basis of these components are in the form given by

$$\varphi_n \stackrel{\text{def}}{=} \left\{e^{jnx}, \text{for all } n \text{ in } \mathbb{Z}\right\}$$

and a set of scalar coefficients defined by the inner product space of f with the basis functions in φ_n ,

$$c_n \stackrel{\text{def}}{=} \int f(x)\overline{\varphi_n}dx = \langle f, \varphi_n \rangle.$$

This allows us to represent f as

$$f(x) \stackrel{\text{def}}{=} \sum_n \langle f, \varphi_n \rangle \varphi_n.$$

Now we propose that the basis φ_n could be given by given by a number of possible orthogonal basis functions or even properly chosen non-orthogonal sets of functions. For all possible functions f in $\mathrm{L}^2(\mathbb{R})$, and a set of functions φ_n a frame can be defined such that

$$A\|f\|^2 \leq \sum_n |\langle f, \varphi_n \rangle|^2 \leq B\|f\|^2$$

where A and B are scalars and $0 < A \leq B < \infty$. These constants are called the *lower and upper frame bounds* respectively. When φ_n is the Fourier basis, $\{e^{j\xi x}$, for all ξ in $\mathbb{R}\}$, or any orthonormal basis A and B will both equal 1. Additionally, we know that when φ_n is the Fourier basis it is possible to perfectly reconstruct f. But what about when φ_n is not a basis? Often times it is not possible or not desirable for φ_n to be orthogonal. In these cases how can we know whether it is possible to reconstruct f? And if it is possible, how? The simple answer is that if φ_n meets the frame condition then it is possible to reconstruct f. The details in understanding how this is possible are not overly difficult, but upon first encounter the understanding how the details fit together is a little confusing. Here we will begin with a some theory.

3.3.9.2 The Analysis and Synthesis Operators. When we transform a function we *map* the function to an inner product space. The map operation is called the *Bessel map* or *analysis operator* and using function notation is denoted as[7],

$$T^*: \mathcal{H} \rightarrow \ell^2(\mathbb{N})$$

$$T^*: f \rightarrow \{\langle f, \varphi_n \rangle\}_{n=1}^{\infty}$$

With the Bessel map the first line means that we are declaring a function T^* that has a Hilbert space input domain and has a discrete Lesbesgue-2 space target codomain. The second line defines the domain as a function f and specifies the codomain as the inner product space of f and the set of functions in φ_n.

When we take the inverse transform we have a second operator known as the pre-frame operator or synthesis operator. The synthesis operator is,

$$T: \ell^2(\mathbb{N}) \rightarrow \mathcal{H}$$

---

[7] This notation describes functions or operators. For example the function $f(x) = x^2$ would be expressed as,
$f: \mathbb{R} \rightarrow \mathbb{R}$ (declares the domain of the input as the field of real numbers and specifies the output codomain as the field of real numbers)
$f: x \rightarrow x^2$ (maps the domain x to x^2)

$$T: \{c_n\} \to \sum_n c_n \varphi_n$$

In frame theory when $A = B = 1$ we call the frame a 1-tight frame or a Parseval frame. A Parseval frame is a normalized frame. All orthonormal bases are Parseval frames but not all Parseval frames are orthonormal bases (Weisstein, Parseval's Theorem, 2011). From what we already know about orthonormal bases we can express f as:

$$f = TT^*.$$

If it isn't apparent why this is true, consider the Fourier series. Give the proper domain and basis this restates the Fourier series.

The frame can be expressed in matrix form where each vector is a row,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$

This is matrix representation is also known as the *frame operator*.

3.4 Fourier Analysis

We begin this section by asking "What is a periodic signal?" A periodic signal is one that repeats itself over time. The length of time before it repeats is called its period. So in mathematical notation we can define a function $f(t)$ periodic if a T exists such that $f(t+T) = f(t)$ for all t. For example, the function $f(t) = \cos(3\pi t)$ is periodic when $T = 2\pi/3\pi$:

$f(\mathrm{t}) = \mathrm{f}(t + 2\pi/3\pi).$

$\cos(3\pi t) = \cos(3\pi(t + 2\pi/3\pi))$

$\cos(3\pi t) = \cos(3\pi t + 6\pi^2/3\pi)$

$\cos(3\pi t) = \cos(3\pi t + 2\pi)$

As a corollary to our definition of a periodic function we can define an aperiodic function if a T does not exists such that $f(t+T) = f(t)$ for all t. For example the Dirac delta function defined as:

$$\delta(t) = \begin{cases} +\infty, & t = 0 \\ 0, & t \neq 0 \end{cases}$$

is obviously aperiodic. Through the course of this section we will show that any signal, whether period or aperiodic, can be represented as a linear combination of periodic signals. In 1807 Fourier's discovery of what we now know as Fourier series essentially provides an algorithm for representing periodic functions as an infinite sum of sine and cosine functions. Breaking up a function into simpler terms can in many cases yield closed form analytic solutions.

3.4.1 *Introduction to the Fourier Series.*

Fourier series decompose functions which are periodic over the domain $[-\pi, \pi]$. For example let's consider the square wave defined by :

$$f(x) = \begin{cases} 0, & \text{if } -\pi < x \leq 0 \\ 1, & \text{if } \quad 0 < x \leq \pi \end{cases}$$

when $f(x + 2\pi) = f(x)$. **3.4.1**

The square wave function, or any function $f(x)$ with a domain bounded by $[-\pi, \pi]$, can be represented using the trigonometric Fourier Series (Weisstein, 2011):

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)] \qquad \textbf{3.4.2}$$

where,

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)dx$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)dx$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx$$.[8]

The *n* parameter iterates through $\mathbb{Z}$ (the set of all integers). As *n* increases the summation yields better and better approximations of the original function. The a_0 coefficient provides a constant offset to the series representation. The coefficients a_n and b_n weight the $\cos(nx)$ and $\sin(nx)$ functions within the summation. One of the handy things about Fourier series is that we can solve the integrals to hopefully find a less cumbersome representation for the coefficients. Remember, these techniques were invented long before the advent of computers. Nowadays if we can specify an equation a computer can solve it, but when everything was done with pencil and paper reducing the amount of work one has to do is critical. When we solve the integrals[9] defining a_0, a_n and b_n for the square-wave function defined in Equation 3.4.1 we see that:

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{0} 0dx + \frac{1}{\pi}\int_{0}^{\pi} 1dx = 0 + \frac{1}{\pi}\pi = 1$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{0} 0\cos(nx)dx + \frac{1}{\pi}\int_{0}^{\pi} 1\cos(nx)dx = 0 + \frac{\sin(\pi n)}{n\pi}$$

---

[8] In some disciplines $F(x)$ is also denoted as $f(x)$. I find this adds confusion because it makes the coefficients seem like they should be recursive. Here I use $f(x)$ to refer to the original time signal and use $F(x)$ to refer to the Fourier series representation. The reader should keep in mind the Fourier series completely represents $f(x)$, so $f(x) = F(x)$.

[9] We are treating the integrals piecewise. When we integrate from $-\pi$ to 0 we substitute 0 for $f(x)$ (since $f(x)$ is 0 over this domain). Likewise, when we integrate over 0 to π we substitute 1 for $f(x)$. The Fourier series can decompose any piecewise periodic function (on the interval – π to π) in this manner. Readers who aren't particularly fond of finding integrals by hand should be aware of WolframAlpha.com. Example syntax: "integrate sin(n x) dx, 0<x<pi"

$$b_n = \frac{1}{\pi}\int_{-\pi}^{0} 0\sin(nx)dx + \frac{1}{\pi}\int_{0}^{\pi} 1\sin(nx)dx = 0 + \frac{1-\cos(\pi n)}{n\pi}$$

Since these coefficients are inside the summation we can think of them as functions of *n*. Let's see what happens to the coefficients of the first example as *n* increases.

$$a_1 = \sin(\pi)/\pi = 0$$
$$a_2 = \sin(2\pi)/2\pi = 0$$
$$a_3 = \sin(3\pi)/3\pi = 0$$
$$a_4 = \sin(3\pi)/4\pi = 0$$
$$\vdots$$
$$b_1 = (1-\cos(1\pi))/\pi = 2/\pi$$
$$b_2 = (1-\cos(2\pi))/2\pi = 0$$
$$b_3 = (1-\cos(3\pi))/3\pi = 2/3\pi$$
$$b_4 = (1-\cos(4\pi))/4\pi = 0$$
$$b_5 = (1-\cos(5\pi))/5\pi = 2/5\pi$$
$$b_6 = (1-\cos(6\pi))/6\pi = 0$$
$$b_7 = (1-\cos(7\pi))/7\pi = 2/7\pi$$
$$\vdots$$

So we can see through induction that the a_n coefficients always become zero becomes $\sin(x)$ always crosses the x-axis at each pi interval. The b_n coefficients are always zero when *n* is even and become $2/n\pi$ when *n* is odd. With this information in hand we can redefine a_n and b_n as:

$$a_n = 0$$

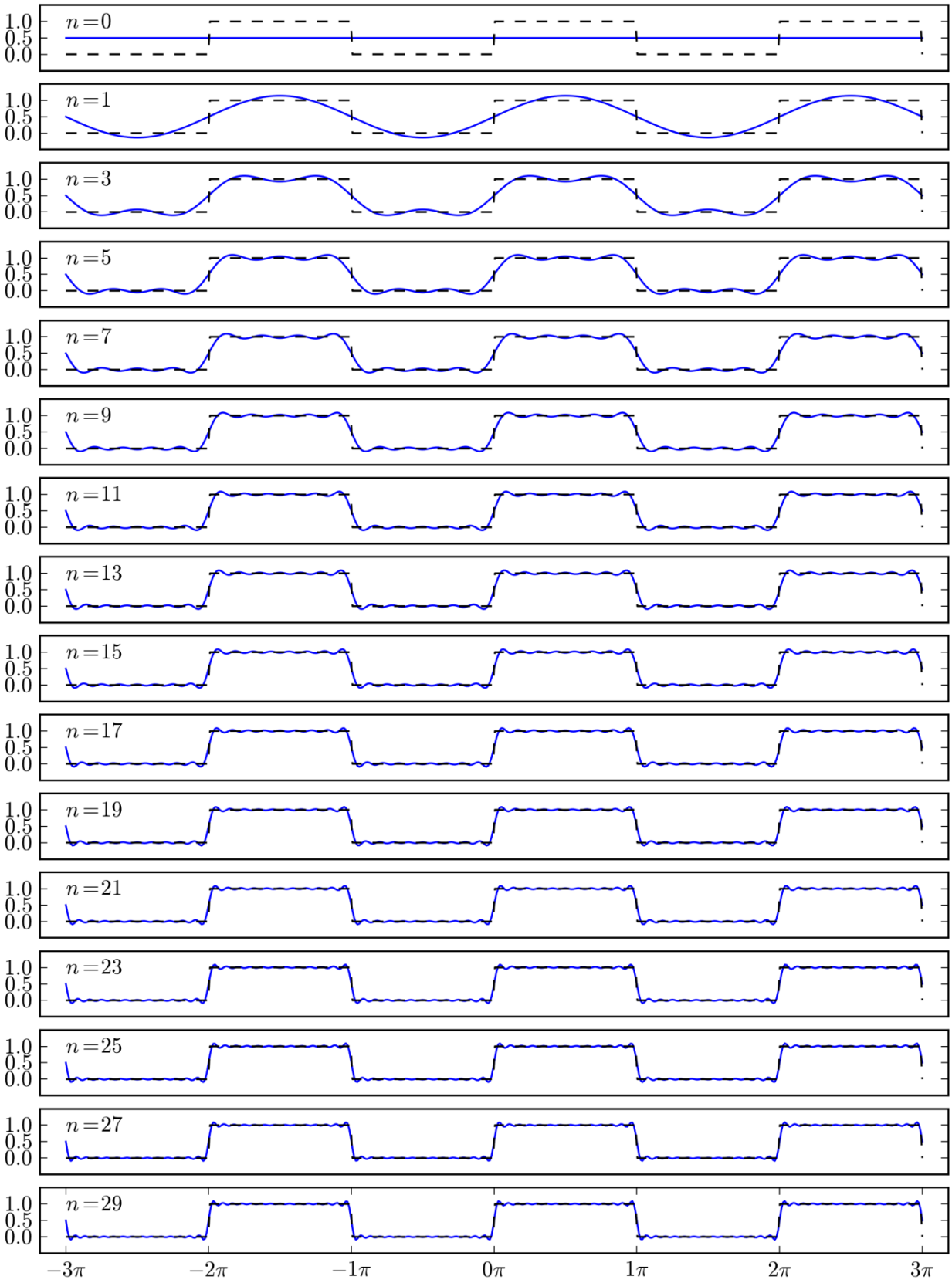
$$b_n = \begin{cases} 0 & n \text{ is even} \\ 2/n\pi & n \text{ is odd} \end{cases}$$

Now when we plug these into $F(x)$ we see the equation is greatly simplified:

$$F(x) = \frac{1}{2} + \sum_{n=1,3,5,7,..}^{\infty} \frac{2}{n\pi}\sin(nx)\,.$$

Figure 3.4.1 plots the first 16 approximations of $F(x)$. As the reader can see even after only a handful of iterations the function is a visually close approximation to the original function. At present this is not much more than a parlor trick. Upcoming sections will discuss why and how it

Figure 3.4.1 *Fourier series approximations of a square wave.*



actually works in more detail.

3.4.1.1 Euler's Formula. Euler's formula which describes a relation between the trigonometric functions and the complex natural logarithm. Euler's formula states that for and real value x,

$$e^{jx} = \cos(x) + j\sin(x).$$

Various proofs of this relationship are possible. One of the more popular shows that the power series of e^{jx} matches the Taylor series of $\cos(x) + j\sin(x)$. First let us define the power series for e^{z} as:

$$e^{z} = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

Now we can substitute jx for z:

$$e^{jx} = 1 + \frac{(jx)}{1!} + \frac{(jx)^2}{2!} + \frac{(jx)^3}{3!} + \frac{(jx)^4}{4!} + \frac{(jx)^5}{5!} + \frac{(jx)^6}{6!} + \frac{(jx)^7}{7!} + \cdots$$

Recall that:

$$j^0 = 1,\ j^1 = j, j^2 = -1, j^3 = -j,$$

$$j^4 = 1,\ j^5 = j, j^6 = -1, j^7 = -j,\ldots$$

So the power series can be simplified to:

$$e^{jx} = 1 + \frac{jx}{1!} - \frac{x^2}{2!} - \frac{jx^3}{3!} + \frac{x^4}{4!} + \frac{jx^5}{5!} - \frac{x^6}{6!} - \frac{jx^7}{7!} + \cdots$$

Now we can group the *j* terms together and group the non *j* terms together:

$$e^{jx} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\right) + j\left(\frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right)$$

From here we can recognize the power series for $\cos(x)$ and for $\sin(x)$ in the respective parentheses and the series can be expressed as:

$$e^{jx} = \cos(x) + j\sin(x)$$

Equation 3.3.1 defined how a vector can be defined as a complex number or an angle and magnitude. The identity can be extended to:

$$a + jb = r[\cos(\alpha) + j\sin(\alpha)] = re^{j\alpha}$$

This representation of a complex number is known as a phase vector (complex phasor, or phasor). Recall from above r denotes the magnitude, and α specifies the phase in radians from the positive real axis in the clockwise direction. With the phasor notation trigonometric problems can be solved using algebra of exponents.

3.4.1.2 Phase Vectors (Phasors). In the previous section we showed how Euler's formula could be derived from the e^z power series by substituting jx for z. It is often more convenient to think about time varying signals in terms of their frequency (f_0) in cycles per second. By substituting $j2\pi f_0 t$ for z we can express sinusoidal waves, with frequency specified by f_0, as functions of time (t):

$$e^{j2\pi f_0 t} = \cos(2\pi f_0 t) + j\sin(2\pi f_0 t)$$

If $-jx$ is substituted for z a second relation is found:

$$e^{-j2\pi f_0 t} = \cos(2\pi f_0 t) - j\sin(2\pi f_0 t)$$

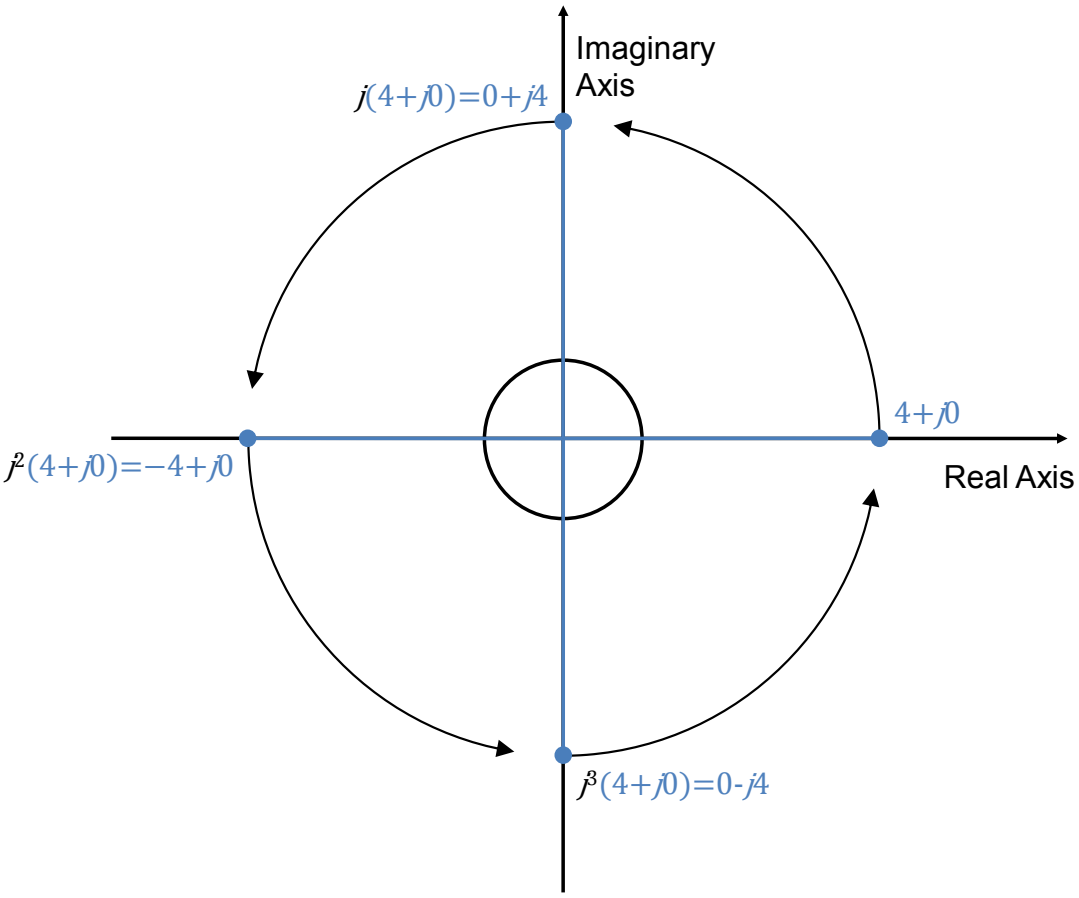
With these two formulas and some algebraic manipulation the following identities can be obtained:

$$\cos(2\pi f_0 t) = \frac{e^{j2\pi f_0 t}}{2} + \frac{e^{-j2\pi f_0 t}}{2} \text{ and} \qquad \textbf{3.4.3}$$

$$\sin(2\pi f_0 t) = \frac{je^{-j2\pi f_0 t}}{2} - \frac{je^{j2\pi f_0 t}}{2} \qquad \textbf{3.4.4}$$

To understand what is going on here let us examine the phasors associated with the function $2\cos(2\pi t)$. From the equation we can see the wave has an amplitude of 2 and a frequency of 1 Hz. These values simplify the first term to $e^{j2\pi t}$and the second term to $e^{-j2\pi t}$. On the complex plane the first phasor will follow the unit circle on the complex plane in the counter-clockwise (CCW) direction. The second phasor follows the unit circle in the clockwise (CW) direction. Because $f_0 = 1$

Figure 3.4.2 *The j-operator. The j-operator. shifts a vector or a function by 90° counterclockwise in the complex plane. Karl Gauss called the j-operator "the shadow of shadows."*



in the phasors will each travel around the unit circle one time every second. As these phasors travel around the unit circle their vector sums are such that the imagery parts cancel one another and the real parts sum such that the real parts equal $2\cos(2\pi t)$**.**

Figure 3.4.3 *2cos(2πt) can be represented by the sum of two phasors: ej2πt (blue trace) and e-j2πt (green trace). In the imaginary axis the phasors are anti-phase and cancel one another when summed (the dotted traces are all on the imaginary plane), while the real axis the phasors are in phase and when summed compose 2cos(2πt).*

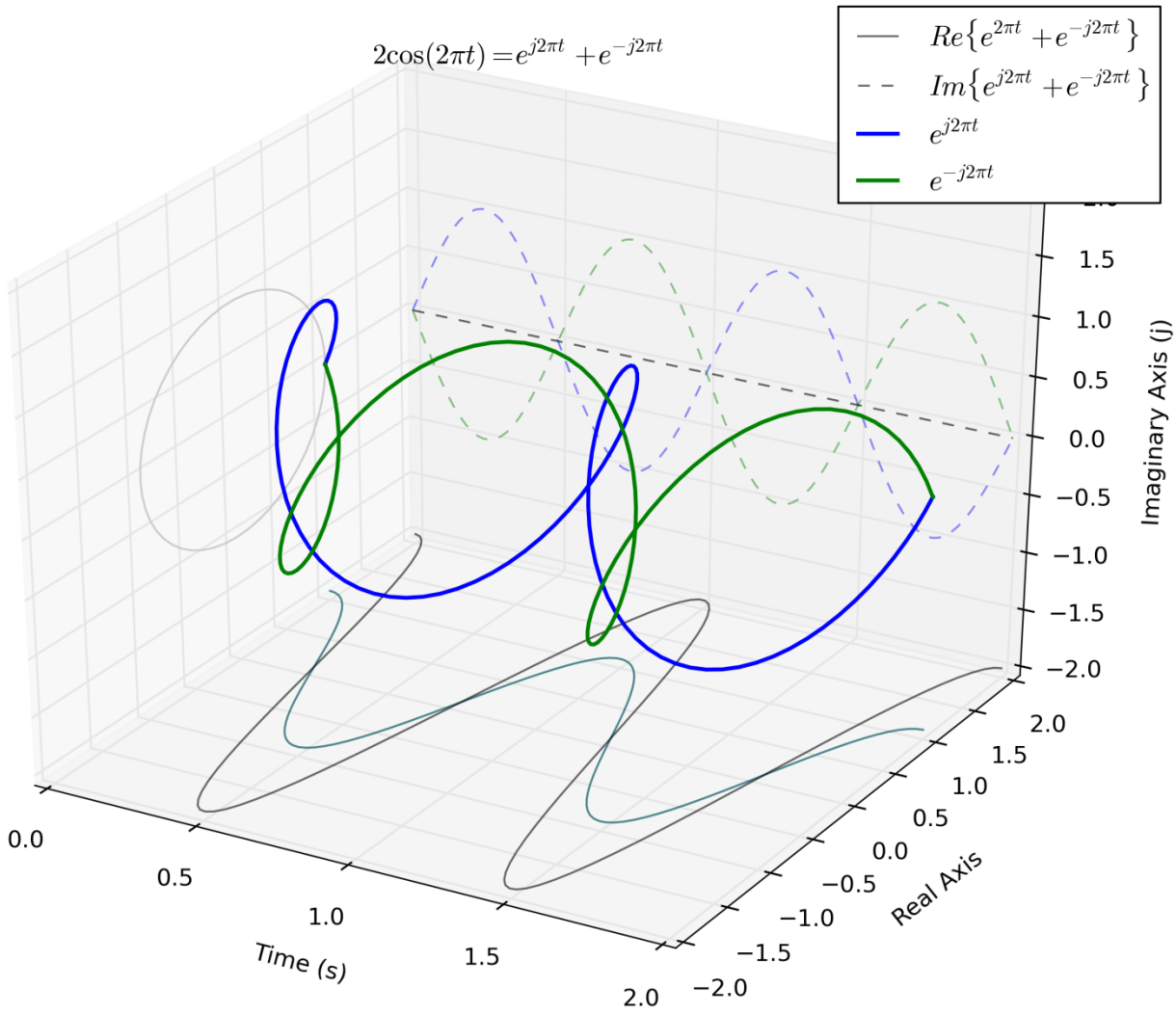
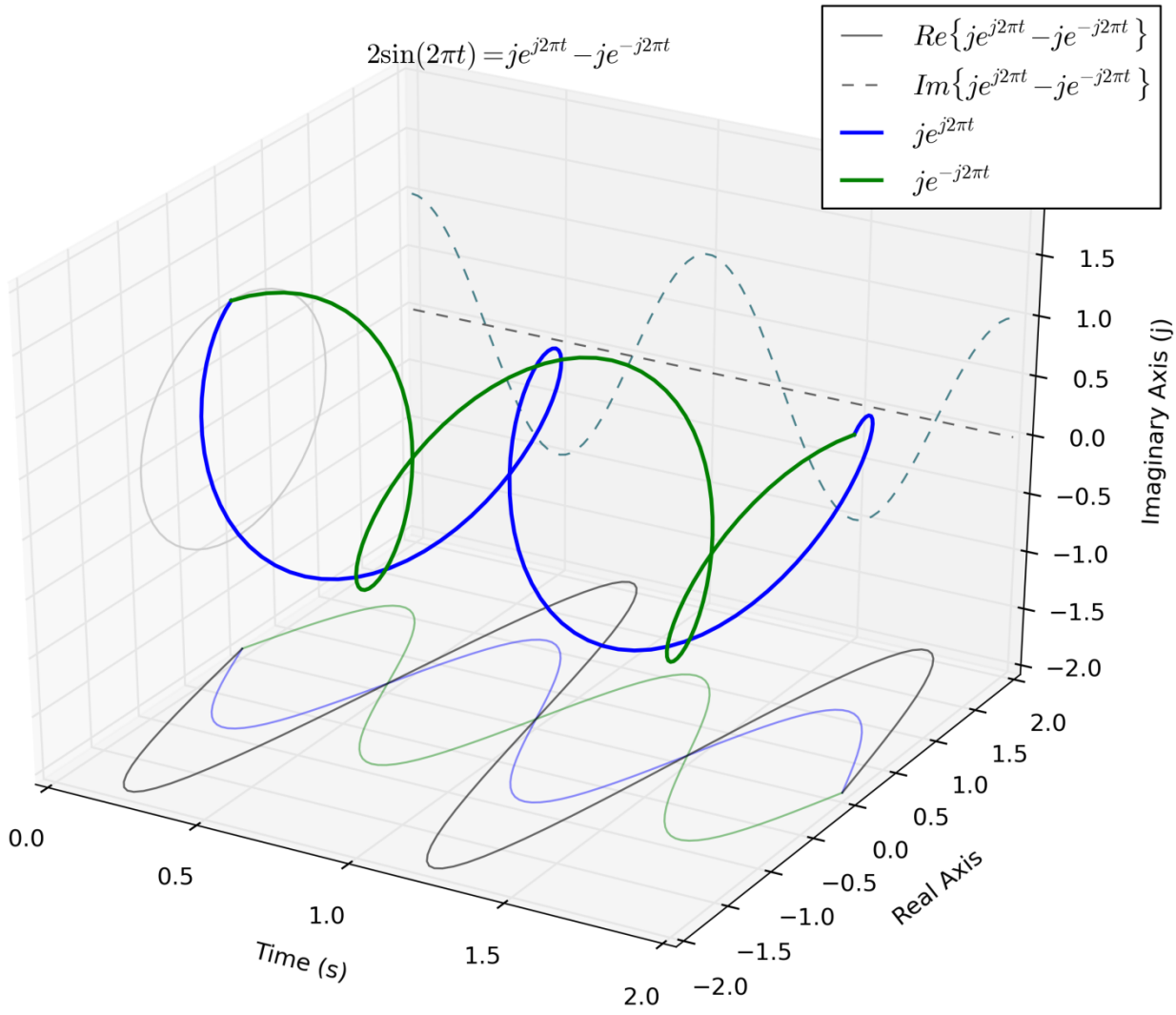


Figure 3.4.4 *2sin(2πt) can be represented as the difference between phasors: jej2πt (blue trace) and je-j2πt (green trace) The leading j terms rotate the phasors 90° counter-clockwise about the unit circle . In the imagery plane the phasors are identical so their difference is 0. In the real plane the difference between the phasors reaches a maximum when t = 0.5 and 1.5 (1/4 and 3/4 of the sine's period).*



See Figure 3.4.2.[10]

Now let us examine the phasors associated with $2\sin(2\pi t)$ (Equation 3.4.4). From the identity above we can see that the phasors become $je^{-j2\pi t}$ and $je^{j2\pi t}$. We can also see that instead of summing the second phasor is subtracted from the first. The *j* term is known as the j-operator. Say we have a complex number $\mathbf{c} = 4 + j0$. When we multiply **c** by *j* the result result is 0 + *j*4. When we multiply **c** by j^2 the result is -4 + j0 because j^2=-1. When we multiply **c** by j^3 the result is 0 – 4*j,* and lastly **c** times j^4 brings us back to c (see Figure 3.4.3). On the complex plane we can see that multiplying a phasor by *j* rotates the phasor by 90° in the CCW direction. This is the *j-operator* (Lyons, 2008). Through Euler's formula we can also see that since $e^{j\pi/2} = \cos\left(\frac{\pi}{2}\right) + j\sin\left(\frac{\pi}{2}\right) = 0 + j$ multiplying a phasor by $e^{j\pi/2}$ has the same effect. Figure 3.4.4 depicts the phasors associated with $2\sin(2\pi t)$ The j-operators cause the imaginary parts to remain equivalent over time while the difference between the real parts equal $2\sin(2\pi t)$. Now that we have an idea of how phasors can represent harmonic signals we can begin to examine how the Fourier transformation extracts the frequency content from a time-varying signal.

3.4.1.3 The Exponential Fourier Series. Now that we have some experience working with phasors we can represented the trigonometric Fourier Series presented in Equation 3.4.2 as:

$$F(x) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{n=-\infty}^{\infty} c_n e^{jnx} \qquad \textbf{3.4.5}$$

where,

$$c_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)e^{-jnx}dx \quad .$$

[10] These phasor representations are not universally appealing. They are a bit easier to decipher when viewed interactively and one can orbit the axes to perceive the depth relations of the spirals.

Figure 3.4.3 *2cos(2πt) can be represented by the sum of two phasors: $e^{j2\pi t}$ (blue trace) and $e^{-j2\pi t}$ (green trace). In the imaginary axis the phasors are anti-phase and cancel one another when summed (the dotted traces are all on the imaginary plane), while the real axis the phasors are in phase and when summed compose 2cos(2πt).*

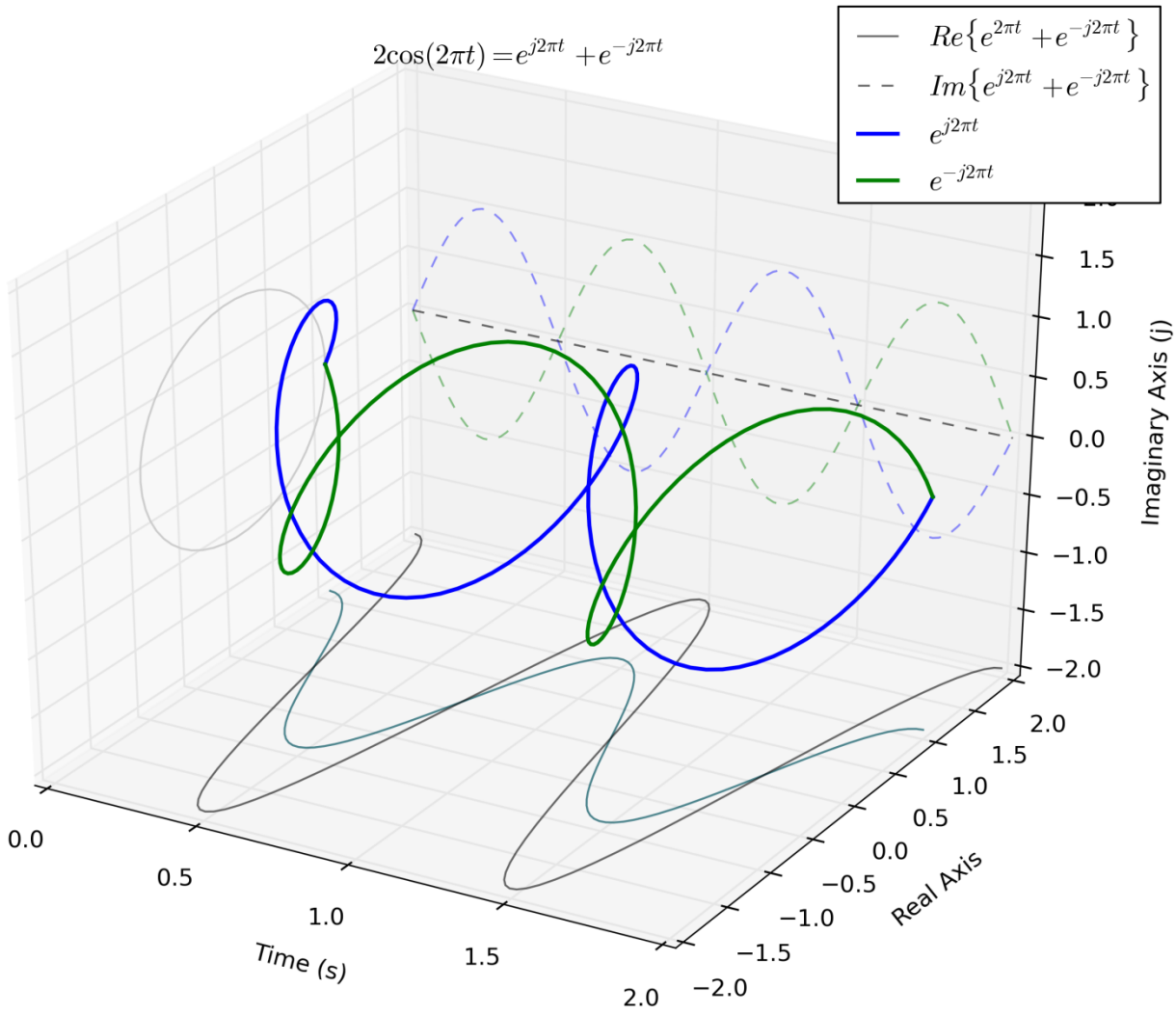
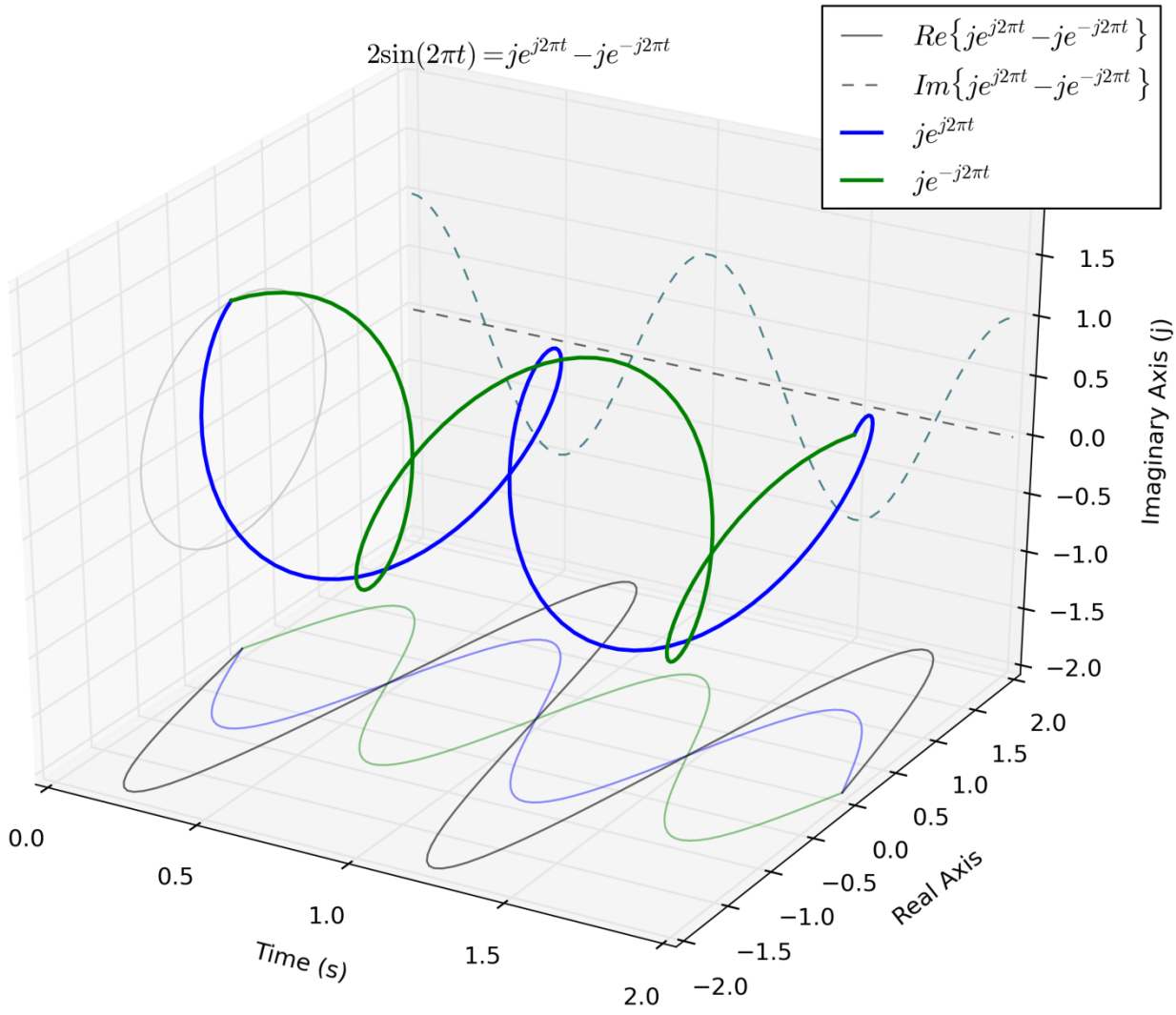


Figure 3.4.4 *2sin(2πt) can be represented as the difference between phasors: $je^{j2\pi t}$ (blue trace) and $je^{-j2\pi t}$ (green trace) The leading j terms rotate the phasors 90° counter-clockwise about the unit circle . In the imagery plane the phasors are identical so their difference is 0. In the real plane the difference between the phasors reaches a maximum when t = 0.5 and 1.5 (1/4 and 3/4 of the sine's period).*



How wonderfully succinct! This is possible by taking advantage of symmetries of cosine and sine:

$$\cos(x) = \cos(-x)$$

$$\sin(-x) = -\sin(x) \qquad .$$

These identities reveal symmetry of the coefficients of the trigonometric Fourier Series:

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(-nx)dx = a_{-n}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx = \frac{1}{\pi}\int_{-\pi}^{\pi} -f(x)\sin(-nx)dx = -\,b_{-n}$$

$$b_{-n} = -b_n$$

Our goal is to replace $\cos(nx)$ and $\sin(nx)$ with e^{jnx}. If we define the Fourier coefficients as a complex scalar

$$c_n \stackrel{\text{def}}{=} a_n - jb_n$$

$$= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)dx - j\frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx$$

$$= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)[\cos(nx) - j\sin(nx)]dx$$

Euler's tells us that $\cos(nx) - j\sin(nx) = e^{-jnx}$ so after substitution

$$c_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)e^{-jnx}dx.$$

The symmetries of cosine and sine are still embedded in this coefficient. The consequence is c_{-n} is identical to the complex conjugate of c_n

$$c_{-n} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)e^{-j(-n)x}dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)e^{jnx}dx = \overline{c_n}.$$

Now let's see what happens when we expand the Exponential Fourier Series (Equation 3.4.5)

$$F(x) = \frac{1}{2}\left[(c_0e^0) + \left(c_1e^{jx} + \overline{c_1}e^{-jx}\right) + \left(c_2e^{j2x} + \overline{c_2}e^{-j2x}\right) + \cdots\right]$$

When we substitute in for our a_0, a_n and b_n for our c coefficients and expand only $\left(c_1 e^{jx} + \overline{c_1} e^{-jx}\right)$ we can obtain the Trigonometric Fourier Series:

$$F(x) = \frac{1}{2}\left[(a_0) + \left((a_1 - jb_1)e^{jx} + (a_1 + jb_1)e^{-jx}\right) + \cdots\right]$$

$$= \frac{1}{2}\left[(a_0) + \left(a_1 e^{jx} - jb_1 e^{jx} + a_1 e^{-jx} + jb_1 e^{-jx}\right) + \cdots\right]$$

$$= \frac{1}{2}\left[(a_0) + \left(a_1(e^{jx} + e^{-jx}) + b_1(je^{-jx} - je^{jx})\right) + \cdots\right]$$

$$= \frac{1}{2}\left[(a_0) + \left(2a_1\left(\frac{e^{jx}}{2} + \frac{e^{-jx}}{2}\right) + 2b_1\left(\frac{je^{-jx}}{2} - \frac{je^{jx}}{2}\right)\right) + \cdots\right]$$

Here we can recognize Euler's formulas for cosine and sine, and after substitution we have:

$$F(x) = \left(\frac{a_0}{2}\right) + (a_1 \cos(x) + b_1 \sin(x)) + (a_2 \cos(2x) + b_2 \sin(2x)) + \cdots.$$

The reader should be able to finish the induction from here if they still need more convincing.

The $\frac{1}{2}$ term from outside the summation to inside c_n. It makes the presentation a little nicer and is more consistent with existing definitions.

$$F(x) = \sum_{n=-\infty}^{\infty} c_n e^{jnx}, \quad \text{where } c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-jnx}dx.$$

3.4.1.4 The Complex Fourier Coefficients Define a Hilbert Space. If the reader will recall the inner product of continuous complex functions $f(x)$ and $g(x)$ is defined as

$$\langle f, g\rangle \stackrel{\text{def}}{=} \int_a^b f(x)\overline{g(x)}.$$

If the inner product space is complete and normed such that $\|f\| = \sqrt{\langle f, f\rangle}$ the inner product space is a Hilbert space. With a little imagination we can see that the c_n formula from the exponential Fourier series closely resembles our continuous inner product definition. In fact, c_n actually defines an inner product space. To make it a little more obvious let's define a function

$e_n \stackrel{\text{def}}{=} e^{jnx}$.[11]

This allows us to define what was previously c_n as an inner product space

$$\langle f, e_n \rangle \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{e_n}dx,$$

and the Fourier series becomes

$$F(x) \stackrel{\text{def}}{=} \sum_{n=-\infty}^{\infty} \langle f, e_n \rangle e_n.$$

Let's step back and think through what is happening here. While it might not be immediately apparent, the inner product between $\langle f, e_n \rangle$ is expressing the amount of similarity between $f(x)$ and a complex sinusoid with a frequency of *n*.[12] Therefore, the space of all the inner products, the Hilbert space, $\langle f, e_n \rangle$ reflects the amount of similarity between $f(x)$ at all frequencies in $\mathbb{Z}$. The inner product is transforming $f(x)$ from the time domain to the frequency domain. The Hilbert space represents $f(x)$ in the frequency domain. To re-state this idea we can say that the inner product encodes $f(x)$ as a Hilbert space. When the Fourier series calculates $\langle f, e_n \rangle e_n$ it is decoding the Hilbert space' frequency domain representation back to a time domain representation.

3.4.1.5 Parseval's Identity. Parseval's Identity shows that the sum of the squared coefficients of the Fourier series is equal to the integral of $f(x)$ squared (Weisstein, 2011f),

$$\sum_{n=-\infty}^{\infty} |e_n|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2 dx.$$

The implication of Parseval's Identity is that the amount of energy in the signal is maintained from one representation to the other in a manner that guarantees reproducibility.

3.4.1.6 The Fourier Coefficients have an Orthonormal Basis. One of the key concepts that makes the Fourier series work is that the e_n terms are orthogonal to one another. If they were not

[11] The complex conjugate of e_n is e^{-jnx}

[12] This idea is elaborated on in future sections.

orthogonal it would imply that frequencies in $f(x)$ exist that are dependent on other frequencies in $f(x)$ and we would not be able to treat them as independent sinusoids. We can define the basis as a set

$$B \stackrel{\text{def}}{=} \{e_n = e^{jnx}, \text{for all } n \text{ in } \mathbb{Z}\}.$$

To show that these are indeed a basis we need to show that each element is orthogonal to all elements of the set except themselves over $L^2[-\pi, \pi]$. Recall that two vectors are orthogonal to one another if their inner product is zero:

$$\mathbf{u} \perp \mathbf{v} \text{ if } \langle \mathbf{u}, \mathbf{v} \rangle = 0.$$

Since there are an infinite number of vectors it really isn't possible to check all $\frac{\infty!}{2!(\infty-2)!}$ combinations. Instead we can show that $\langle e_n, e_m \rangle = 0$ for any *n* and *m* in $\mathbb{Z}$ as long as $n \neq m$.

$$\langle e_n, e_m \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e_n \overline{e_m} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{jnx} e^{-jmx} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{(n-m)jx} dx$$

Whenever $n \neq m$ the difference between *n* and *m* will always be an integer since $\mathbb{Z}$ is closed on addition (adding and subtracting integers will always result in another integer). We can define this difference as $k \stackrel{\text{def}}{=} n - m$. The consequence of this closure is that

$$\langle e_n, e_m \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{jkx} dx = 0, \qquad \text{when k} \neq 0.$$

Why?

When $k = 1$ or -1 the complex phasor goes through exactly 1 cycle over $L^2[-\pi, \pi]$.

When $k = 2$ or -2 the complex phasor goes through exactly 2 cycles over $L^2[-\pi, \pi]$.

When $k = 3$ or -3 the complex phasor goes through exactly 3 cycles over $L^2[-\pi, \pi]$.

⋮

As long as the phasor goes through full cycles over the bounds of the integral the real and imaginary positive parts will cancel out the real and imaginary negative parts and the integral results in 0. Now let's see what happens when $n = m$. In this case the inner product does not equal zero because:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{0jx}dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} 1\,dx = \frac{1}{2\pi}2\pi = 1.$$

Since $\langle e_n, e_m\rangle = 0$ for any n and m in $\mathbb{Z}$ as long as $n \neq m$ and we can conclude the elements of B are indeed orthogonal to one another.

Now we can go on to show that they are all unit vectors (vectors with a length of 1). In Hilbert space the length of a vector $\mathbf{x}$ is given by its norm $\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$. We just showed that the inner product of e_n with itself is 1 (this is what happens when $n = m$) and $\sqrt{1} = 1$ so we can conclude the elements of B are indeed unit vectors, and B is therefore not only a basis it is an orthonormal basis for the Hilbert space. The basis B we have been discussing is one of many possible bases for the Fourier series. Alternative bases can be obtained by changing how the inner product is defined. For instance, we could define the inner product as

$$\langle f, e_n\rangle \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_0^{2\pi} f(x)\overline{e_n}dx\,, \qquad \text{where } B = \{e_n = e^{jnx}, \text{for all } n \text{ in } \mathbb{Z}\} \text{ for } L^2[0,2\pi].$$

Or we could also define the inner product as

$$\langle f, e_n\rangle \stackrel{\text{def}}{=} \int_0^1 f(x)\overline{e_n}dx\,, \qquad \text{where } B = \{e_n = e^{j2\pi nx}, \text{for all } n \text{ in } \mathbb{Z}\} \text{ for } L^2[0,1].$$

Or we can generalize this by defining the inner product in terms of the period length T

$$\langle f, e_n\rangle \stackrel{\text{def}}{=} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\overline{e_n}dx\,, \text{where } B_T = \{e_n = e^{j2\pi nx/T}, \text{for all } n \text{ in } \mathbb{Z}\} \text{ for } L^2[-\pi,\pi].$$

The point of being able to modify the period is that we can make it fit the time units we want to use. We could have T equal 1 second, or have T equal one year. We can make the Fourier series fit what we are interested in instead of having to think in terms of radians. To conclude our discussion of Fourier series I shall point out that Fourier series are intended to examine Periodic signals. What if we have signals that are aperiodic or signals that we think might be periodic but we don't know the time interval they are periodic over? As we will soon see the Fourier transform will address these questions.

3.4.2 ***Introduction to the Fourier Transform.*** The Fourier transform is in many ways similar to the exponential Fourier series. The importance of the Fourier transform is that it is able to fully represent any aperiodic or periodic signal as an infinite sum of sinusoidal functions. In the previous section we demonstrated how we could modify the Fourier series to suit the periodicity of the signals we are interested in representing. The Fourier transform essentially assumes the time domain has no periodicity. In the conclusion of the previous section we showed that the Fourier coefficients can be generalized over any period length *T,*

$$\langle f, e_n \rangle \stackrel{\text{def}}{=} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\overline{e_n}dx.$$

The Fourier transform generalizes the Fourier coefficient definition by taking the limit of T to infinity. With the Fourier series the $\frac{1}{T}$ term ensures the basis functions have unit length. But, the $\lim_{T\to\infty}\frac{1}{T}$ is 0 which makes the inner product terribly uninformative. To resolve this limit mathematical analysis[13] is required.

Suppose $\omega_0 = \frac{2\pi}{T}$.

The difference between harmonic frequencies can then be defined as,

$$\Delta\omega = (n+1)\omega_0 - n\omega_0 = \omega_0.$$

Substituting into the basis gives,

$$e_n = e^{j2\pi nx/T} = e^{jn\omega_0 x}.$$

Now, as T becomes large $\Delta\omega$ becomes a differential separator,

$$\Delta\omega \longrightarrow \mathrm{d}\omega.$$

[13] Analysis is a branch of mathematics that concerns itself with transfinite operations and involves procedures of making successive approximations. Without analysis, ∞ tends to "break the field." The branch was developed largely in response to Berkeley's 1734 work "The Analyst: A Discourse Addressed to an Infidel Mathematician." The "infidel" in question was Edmund Halley who is credited for calculating the orbit of the Halley comet. Berkeley objected to the use of use of 0 ÷ 0 used in Newton's limit argument. Mathematics was forced to invent Analysis in order to save Newton's Calculus. This took a little over a 100 years to fully address Berkeley's objection (Wells, Personal Communication, 2012).

Furthermore,

$$\frac{1}{T} = \frac{\omega_0}{2\pi} = \frac{\Delta\omega}{2\pi} \longrightarrow \frac{\mathrm{d}\omega}{2\pi} \text{ and}$$

$$\langle f, e_n \rangle = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)e^{-jn\omega_0 x}dx.$$

Let $\omega = n\omega_0$. Then

$$\langle f, e_n \rangle \longrightarrow \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)e^{j\omega x}dx$$

and

$$T \cdot \langle f, e_n \rangle \longrightarrow \int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)e^{-j\omega x}dx \longrightarrow \int_{-\infty}^{\infty} f(x)e^{-j\omega x}dx \text{ as } T \longrightarrow \infty$$

The Fourier transform merely replaces $\langle f, e_n \rangle$ by $T \cdot \langle f, e_n \rangle$ as $T \longrightarrow \infty$ and $T \cdot d\omega \longrightarrow 2\pi$ as $T \longrightarrow \infty$. The analysis demonstrates that the basis length appropriate for differential frequency spacings is $T \cdot \langle f, e_n \rangle$ because as $T \longrightarrow \infty$, $\langle f, e_n \rangle \longrightarrow 0$ but $T \cdot \langle f, e_n \rangle \longrightarrow 1$. Therefore, the Fourier transform merely becomes another form of basis function expression that is analytic for $T \longrightarrow \infty$.

The Fourier transform can then be defined as,

$$X(\omega) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,\mathrm{d}t.$$

The parameter ω in angular frequency (specified in radians/unit time). The relation between angular frequency (cycles/unit time) is $\omega = 2\pi f$ where f is ordinary frequency. The function $x(t)$ is our original time-varying function. The function $X(\omega)$ is our frequency-varying function. The term $e^{-j\omega t}$ is a phasor describing a complex sinusoid of frequency ω. The integral is over $-\infty$ to ∞ with respect to t. Since the $e^{-j\omega t}$ corkscrews over all of time the function $x(t)$ must have finite energy in order for the integral to converge. That is to say:

$$\int_{-\infty}^{\infty} |x(t)|^2\,\mathrm{d}t < \infty.$$

3.4.2.1 Fourier Transform has a Complete Basis. The fact that the Fourier transform is unbounded means the basis must be able to span all possible functions definable over $L^2[-\infty, \infty]$ (or more simply $L^2\mathbb{R}$). To be able to do this the basis set must be able to specify all cosine and sine waves with infinite precision. In other words the basis must be mathematically complete. The basis for the Fourier transform is given by,

$$B = \left\{e^{j\omega t}, \text{for all } \omega \text{ in } \mathbb{R}\right\} \text{ for } L^2\mathbb{R}.$$

At this point the reader is probably wondering how can functions which are nearly identical to one another be orthogonal to one another? The insight is in understanding that the integral defining the inner product is infinite. Two complex exponentials e^{jt} and $e^{j1.000000000000000000000000000001t}$ won't be orthogonal to one another if the inner product integral is bounded about zero, but if the inner product integral is unbounded they in fact are orthogonal as they drift out of phase over infinite time in both directions (see Figure 3.4.5).

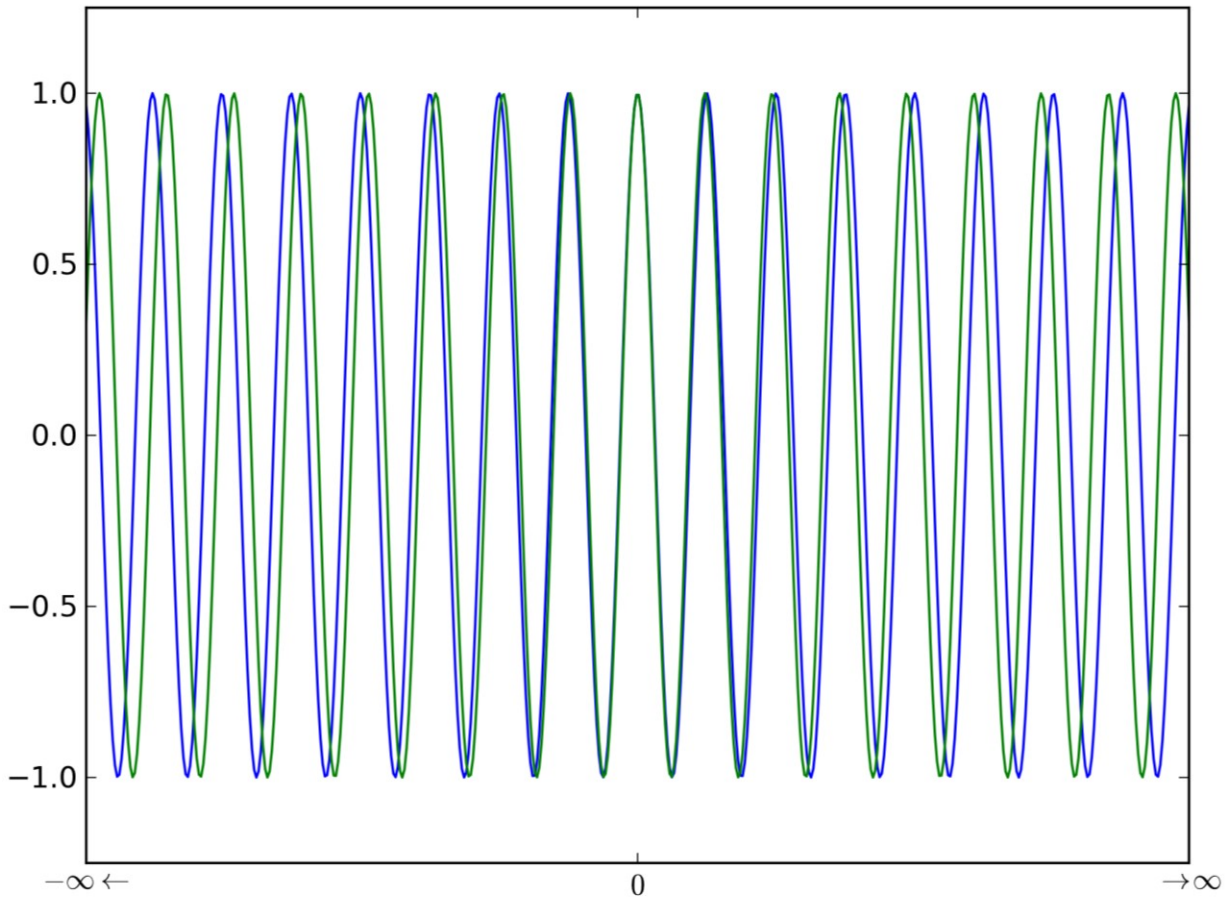
3.4.2.2 The Inverse Fourier Transform. A consequence of having a complete basis set is that recovering the original function $x(t)$ requires integration. What was formally the Fourier series summation becomes the inverse Fourier transform,

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} \, d\omega.$$

Scaling the integral by $\frac{1}{2\pi}$ maintains the power in the original signal. The Fourier transform presented above is non-unitary. Unitary transforms normalize the output such that it contains power equal to that of the input. The Fourier transform and its inverse could be presented as unitary transforms by applying a scaling factor (the term outside the integral) of $\frac{1}{\sqrt{2\pi}}$ to both transform and inverse transform. [14]

[14] Here we present the non-unitary transform because as we will see the discrete Fourier transform is usually presented as non-unitary which makes essentially almost all fast Fourier transform algorithms non-unitary.

Figure 3.4.5 *The Fourier transform has a complete basis space. Basis functions may look similar about 0 but drift out of phase as they approach $\pm\infty$.*



3.4.2.3 Plancherel's Theorem. Plancherel's theorem holds that the integral of the square of the X(ω) is equal to the integral of the square of x(t) (Weisstein, Plancherel's Theorem, 2011) [After norming non-unitary transformations of course.],

$$\int |X(\omega)|^2 \, d\omega = \int |x(t)|^2 \, dt$$

The has similar implications as Parseval's identity for the Fourier series in that the total amount of energy across the time domain is equal to the total energy across the spectral domain. This is also referred to as Rayleigh's energy theorem (Rayleigh, 1889). Now that we have the introduction out of the way we can start looking at what exactly is occurring when a function is multiplied by a complex sinusoid and integrated.

3.4.2.4 Illustrating the real and imaginary parts of the transform. It is often common practice to drop the bounds when they extend from over $-\infty$ to ∞ and it reduces the visual clutter so we will follow this convention where it aids the presentation.

Using Euler's Formula the complex integral can be broken into two definite integrals:

$$\begin{aligned} X(\omega) &= \int x(t) e^{-j\omega t} dt \\ &= \int x(t)[\cos(\omega t) - j\sin(\omega t)]dt \\ &= \int x(t)\cos(\omega t)\, dt - j\int x(t)\sin(\omega t)\, dt \end{aligned}$$

From this we can see that the Fourier transform returns a complex number. The integral with the cosine function calculates the real part and the integral with the sine function calculates the imaginary part. **The magnitudes of the complex number reflect the degrees of similarity between $x(t)$ and the frequency ω .** A concrete example at this point might shed some light on what is occurring.

To give a visual representation to what the Fourier Transform is doing we first need a function of time. So we begin by defining $x(t) = \cos(3 * 2\pi t)e^{-\pi t^2}$. The cosine function oscillates with an

amplitude of 1 at a frequency of 3 Hertz. The exponential portion just provides a time window that ramps the function up and then ramps the function down over a couple of seconds. $x(t)$ is depicted as Panel *A* of Figure 3.4.6. From here we can calculate the product of $x(t)\cos(\omega t)$ and the product of $x(t)\sin(\omega t)$ for $\omega = 6\pi$. Panel B displays the real portion of $X(6\pi)$. If we mentally integrate over the product represented by the magenta area it should be clear that the value is positive. The non-zero value indicates that the function $x(t)$ correlates with $\cos(6\pi t)$ which make intuitive sense. Panel C displays the imaginary portion of $X(6\pi)$. Now if we mentally integrate the product the portions below zero cancel one another out and the integral is close to zero. Now let's see what happens when $\omega = 1.3\pi$. Figure 3.4.7 depicts this case and we can see that both the real and imaginary integrals are close to zero in this case. To make the point that the real and imaginary portions represent a complex signal Figure 3.4.8 depicts $X(6\pi)$ in the three-dimensions (real, imaginary, time). The information contained in this figure is essentially the same as the information in Figure 3.4.6, it just makes it easier to visualize the complex path over time. The trace in the real plane is the purple area in the 2nd panel of Figure 3.4.6, and the trace in the imaginary plane is the purple area in the bottom panel. If we So far we have only examined positive frequencies. Fourier Transform can also be examined at negative frequencies.

3.4.2.5 What happens when ω is negative? Earlier we showed how Euler's identity $\cos(2\pi f_0 t)$ consists of two phasors. One circles about the complex plane in the CW direction while the other circles CCW. The reader should recall that the imaginary parts cancel while the real parts sum. With the Fourier transform the sign of the frequency term ω dictates the direction of the phasor as it rotates about the unit circle. When ω is positive the phasor is $e^{-j\omega t}$ and circles in the CW direction. When ω is negative the phasor becomes $e^{j\omega t}$ and circles in the CCW direction. What happens to the real and imaginary components of the Fourier Transform when ω is negative? To answer this we can address the integrals separately. If reader's are intimately familiar with their trigonometric identities the answer might reveal itself. Those that need a refresher should now

refer to Figure 2.3.1 and Figure 3.4.9.

If we look at the real part we see that because $\cos(-\alpha) = \cos(\alpha)$ the real portion of the Fourier Transform does not change. However the sign of the imaginary portion changes because $\sin(-\alpha) = -\sin(\alpha)$. Let's look at the complex phasors associated with the Fourier Transform of $x(t) = \cos(6\pi t)e^{-\pi t^2}$ when $\omega = 6\pi$ and $\omega = -6\pi$. In Figure 3.4.10 we can see that the real portion of the innerproduct remains identical while the imaginary portion is inverted. The interpretation is that $x(t)$ is equally similar at a frequency of 6π as it is at -6π. In the imaginary plane the similarity is opposite at negative frequencies, but the about of similarity (the integrals of the traces on the imaginary plane) are the same. If the reader is not yet convinced we can make this inversion less subtle by redefining $x(t) = \sin(6\pi t)e^{-\pi t^2}$. This is depicted in Figure 3.4.11. In this case the real portions of the inner procuct at both negative and positive frequencies are identical. However the imaginary ports are different. When ω is negative, the imaginary portion becomes inverted. If one were to integrate the blue and green traces they would find $x(t)$ is equally similar in magnitude at 6π as it is at -6π.

This operation exactly describes complex conjugation. For a complex $z = a + jb$ the complex conjugate is $\bar{z} = a - jb$. **The Fourier Transform of a purely real signal is Hermitian** (See Section 3.3.8.4)**.** A Hermitian form or symmetric sequilinear form is one in where$f(-x) = \overline{f(x)}$, which is exactly what we what we just demonstrated, essentially that $X(-\omega) = \overline{X(\omega)}$. The

Figure 3.4.6 *Panel A displays our original function x(t) of time. Panel B depicts the product between our original function and cos(6πt). If the purple area representing product is integrated the value is non-zero indicating the original function and cos(6πt) are similar. Panel C depicts the product between the original function and sin(6πt). In this case the integral is close to zero indicating sin(6πt) is not similar to x(t)*

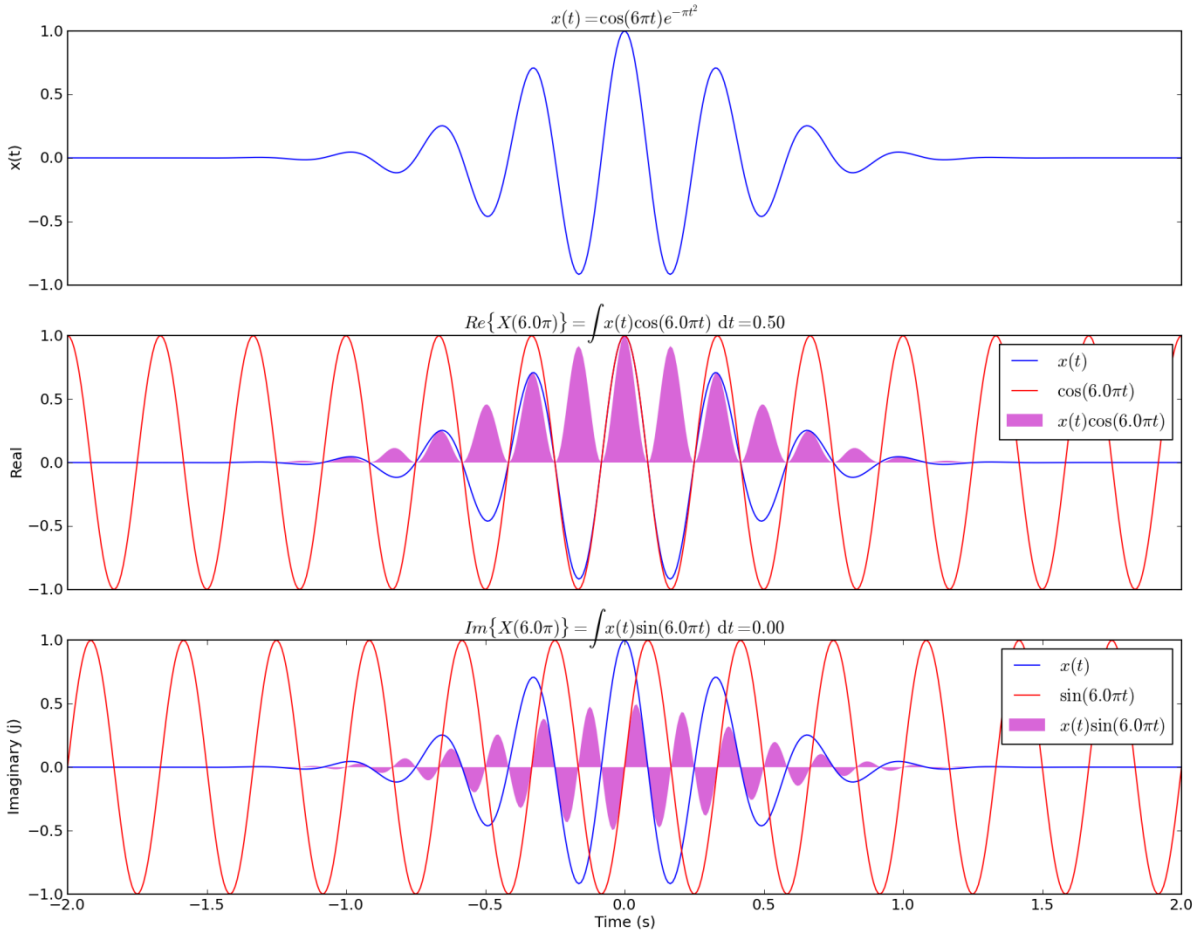


Figure 3.4.7 *The real and imaginary components of X(1.3πt). We can see the compared to cos(6.0πt) and sin(6.0πt), x(t) does not share much similarity with cos(1.3πt) and sin(1.3πt).*

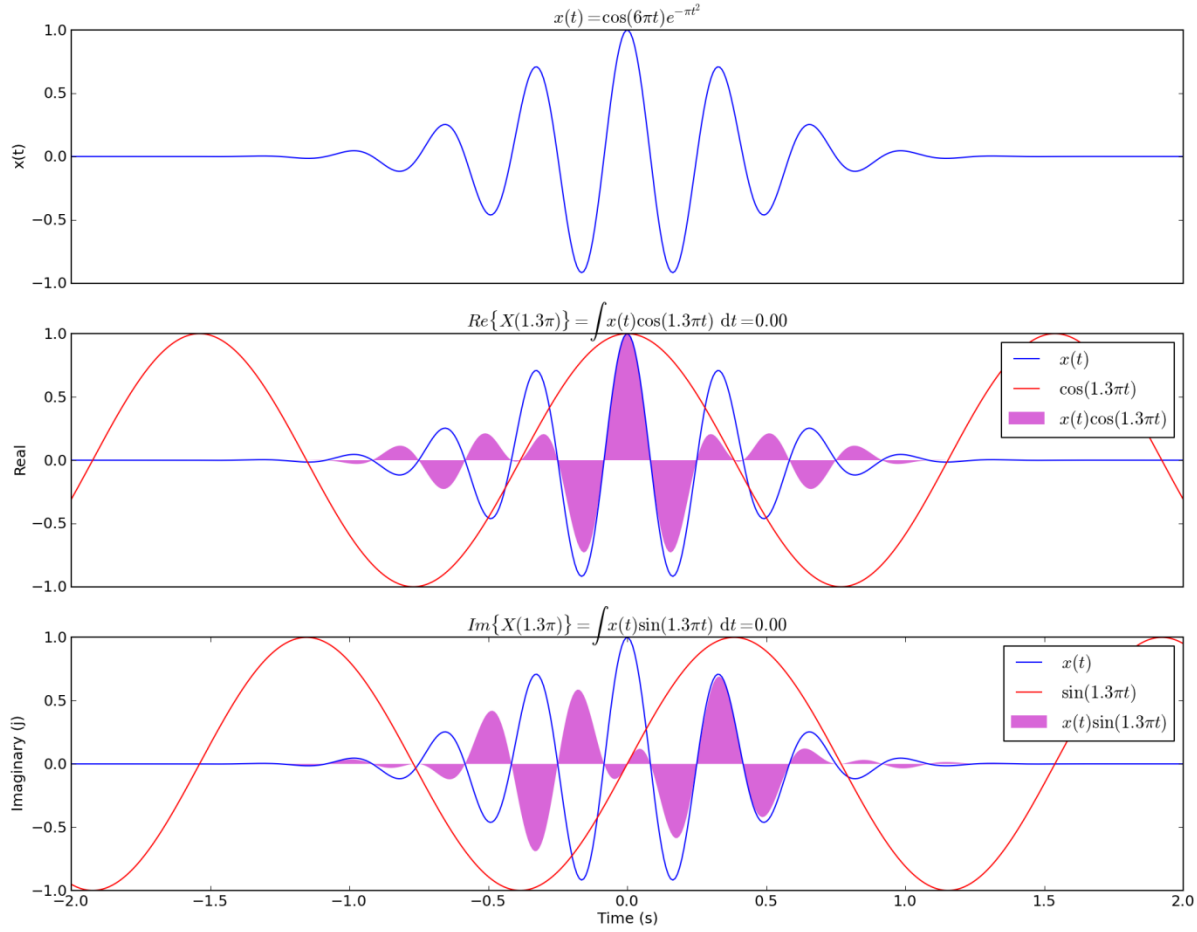


Figure 3.4.8 *This graph depicts X(6πt) as a complex phasor. The real and imaginary portions are projected as the dashed and dotted-and-dashed lines.*

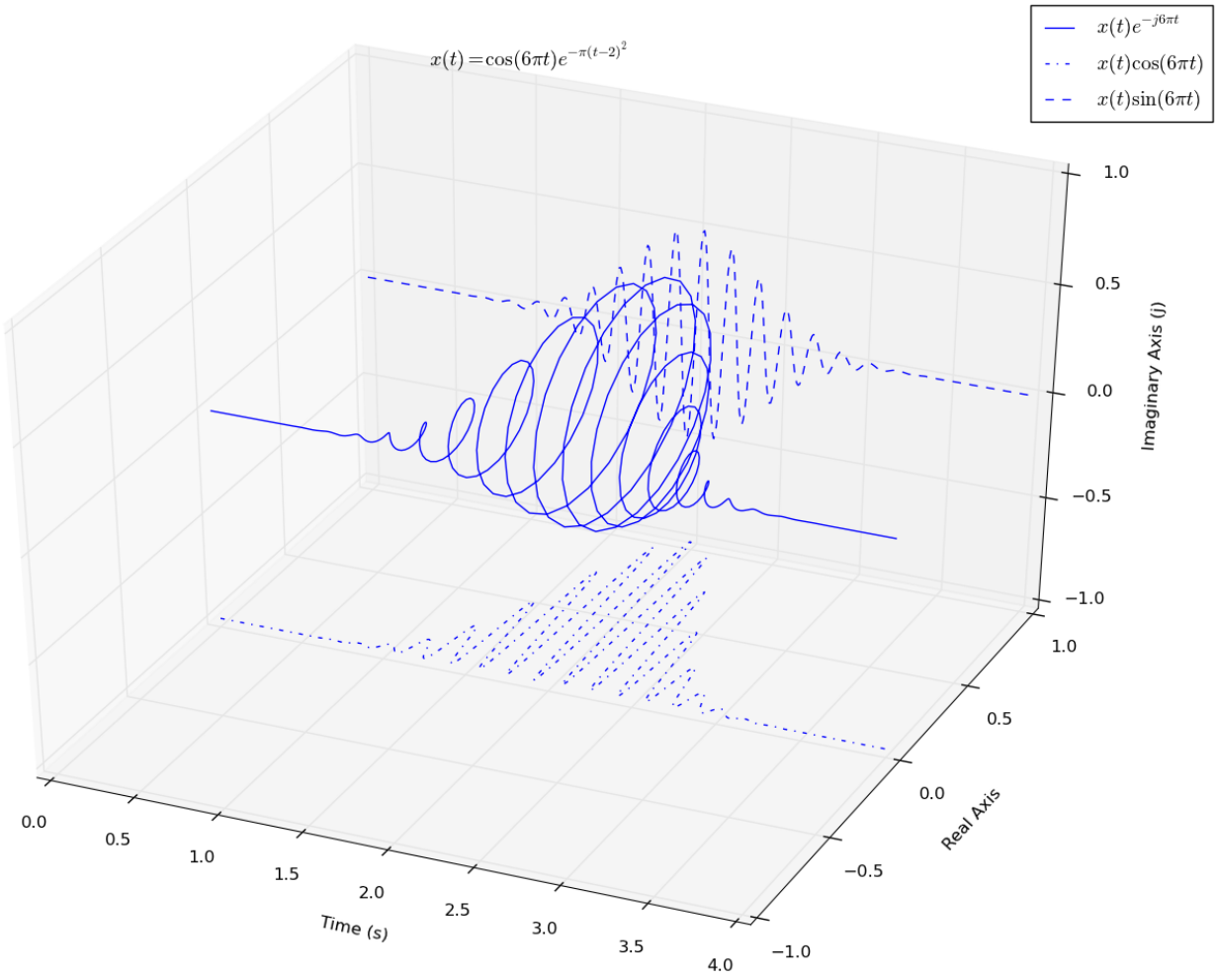


Figure 3.4.9 *Complex numbers can be transformed into polar coordinates. Using the following identities:*

$a + jb = r[\cos(\alpha) + j\sin(\alpha)]$

where,

α is the angle

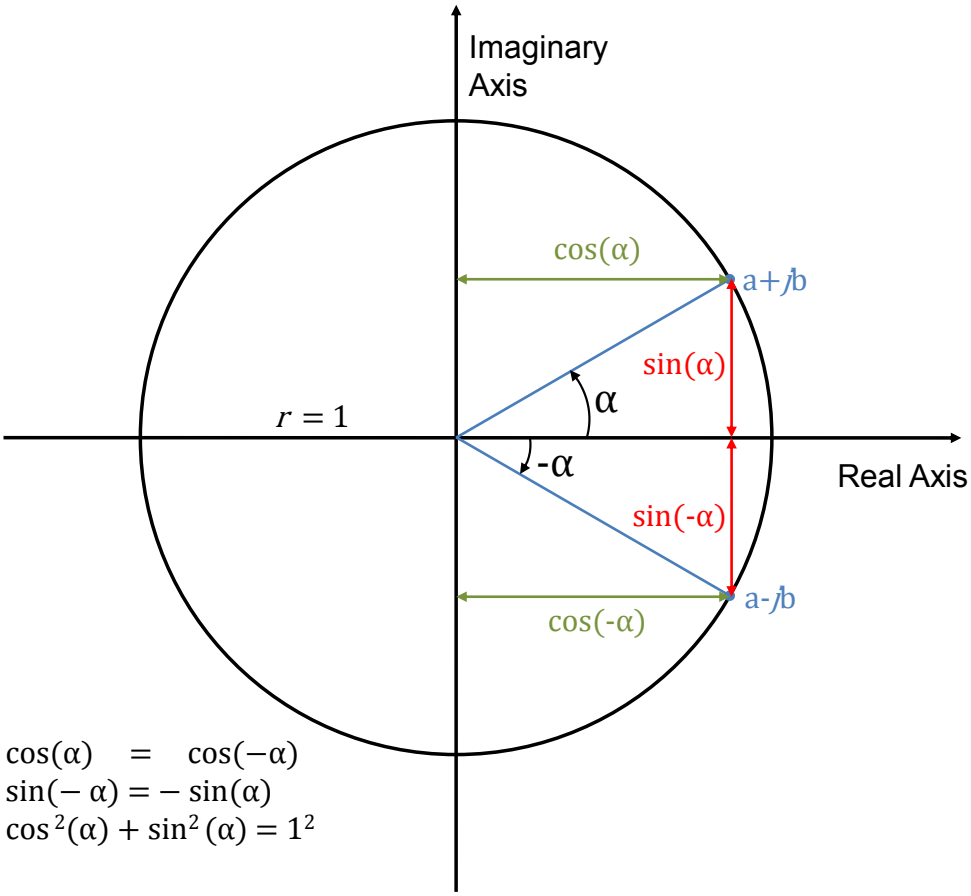
r is the magnitude

$a + jb \neq 0$

$r = \sqrt{(a + jb)(a - jb)} = \sqrt{a^2 + b^2}$

$\tan(\alpha) = \frac{b}{a}, \sin(\alpha) = \frac{b}{r}, \cos(\alpha) = \frac{a}{r}$

When r is fixed at 1 we can see that cos(α) *reflects the real portion of* a + *j*b *while* sin(α) reflects *the imaginary portion of* a + *j*b.



$\cos(\alpha) = \cos(-\alpha)$

$\sin(-\alpha) = -\sin(\alpha)$

$\cos^2(\alpha) + \sin^2(\alpha) = 1^2$

Figure 3.4.10 *Fourier transform symmetry of a enveloped cosine function. Complex representations of the innerproduct between x(t) and* $e^{-j\omega t}$ *when* $\omega = 6\pi$ *(blue) and when* $\omega = -6\pi$ *(green). We can see that the real portion is identical (because* $\cos(-\alpha) = \cos(\alpha)$*), while the imaginary portion becomes mirrored (because* $\sin(-\alpha) = -\sin(\alpha)$*). The implication is that x(t) is identically similar in the real domain at frequencies of* 6π *and* -6π*. In the imaginary domain x(t) is equally similar, or apply disimiliar, in magnitude at* 6π *and* -6π *but in an opposite fashion.*

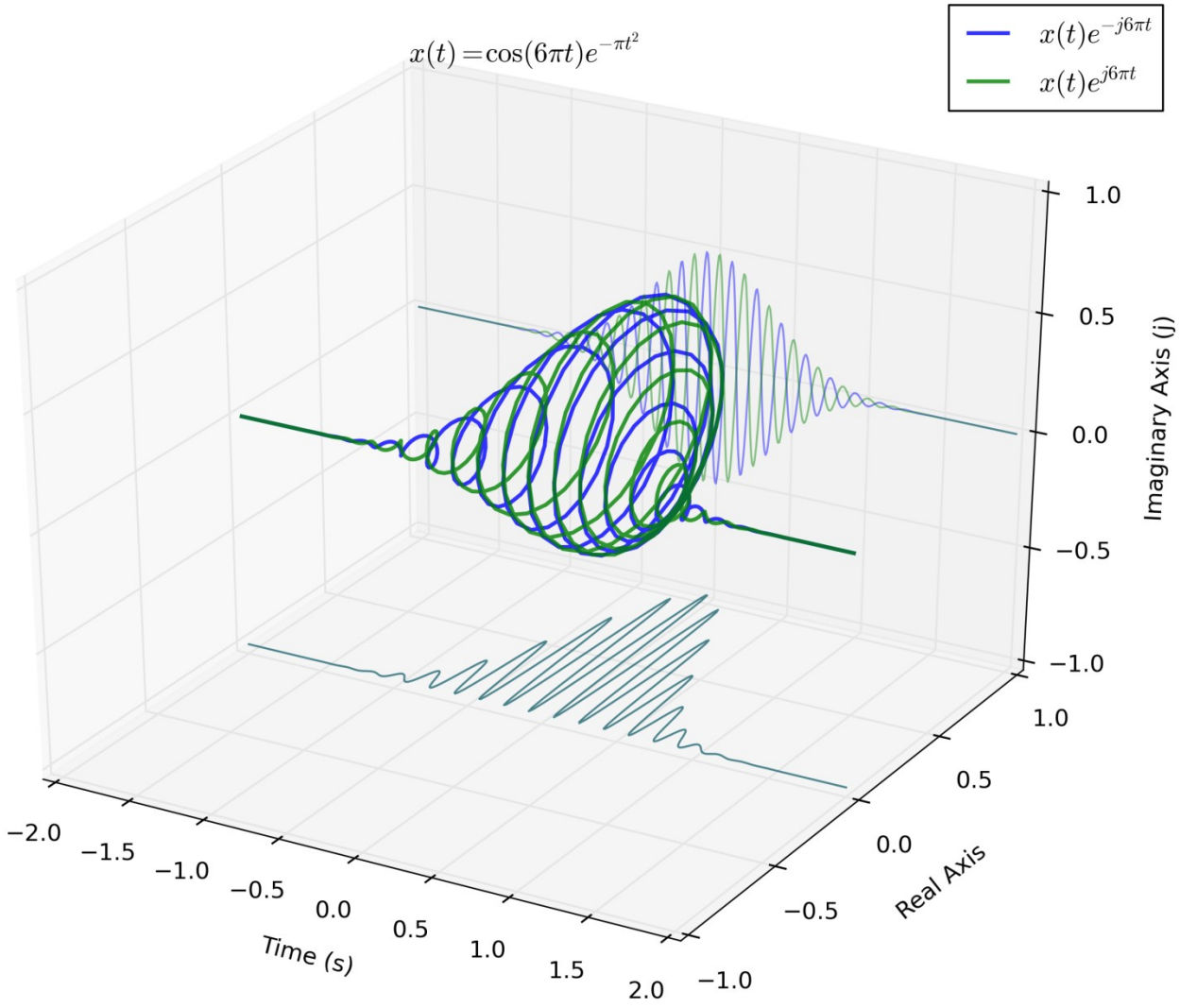


Figure 3.4.11 *Fourier transform symmetry of a enveloped sine function. Complex representations of the innerproduct between x(t) and e^(-jωt) when ω=6π (blue) and when ω=-6π (green).Once again the real portions are identical, while the imaginary portions are inverted.*

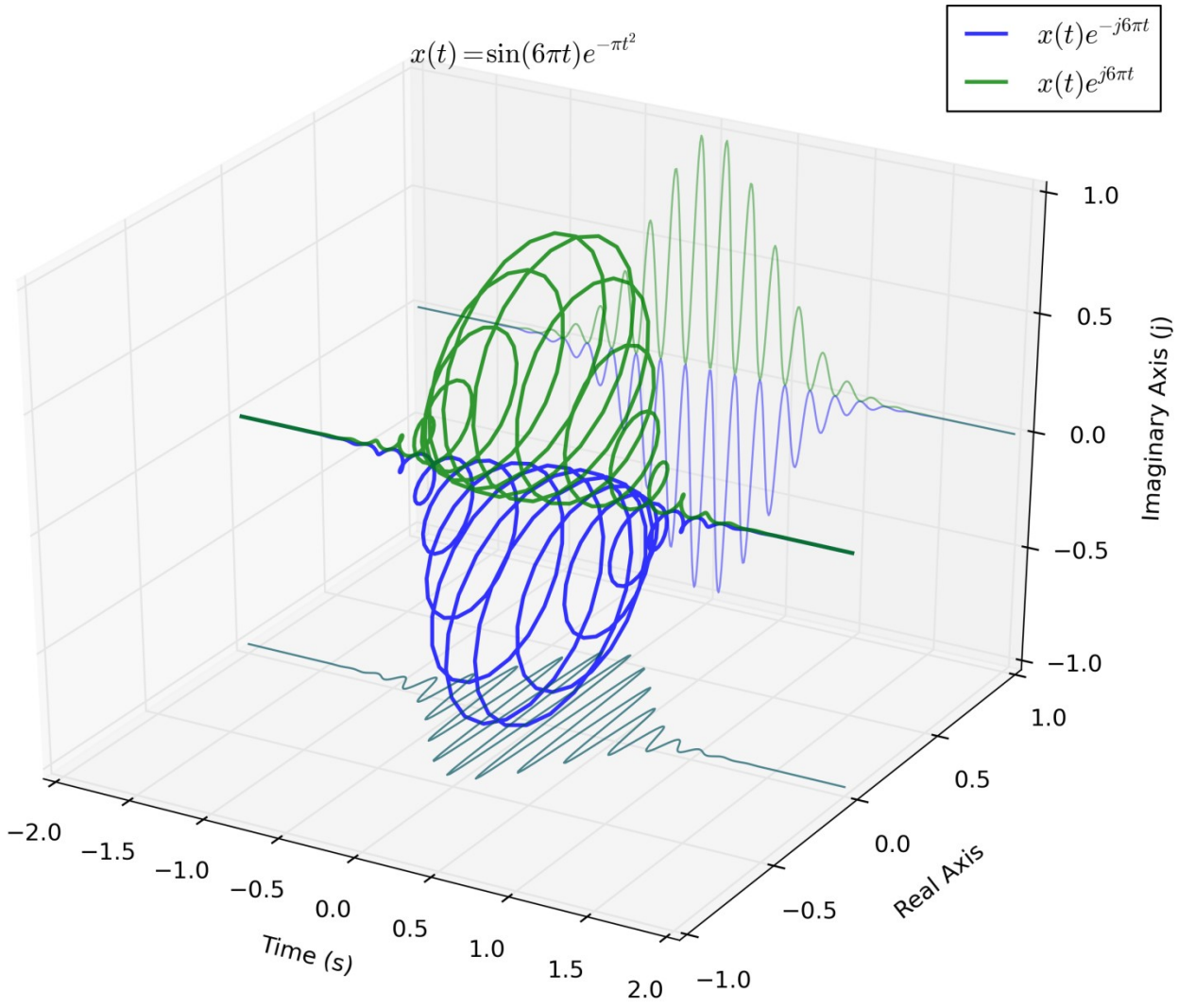
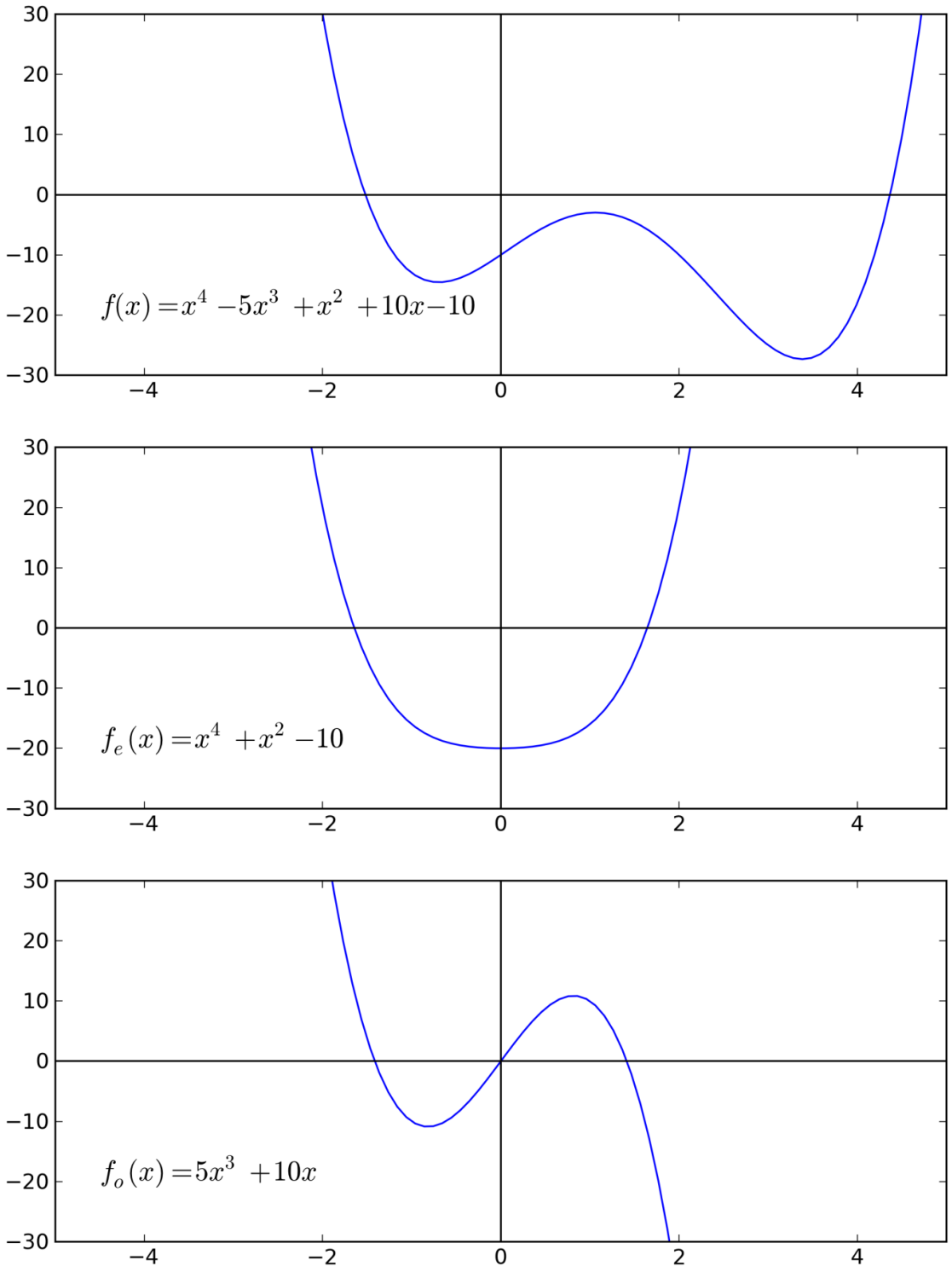


Figure 3.4.12 *Even and odd symmetry. The function in the middle panel has even symmetry,* $\boldsymbol{x(t) = x(-t)}$*. The function in the bottom panel has odd symmetry,* $\boldsymbol{x(t) = -x(t)}$*. The function in the top panel is neither even or odd, but can decomposed into even and odd functions.*



reader should take a moment to digest this insight and then think about what happens when the original signal, $x(t)$, is Hermitian.

3.4.2.6 Accounting for Even and Odd Symmetry in x(t). In the examples so far $x(t)$ has always been real-valued so $x(t)$ meets the requirement for being Hermitian when $x(-t) = x(t)$. Visually this just implies $x(t)$ has even symmetry about time 0 (see Figure 3.4.12). Let's look at our original Fourier Transform equation:

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,\mathrm{d}t$$

Since $x(t)$ has even symmetry we can change the lower bound to 0 if we account for the cases where t makes the phasor positive:

$$X(\omega) = \int_{0}^{\infty} \left[x(t)e^{-j\omega t} + x(t)e^{j\omega t}\right]\mathrm{d}t$$

$$X(\omega) = \int_{0}^{\infty} x(t)\left[e^{-j\omega t} + e^{j\omega t}\right]\mathrm{d}t$$

$$\frac{X(\omega)}{2} = \int_{0}^{\infty} x(t)\left[\frac{e^{-j\omega t}}{2} + \frac{e^{j\omega t}}{2}\right]\mathrm{d}t$$

The terms in the square brackets should look familiar. They are exactly Euler's formula for the cosine function. So:

$$X(\omega) = 2\int_{0}^{\infty} x(t)\cos(\omega t)\,\mathrm{d}t, \qquad \text{when } x(t) = x(-t)$$

When $x(t)$ has even symmetry the imaginary portion of the transform completely drops out and $X(\omega)$ is real. This is what we observed when $x(t) = \cos(3 * 2\pi t)e^{-\pi t^2}$. Now let's examine analytically what happens when $x(t)$ has odd symmetry. That is to say when $x(-t) = -x(t)$. In a fashion similar to above we can change the lower bound to 0 if we account for the cases where t makes the phasor positive and makes $x(t)$ negative:

$$X(\omega) = \int_0^\infty \left[x(t)e^{-j\omega t} - x(t)e^{j\omega t}\right] \mathrm{d}t$$

$$X(\omega) = \int_0^\infty x(t)\left[e^{-j\omega t} - e^{j\omega t}\right] \mathrm{d}t$$

$$\frac{X(\omega)}{2j} = \int_0^\infty x(t)\left[\frac{e^{-j\omega t}}{2j} + \frac{e^{j\omega t}}{2j}\right] \mathrm{d}t$$

Now the terms in the square brackets are exactly Euler's formula for the sine function. So:

$$X(\omega) = 2j\int_0^\infty x(t)\sin(\omega t)\,\mathrm{d}t, \qquad \text{when } x(-t) = -x(t)$$

In this case we can see that when $x(t)$ has odd symmetry the real portion of the transform drops out and $X(\omega)$ is completely imaginary. And this is what we observed when $x(t) = \sin(3 * 2\pi t)e^{-\pi t^2}$. Now obviously not all real functions will be purely even or purely odd but all functions can be expressed as a linear combination of even functions and odd functions. This allows any function to be expressed as the sum of an even function and an odd function (Smith, 2007).

$$f(x) = \frac{f(x)}{2} + \frac{f(-x)}{2} - \frac{f(-x)}{2} + \frac{f(x)}{2}$$

Notice the middle terms cancel and the outer terms sum to $f(t)$.

$$f(x) = \left(\frac{f(x)}{2} + \frac{f(-x)}{2}\right) + \left(\frac{f(x)}{2} - \frac{f(-x)}{2}\right)$$

$$f_e(x) \stackrel{\text{def}}{=} \frac{1}{2}\big(f(x) + f(-x)\big)$$

$$f_o(x) \stackrel{\text{def}}{=} \frac{1}{2}\big(f(x) - f(-x)\big)$$

$$f(x) = f_e(x) - f_o(x)$$

For example the polynomial $f(x) = x^4 - 5x^3 + x^2 + 10x - 10$ can be decomposed to

$$f_e(x) = x^4 + x^2 - 10$$

$f_o(x) = 5x^3 + 10x.$

Figure 3.4.11 Figure 3.4.12 plots $f(x)$ and its decomposed even and odd functions. This is what Euler's Formula is doing. Cosine is an even function, while sine is an odd function. We could plug $e^{j\omega t}$ into the above formulas for $f_e(x)$ and $f_o(x)$ we will arrive directly at the Euler's phasor representations for cosine and sine respectively. If we express $x(t)$ as a sum of even and odd functions the Fourier Transform becomes:

$$X(\omega) = \int [x_e(t) + x_o(t)] \cos(\omega t)\, \mathrm{d}t - j \int [x_e(t) + x_o(t)] \sin(\omega t)\, \mathrm{d}t$$

$$= \int x_e(t) \cos(\omega t)\, \mathrm{d}t + \int x_o(t) \cos(\omega t)\, \mathrm{d}t$$

$$-j \int x_e(t) \sin(\omega t)\, \mathrm{d}t - j \int x_o(t) \sin(\omega t)\, \mathrm{d}t$$

We have previous shown that $\int x_o(t) \cos(\omega t)\, \mathrm{d}t = 0$ and that $\int x_e(t) \sin(\omega t)\, \mathrm{d}t = 0$ so the transform can be simplified to:

$$X(\omega) = \int x_e(t) \cos(\omega t)\, \mathrm{d}t - j \int x_o(t) \sin(\omega t)\, \mathrm{d}t$$

When $x(t)$is purely real the even symmetry of $x(t)$ translates to the real axis while the odd symmetry of $x(t)$ translates to the imaginary axis. All the examples we have provided so far used real valued functions of $x(t)$. What happens when $x(t)$ is purely imaginary?

3.4.2.7 What happens when x(t) is purely imaginary? Now we can briefly address what happens when $x(t)$ is imaginary. To conceptualize $x(t)$ as imaginary we need only to multiply $x(t)$ by *j*. Doing so results in the Fourier Transform becoming:

$$X(\omega) = \int j\, \mathrm{Im}\{x_e(t)\} \cos(\omega t)\, \mathrm{d}t - j \int j \mathrm{Im}\{x_o(t)\} \sin(\omega t)\, \mathrm{d}t$$

$$= j \int \mathrm{Im}\{x_e(t)\} \cos(\omega t)\, \mathrm{d}t - j^2 \int \mathrm{Im}\{x_o(t)\} \sin(\omega t)\, \mathrm{d}t$$

$$= j \int \mathrm{Im}\{x_e(t)\} \cos(\omega t)\, \mathrm{d}t + \int \mathrm{Im}\{x_o(t)\} \sin(\omega t)\, \mathrm{d}t$$

when $x_e(t)$and $x_o(t)$ are imaginary

The Im{ . } function returns the imaginary portion of a complex scalar as a real scalar. For $a + jb$ in $\mathbb{C}$ the function is defined as,

$$\mathrm{Im}\{a + jb\} \stackrel{\text{def}}{=} b.$$

In a similar vein the function Re{ . } returns the real portion of a complex scalar as a real scalar such that,

$$\mathrm{Re}\{a + jb\} \stackrel{\text{def}}{=} a.$$

When the argument given Im{ . } or Re{ . } is a complex vector the functions can be defined as follows:

$$\mathrm{Im}\{(a_1 + jb_1, a_2 + jb_2, \dots, a_n + jb_n)\} \stackrel{\text{def}}{=} (b_1, b_2, \dots, b_n),$$

$$\mathrm{Re}\{(a_1 + jb_1, a_2 + jb_2, \dots, a_n + jb_n)\} \stackrel{\text{def}}{=} (a_1, a_2, \dots, a_n).$$

When we treat the imaginary functions $x_e(t)$and $x_o(t)$ as if they were real we can better illustrate the directions of the complex components. This shows that when $x(t)$ is imaginary the even portion of the transform is imaginary and the odd portion becomes real and is inverted. When ω is negative the Fourier transform becomes:

$$X(\omega, \omega < 0) = j \int \mathrm{Im}\{x_e(t)\} \cos(|\omega|t)\, \mathrm{d}t - \int \mathrm{Im}\{x_o(t)\} \sin(|\omega|t)\, \mathrm{d}t$$

when $x_e(t)$and $x_o(t)$ are imaginary

because $\cos(-x) = \cos(x)$ and $\sin(-x) = -\sin(x)$. **From these identities we can infer that when $x(t)$ is imaginary the Fourier transform is anti-Hermitian.** An anti-Hermitian, skew-Hermitian, or antisymmetric sequilinear form is one where $X(-\omega) = \overline{-X(\omega)}$. By this we mean that the real part is inverted while the imaginary portion remains the same.

3.4.2.8 What happens when x(t) is complex? So far we have looked at input functions that have been completely real or completely imaginary. What happens when the signals have both real and imaginary components? Many real-world applications are treated as complex signals and benefit from the analytical tools provided by Fourier Analysis. These include but are not limited to radar, digital and analog communication systems, coherent pulse measurement systems, antenna

beamforming, and so on (Lyons, 2008). In these applications the real and imaginary transforms might have specific purposes. Quadrature amplitude modulation (QAM) provides distinct channels for communicating information on the real and imaginary axes.

Furthermore complex signals often need to be analyzed at both positive and negative frequencies because their positive and negative spectra do not share the redundancy inherent in real or imaginary signals. To illustrate this point let's examine the transform of $x(t) = e^{j6\pi t}e^{-\pi t^2}$ when $\omega = 6\pi$ and $\omega = -6\pi$. From Figure 3.4.13 we can see that $X(-\omega)$ is approximately 0 + j0 while the $X(\omega)$ has a positive real portion and no imaginary portion. We can verify this analytically by substituting $e^{j6\pi t}e^{-\pi t^2}$ in for $x(t)$ in the Fourier Transform equation:

$$X(6\pi) = \int e^{j6\pi t}e^{-\pi t^2}\, e^{-j6\pi t}\, \mathrm{d}t$$

$$= \int \frac{e^{j6\pi t}e^{-\pi t^2}}{e^{j6\pi t}}\, \mathrm{d}t$$

$$= \int e^{-\pi t^2}\, \mathrm{d}t$$

$$= 1 + j0$$

The real component is 1 and the imaginary component is 0. Now when we examine the Fourier transform at a negative frequency of -6 π we obtain:

$$X(-6\pi) = \int e^{j6\pi t}e^{-\pi t^2}\, e^{j6\pi t}\, \mathrm{d}t$$

$$= \int e^{j12\pi t}e^{-\pi t^2}\, \mathrm{d}t$$

$$= \int e^{-\pi t(t-12j)}\, \mathrm{d}t$$

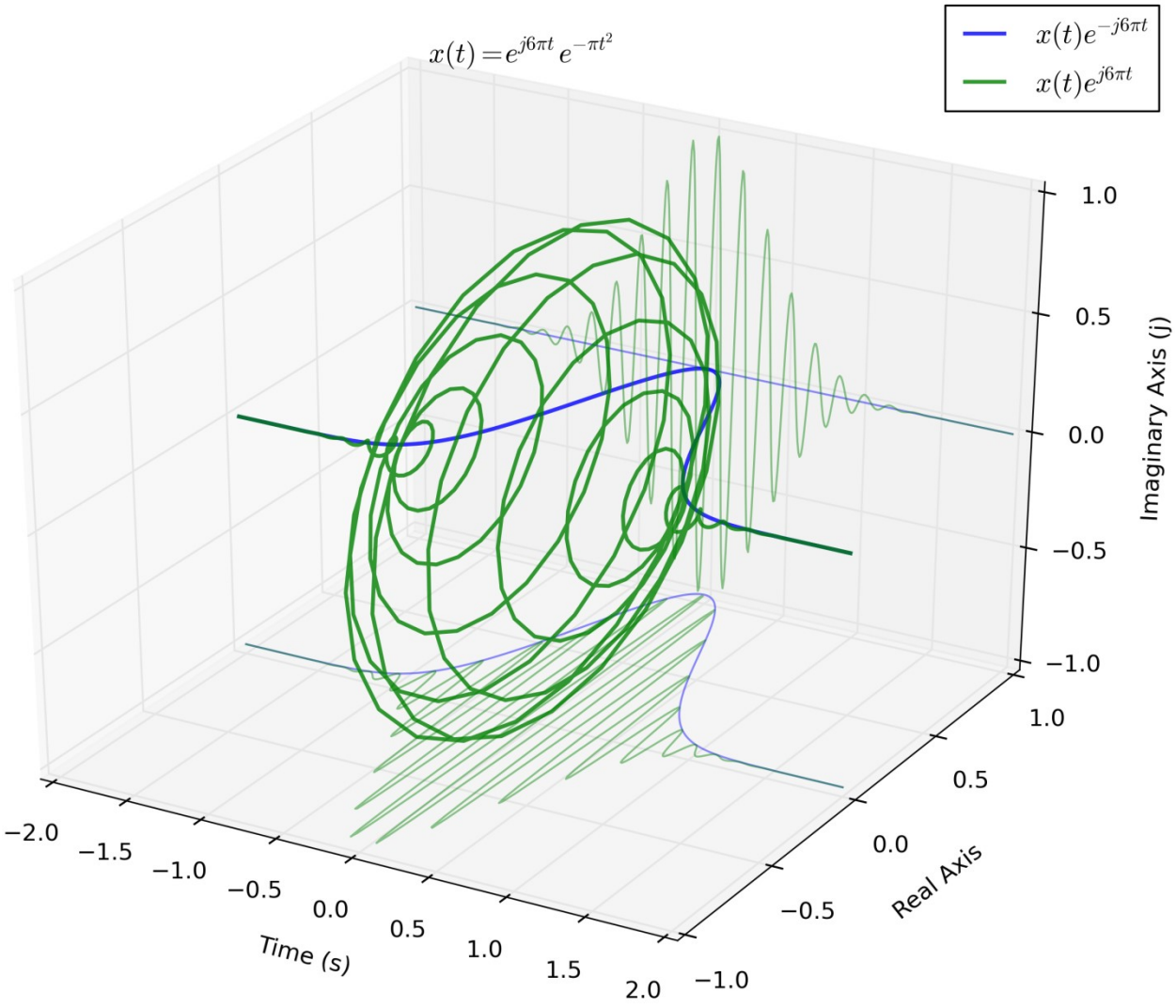
$$\cong 0 + j0$$

In this case the real portion is zero and the imaginary portion is also zero. This can be visually verified in Figure 3.4.13.

Figure 3.4.13 *Fourier transform symmetry of a complex function. Plot depicts Fourier Transform of an enveloped phasor at frequencies of* **6π** *and* **-6π**. *When ω is positive the phasors x(t) and* $e^{-j\omega t}$ *rotate in the opposite direction. When ω is negative they rotate in the same direction.*

At $\omega = 6\pi$, $x(t)e^{-j6\pi t} = e^{j6\pi t}e^{-\pi t^2}e^{-j6\pi t} = e^{-\pi t\text{^}2}$ *(blue)*

At $\omega = -6\pi$, $x(t)e^{j6\pi t} = e^{-j6\pi t}e^{-\pi t^2}e^{-j6\pi t} = e^{-j12\pi t}e^{-\pi t\text{^}2}$ *(green)*



It is nice to see that these integrals agree with our graphical interpretation, but they don't really reveal what is going on here in an obvious manner (at least to novice eyes). To understand what is happening let's look at complex signals of the form:

$$x(t) = e^{jf_0t} + je^{jf_1t} \text{ where } f_0 \neq f_1$$

If we expand these phasors using Euler's Formula we see that:

$$\begin{aligned} x(t) &= \cos(f_0t) + j\sin(f_0t) + j[\cos(f_1t) + j\sin(f_1t)] \\ &= \cos(f_0t) + j\sin(f_0t) + j\cos(f_1t) - \sin(f_1t) \\ &= \cos(f_0t) - \sin(f_1t) + j[\cos(f_1t) + \sin(f_0t)] \end{aligned}$$

In this form we can see that a signal $x(t)$ can have both even and odd symmetry on the real axis and both even and odd symmetry on the complex axis. Before it was shown how a generic function could be represented by the sum of even and odd functions. We can now treat the real and imaginary components in a similar fashion:

$$\begin{aligned} x(t) &= \mathrm{Re}\{x(t)\} + j\mathrm{Im}\{x(t)\} \\ &= \mathrm{Re}\{x_e(t)\} + \mathrm{Re}\{x_o(t)\} + \mathrm{jIm}\{x_e(t)\} + j\mathrm{Im}\{x_o(t)\} \end{aligned}$$

Next we can substitute these into the Fourier transform:

$$\begin{aligned} X(\omega) = &\int [\mathrm{Re}\{x_e(t)\} + \mathrm{Re}\{x_o(t)\} + \mathrm{jIm}\{x_e(t)\} + j\mathrm{Im}\{x_o(t)\}]\cos(\omega t)\,\mathrm{d}t \\ &-j\int [\mathrm{Re}\{x_e(t)\} + \mathrm{Re}\{x_o(t)\} + \mathrm{jIm}\{x_e(t)\} + j\mathrm{Im}\{x_o(t)\}]\sin(\omega t)\,\mathrm{d}t \end{aligned}$$

After removing zero integrals and simplifying we obtain:

$$\begin{aligned} X(\omega) = &\int \mathrm{Re}\{x_e(t)\}\cos(\omega t)\,\mathrm{d}t + \int \mathrm{jIm}\{x_e(t)\}\cos(\omega t)\,\mathrm{d}t \\ &-j\int \mathrm{Re}\{x_o(t)\}\sin(\omega t)\,\mathrm{d}t - j\int \mathrm{jIm}\{x_o(t)\}\sin(\omega t)\,\mathrm{d}t \\ = &\int \mathrm{Re}\{x_e(t)\}\cos(\omega t)\,\mathrm{d}t + \int \mathrm{Im}\{x_o(t)\}\sin(\omega t)\,\mathrm{d}t \\ &+j\int \mathrm{Im}\{x_e(t)\}\cos(\omega t)\,\mathrm{d}t - j\int \mathrm{Re}\{x_o(t)\}\sin(\omega t)\,\mathrm{d}t \end{aligned}$$

Now we can treat the result as piecewise to account for the Hermitian and anti-Hermitian symmetry about the real and imaginary axes respectively:

$$X(\omega)=\begin{cases}\int \mathrm{Re}\{x_e(t)\}\cos(\omega t)\,\mathrm{d}t+\int \mathrm{Im}\{x_o(t)\}\sin(\omega t)\,\mathrm{d}t \\ +j\int \mathrm{Im}\{x_e(t)\}\cos(\omega t)\,\mathrm{d}t-j\int \mathrm{Re}\{x_o(t)\}\sin(\omega t)\,\mathrm{d}t\,, & \omega\geq 0 \\ \int \mathrm{Re}\{x_e(t)\}\cos(|\omega|t)\,\mathrm{d}t-\int \mathrm{Im}\{x_o(t)\}\sin(|\omega|t)\,\mathrm{d}t \\ +j\int \mathrm{Im}\{x_e(t)\}\cos(|\omega|t)\,\mathrm{d}t+j\int \mathrm{Re}\{x_o(t)\}\sin(|\omega|t)\,\mathrm{d}t\,, & \omega<0\end{cases} \quad \textbf{3.4.6}$$

If the reader can understand this representation of the Fourier Transform and its rather intuitive derivation then the reader should have a grasp of all the topics discussed thus far.

We should now turn our attention back to $x(t)=e^{j6\pi t}e^{-\pi t^2}$. Through some manipulation:

$$\begin{aligned}x(t)&=[\cos(6\pi t)+j\sin(6\pi t)]e^{-\pi t^2}\\&=\cos(6\pi t)\,e^{-\pi t^2}+j\sin(6\pi t)\,e^{-\pi t^2}\\&=\mathrm{Re}\{x_e(t)\}+j\mathrm{Im}\{x_o(t)\}\end{aligned}$$

where:

$$\mathrm{Re}\{x_e(t)\}=\cos(6\pi t)\,e^{-\pi t^2}$$

$$\mathrm{Im}\{x_o(t)\}=\sin(6\pi t)\,e^{-\pi t^2}$$

In this instance $\mathrm{Re}\{x_0(t)\}$ and $\mathrm{Im}\{x_e(t)\}$ are both zero since this is necessary to specify $x(t)$. When we take the Fourier Transform at $\omega=6\pi$ the four definite integrals are reduced to two:

$$\begin{aligned}X(6\pi)&=\int \mathrm{Re}\{x_e(t)\}\cos(6\pi t)\,\mathrm{d}t+\int \mathrm{Im}\{x_o(t)\}\sin(6\pi t)\,\mathrm{d}t\\&=\int \cos(6\pi t)\,e^{-\pi t^2}\cos(6\pi t)\,\mathrm{d}t+\int \sin(6\pi t)\,e^{-\pi t^2}\sin(6\pi t)\,\mathrm{d}t\\&=\int \cos^2(6\pi t)\,e^{-\pi t^2}\,\mathrm{d}t+\int \sin^2(6\pi t)\,e^{-\pi t^2}\,\mathrm{d}t\\&=\frac{\left(1+e^{-36\,pi}\right)}{2}+\frac{\left(1+e^{-36\,pi}\right)}{2}=\left(1+e^{-36\,pi}\right)\approx 1+j0\end{aligned}\quad .$$

This result agrees with the integral we obtained by through exponential integration as well as with our visual interpretation. Now let's see what happens when we compute the Fourier Transform at $\omega = -6\pi$.

$$X(-6\pi) = \int \mathrm{Re}\{x_e(t)\}\cos(|-6\pi|t)\,\mathrm{d}t - \int \mathrm{Im}\{x_o(t)\}\sin(|-6\pi|t)\,\mathrm{d}t$$

$$= \int \cos(6\pi t)\, e^{-\pi t^2}\cos(|-6\pi|t)\,\mathrm{d}t - \int \sin(6\pi t)\, e^{-\pi t^2}\sin(|-6\pi|t)\,\mathrm{d}t$$

$$= \int \cos^2(6\pi t)\, e^{-\pi t^2}\,\mathrm{d}t - \int \sin^2(6\pi t)\, e^{-\pi t^2}\,\mathrm{d}t$$

$$= \frac{\left(1 + e^{-36\,pi}\right)}{2} - \frac{\left(1 + e^{-36\,pi}\right)}{2} = 0 + j0 \qquad .$$

This also agrees with our previous result obtained through exponential integration yet it is much easier to get an idea of why $X(6\pi)$ is so different from $X(-6\pi)$. At first glance, decomposing complex signals into their real and imaginary parts as well as into their even and odd parts seems complicated but it yields greater insight into Fourier Transformation.

The piecewise formulation of the Fourier Transform presented in Equation 3.4.6 reveals the complex symmetry (pun intended) between even and odd functions, real and imaginary signal components, and positive and negative frequencies.

- When $x(t)$ is purely real only the 1st and 4th terms remain and we can see $X(\omega)$ is Hermitian -> $X(-\omega) = \overline{\mathrm{X}(\omega)}$
- When $x(t)$ is purely imaginary only the 2nd and 3rd terms remain and we can see $X(\omega)$ is anti-Hermitian -> $X(-\omega) = \overline{-\mathrm{X}(\omega)}$
- When $x(t)$ is complex the real and imaginary and even odd symmetries can be easily accounted for by the piecewise representation
- The symmetry explored here apply equally to the Fourier series

So far we have been examining in minutia how the real and imaginary components of an input interact with the real and imaginary components of a basis function to form a complex scalar

output. This understanding will come in handy when we start to look not just at single elements or pairs of elements within the inner product space, but examine the inner product space as a whole. It is the entirety of the inner product space which holds the spectral representation of $x(t)$. As we will soon see the discrete wavelet transform lends itself well to just this.

3.4.3 ***Introduction to the Discrete Fourier Transform.*** Working directly with the Fourier transform offers some analytic advantages due to the fact we can find closed form solutions, and gain insight through mathematical manipulation. However, in the real world empirically recorded signals will almost always be discretely sampled. In these cases the discrete Fourier transform (DFT) is the tool of choice for calculating the spectra of a discrete input. The transform is defined as

$$\mathbf{X}_n \stackrel{\text{def}}{=} \sum_{\mathrm{k}=0}^{\mathrm{N}-1} \mathbf{x}_k e^{-j2\pi kn/\mathrm{N}}, \text{for } n = 0, \dots, \mathrm{N}-1.$$

The vector $\mathbf{x}_k$ is the original signal with *N* samples. The parameter *k* is our "time" parameter, it specifies $\mathbf{x}_k$ where $\mathbf{x}_0$ refers to the first sample and $\mathbf{x}_{N-1}$ refers to the last sample. The vector $\mathbf{X}_n$ is an inner product space holding the spectral representation of $\mathbf{x}_k$. The *n* parameter specifies frequency, but how it specifies it will require some expounding. As presented above the transform is non-unitary and the corresponding inverse transform is

$$\mathbf{x}_n \stackrel{\text{def}}{=} \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=0}^{\mathrm{N}-1} \mathbf{X}_n e^{j2\pi kn/\mathrm{N}}, \text{for } n = 0, \dots, \mathrm{N}-1.$$

We can see from the transform is non-unitary since the inverse carries the normalization factor of $\frac{1}{\mathrm{N}}$. The discrete Fourier transform shares the most likeness to the generalized Fourier series coefficients equation we presented as

$$\langle f, e_n \rangle \stackrel{\text{def}}{=} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)\overline{e_n} dx, \text{where } B_T = \left\{ e_n = e^{j2\pi nx/T}, \text{for all } n \text{ in } \mathbb{Z} \right\} \text{for } L^2[-\pi, \pi].$$

We can modify the domain to $L^2[0, 2\pi]$

$$\langle f, e_n \rangle \stackrel{\text{def}}{=} \frac{1}{T}\int_0^{\mathrm{T}} f(x)\overline{e_n}dx \text{ , where } B_T = \left\{e_n = e^{j2\pi nx/T}, \text{for all } n \text{ in } \mathbb{Z}\right\} \text{ for } L^2[0, 2\pi].$$

The discrete Fourier transform can be derived from this definition of the Fourier series coefficients by treating $f(x)$as a discrete sequence of finite length. The DFT, like the Fourier series, treats the input signal as if it were periodic. When using the DFT it is important to remember that it always has the underlying assumption that the input sequence is periodic over its length. Discontinuities in the input signal result in spectral noise. These factors will be discussed in more detail.

3.4.3.1 Fast Fourier Transform is the Discrete Fourier Transform. The discrete Fourier transform may colloquially be referred to as the fast Fourier transform (FFT).[15] The difference is purely algorithmic. The resulting coefficients from the FFT are identical to the *slow* version. The FFT is optimized for power of 2 sample sizes (2, 4, 8, 16, 32, …) although it can handle non-power of 2 samples (usually by using zero-padding). The discrete transform has $O(N^2)$ multiplication operations and $O(N^2)$ addition operations. The FFT has $O(N\log(N))$ multiplication operations and $O(N\log(N))$ addition operations (Smith, 2007).

3.4.3.2 Basis Orthogonality. The discrete Fourier transform has a finite sized basis with the number of basis functions being equal to the length of the input sequence. The basis defined as

$$B \stackrel{\text{def}}{=} \{e_n = e^{j2\pi n/\mathrm{N}}, \text{for } n = 0, \ldots, \mathrm{N}-1\}$$

$$\langle e_n, e_m \rangle = \sum_{\mathrm{k}=0}^{\mathrm{N}-1} e^{j2\pi kn/\mathrm{N}} e^{j2\pi km/\mathrm{N}}$$

$$= \sum_{\mathrm{k}=0}^{\mathrm{N}-1} e^{j2\pi k(n-m)/\mathrm{N}}$$

$$= \frac{1 - e^{j2\pi k(n-m)}}{1 - e^{j2\pi k(n-m)/\mathrm{N}}}$$

The last step uses the geometric series formula. When $n \neq m$ the inner product is 0 numerator becomes zero, and the denominator is non-zero. Because the inner product is zero, e_n and e_m are

[15] Strictly speaking this is describing the Cooley-Tukey algorithm. There are other FFT algorithms as well, but their scopes of application are quite narrow (Haynal & Haynal, 2011; Good, 1958).

orthogonal. When $n = m$ the inner product is becomes $\langle e_n, e_n \rangle$, which yields N. Despite having a finite sized basis sets the discrete wavelet transform is able to fully represent any discretely sampled finite signal. One may wonder "how is this possible?" The key lies in understanding that the bandwidth (frequency range) of a signal is intrinsically linked to its sampling rate through the Nyquist-Shannon sampling theorem.

3.4.3.3 Nyquist-Shannon Sampling Theorem. The Nyquist-Shannon sampling theorem provides rationale for how the discrete Fourier transform can have basis with finite cardinality. The theorem states (Shannon, 1949):

> If a function $x(t)$ contains no frequencies higher than W hertz, then it is completely determined by giving its ordinates at a series of points spaced $1/(2W)$ seconds apart.

Shannon's original proof begins by assuming a function $x(t)$ and its Fourier transform $X(\omega)$ are bandlimited by a frequency W Hertz. By bandlimited we mean that $X(\omega) = 0$ when $|\omega| > W$. Because $x(t)$ and its Fourier transform $X(\omega)$ are bandlimited we can substitute the infinite bounds of the inverse Fourier transform with W ,

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} \, d\omega$$

$$= \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} X(\omega) e^{j\omega t} \, d\omega.$$

Shannon then proposes examining what happens to the signal every $\frac{1}{2W}$ seconds. We can do this by substituting $\frac{n}{2W}$ for *t.* [$n \in \mathbb{Z}$]

$$x\left(\frac{n}{2W}\right) = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} X(\omega) e^{j\omega \frac{n}{2W}} \, d\omega.$$

On the right we can see the integral matches the generalized Fourier series expansion for the coefficients,

$$c_n = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} X(\omega) e^{-j\omega \frac{n}{2W}} \, d\omega$$

$$= \frac{1}{2\pi} \; x\left(\frac{n}{2W}\right).$$

From this we can see that the $c_n$is proportional to the sample at $x\left(\frac{n}{2W}\right)$. When we plug the c_n into the Fourier series,

$$X(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{j\omega \frac{n}{2W}}$$

$$= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2W}\right) e^{j\omega \frac{n}{2W}}$$

it becomes clear that the Fourier transform $X(\omega)$ is completely determined by the discrete samples of $x\left(\frac{n}{2W}\right)$.

The corollary is that is discretely sampled signals are only able to accurately represent frequencies up to half of the sampling rate. The limit is known as the Nyquist frequency (Blackledge, 2003) or the folding frequency and denoted f_s. It is called the folding frequency because high frequency content above f_s becomes reflected below f_s due to aliasing. For example, digital video discs (DVDs) have a sampling rate of 48.0 kHz which means it can represent signals up to 24.0 kHz. A frequency of 27.0 kHz would be aliased to 21.0 kHz. The discrete Fourier transform takes advantage of this aliasing to transform negative frequencies.

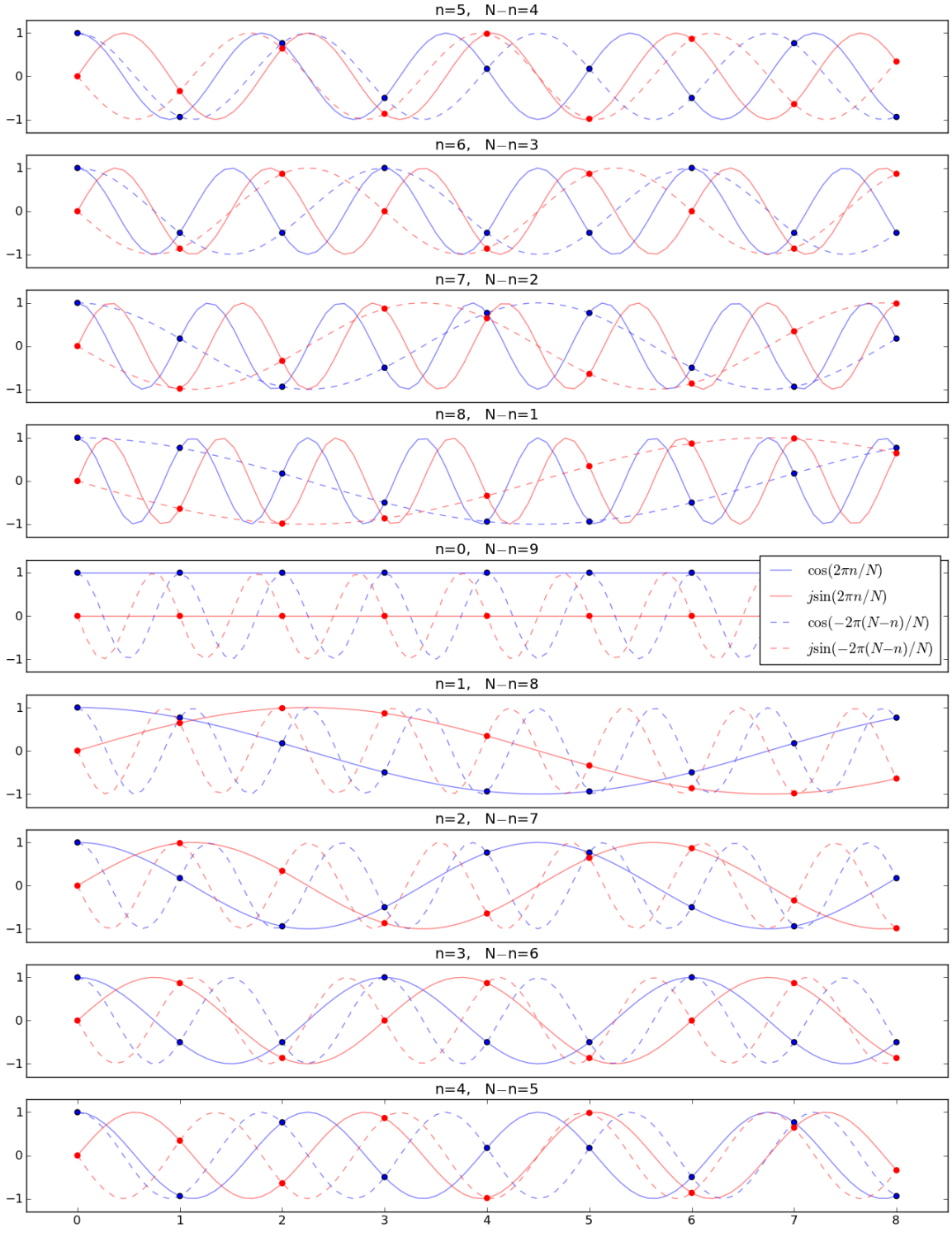
3.4.3.4 Aliasing and Negative Frequencies. The discrete Fourier transform takes advantage of the folding frequency to transform negative frequencies. When *n* is larger than N/2 the inner product is transforming the frequencies about $2\pi n/N$ but at the same time it is transforming frequencies about – $2\pi(\mathrm{N} - n)/N$. Figure 3.4.14 depicts the basis functions when $N = 9$. In each subplot the solid lines plot the $\cos(2\pi n/N)$ and $\sin(2\pi n/N)$. The dashed lines plot $\cos(-2\pi(\mathrm{N} - n)/N)$ and $\sin(-2\pi(\mathrm{N} - n)/N)$. The markers show how the aliasing caused by the discrete

sampling aligns the positive and negative frequencies. From the figure we can see that the basis functions are Hermitian. The real parts remain equivalent at negative and positive frequencies, but the imaginary parts are inverted.

An analogy which might help understand how the negative frequencies are represented is to imagine looking up at a ceiling fan with one blade under a strobe light. The strobe is analogous to discretization in the analogy. If the fan is physically stationary it doesn't matter how fast the strobe light is the fan will appear stationary. Now imagine powering on the fan so its speed is relatively low compared to the speed of the strobe. Under this situation the direction and speed of its motion can be accurately resolved. This is the analogous to when $n = 0$ in Figure 3.4.14. As the fan speeds up its speed and direction become difficult to perceive when the frequency of the fan approaches half the frequency (full rotations per unit time) of the strobe ($n = 1,2,3,4$). When the fan's frequency becomes greater than half the frequency of the strobe the fan will be perceived as moving in opposite direction from which it is actually moving($n = 5,6,7,8$). When the fan's frequency is equal to that of the strobe the fan will once again appear stationary ($n = 9$). The analogy illustrates the basic effect of how *n* causes negative frequencies to be aliased into the transform. A more literal analogy would be to imagine a ceiling fan with two blades at a right angle to one another. Instead of a strobe the fan blades would have lights that flash in synchronization on the tips of their blades, and instead of looking up at the fan you would be looking at the fan on edge.

While the details on how the real and imaginary portions of the transform are calculated may differ between the continuous and discrete Fourier transforms the way those values are interpreted is similar; the discrete Fourier transform maintains the symmetry characteristics we have previously become familiar with.

Figure 3.4.14 *The Nyquist or folding frequency of the discrete Fourier transform. The discrete Fourier transform takes advantage of the Nyquist Frequnecy or folding frequency defined by twice the sampling rate. Signals above the nyquist frequency are mirrored and reflected as negative frequencies (Hermitian: the real parts stay the same the imginary parts are inverted).*



3.4.3.5 Phase and Magnitude. When we examined the continuous Fourier transform our focus was directed at interpreting the real and imaginary components directly. The real and imaginary components can also be expressed in polar coordinates as magnitude and phase. Magnitude is given by the complex modulus; for a complex scalar $\mathbf{z} = a + jb$ the complex modulus is defined as

$$|\mathbf{z}| \stackrel{\text{def}}{=} \sqrt{a^2 + b^2}.$$

The phase is a measure of the angle in from the positive real axis, and principle values are constrained to $(-\pi, \pi]$ such that angles in the counter-clockwise direction are positive, and angles in the clockwise direction are negative. The two-argument arctangent function[16] defined as

$$\operatorname{atan2}(a,b) \stackrel{\text{def}}{=} \begin{cases} \operatorname{atan}\left(\frac{a}{b}\right) & b > 0 \\ \pi + \operatorname{atan}\left(\frac{a}{b}\right) & a \geq 0, b < 0 \\ -\pi + \operatorname{atan}\left(\frac{a}{b}\right) & a < 0, b < 0 \\ \frac{\pi}{2} & a > 0, b = 0 \\ -\frac{\pi}{2} & a < 0, b = 0 \\ undefined & a = 0, b = 0 \end{cases}$$

Empirically recorded signals will almost always yield both real and imaginary components, and the informative value of these components is related to their magnitudes and phases.[17] Because the Fourier Transform of real valued functions is Hermitian the analysis of only positive frequencies is sufficient to characterize the functions.

Figure 3.4.15 through Figure 3.4.17 present discrete Fourier transforms of signals we have some familiarity with in terms of their real and imaginary components and magnitude and phases so the reader can integrate the concepts they have learned so far. Figure 3.4.15 presents the discrete Fourier transform of a pure cosine wave. The top panel presents the original sequence x_n.

---

[16] Many computer implementations treat the atan2(0,0) as 0 instead of undefined. These include Octave/Matlab, Numpy/Python, and C.

[17] With some practical applications it may be necessary to "unwrap" or recover phase information (by taking the 2 π complement of the phase) for discontinuities larger than π.

The 2nd and 3rd panels present the real and imaginary coefficients, and the 4th and 5th panels present the coefficients as magnitudes and phase. The x-axis on panels 2-5 represents the discretized basis function frequencies in terms of n. These are often called frequency bins. Figure 3.4.16 presents the Fourier transform of a complex sinusoid with a frequency of $2\pi 10/N$. As we would expect the sinusoid has a peak in the corresponding positive frequency but zero magnitude over the negative spectrum. Figure 3.4.17 presents the square wave function we previously represented as a Fourier series. The examples depicted in the figures mentioned above are commensurate with our theoretical understanding of spectral analysis. From the plots the reader should be able to grasp how the real and imaginary components can also be represented as magnitude and phase.

3.4.3.6 Spectral Leakage. To be candid, the previous discrete Fourier transform examples are purposefully contrived. Often times the transforms are not as straightforward as these previous examples would suggest. Figure 3.4.18 depicts the discrete Fourier transform of $\cos(7.5\pi k/N)$. From the plot we can see that bins 7 and 8 reflect the energy of the cosine wave, but a significant amount of power is also present in adjacent bins. This phenomena is called *spectral leakage* and is consequence of the input sequence being finite and the sequence as a whole being treated as if it were periodic. From panel A we can see that the sine wave completes 7.5 cycles over the interval. The discontinuity of the half cycles is what generates the spectral leakage. Discontinuities contain broadband energy that gets distributed throughout the spectrum (Harris, 1978). In Figure 3.4.19 we have shifted the input sequence so the end and beginning of the sequence wrap nicely and the discontinuity is in the middle. We can see the phases change to accommodate the shift but the magnitudes and spectral leakage are identical to Figure 3.4.18. To reduce spectral leakage one can apply a variety of techniques. If the input signal is influenced by a trend, the trend should be removed before applying DFT. If the signal under examination is periodic examining an integer number of periods can reduce spectral leakage. Also analyzing longer intervals of time will reduce spectral leakage.

Figure 3.4.15 *The discrete Fourier transform of a cosine wave.*

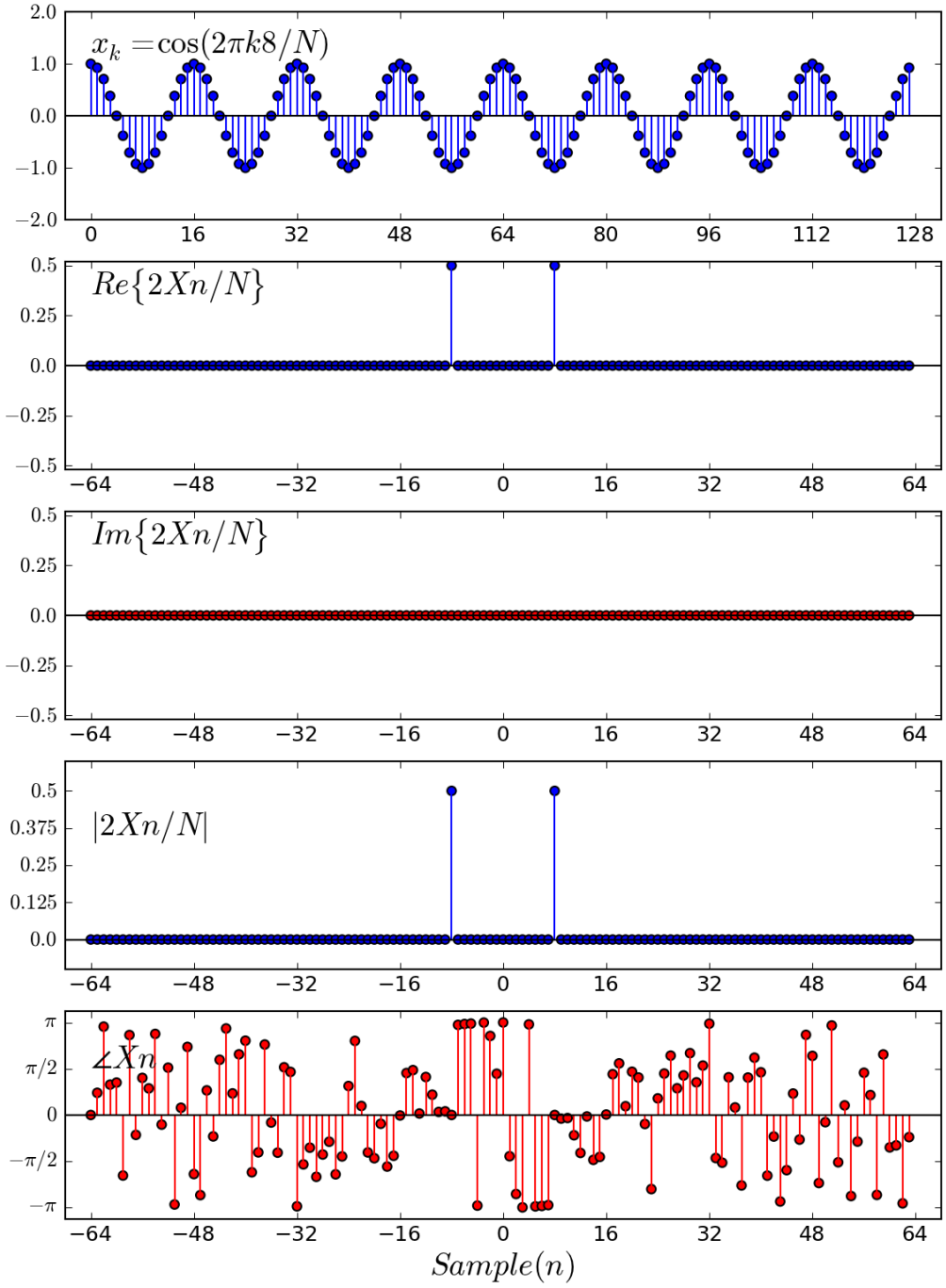


Figure 3.4.16 *The discrete Fourier transform of a complex sinusoid.*

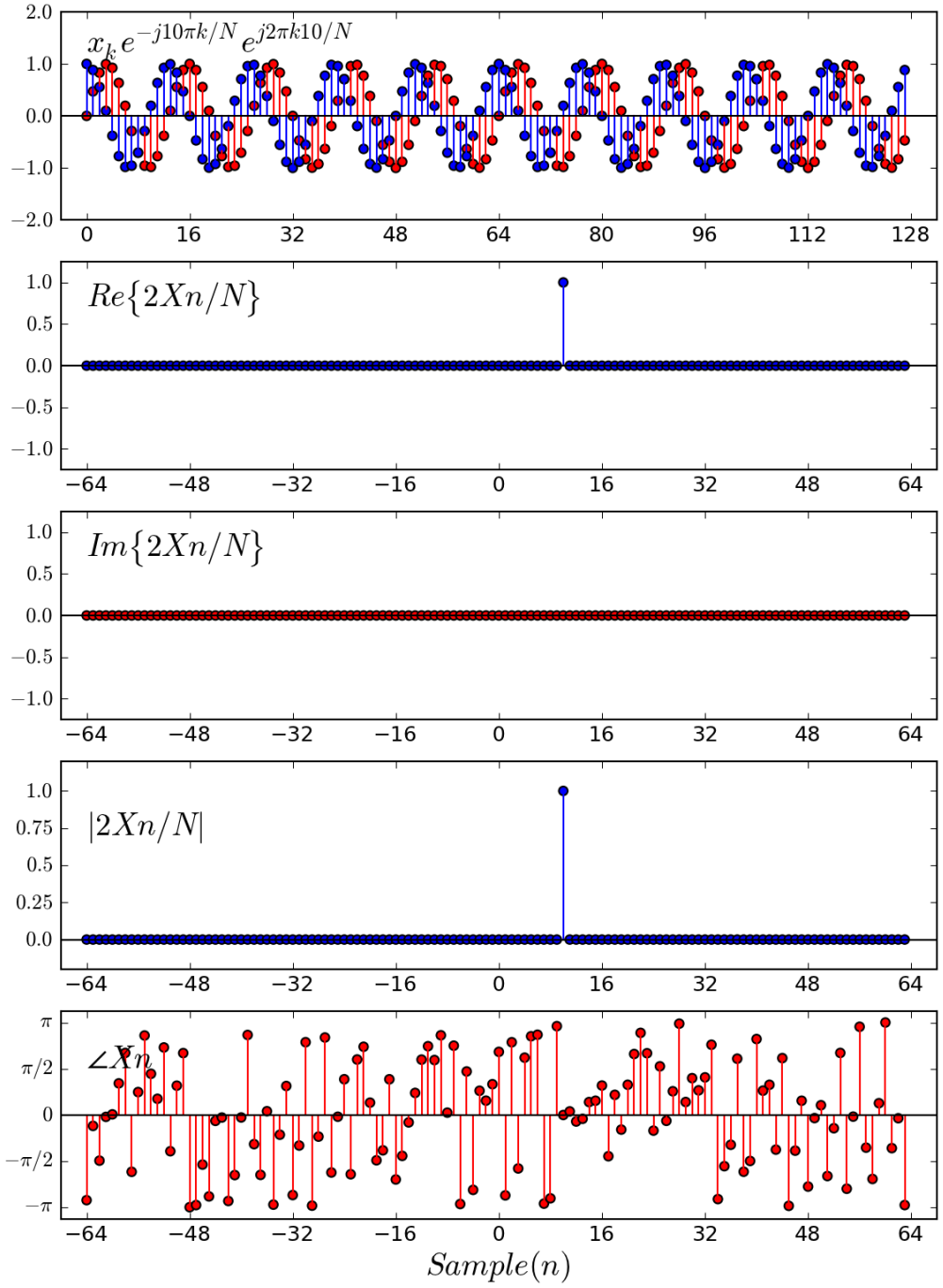


Figure 3.4.17 *The discrete Fourier transform of a square wave.*

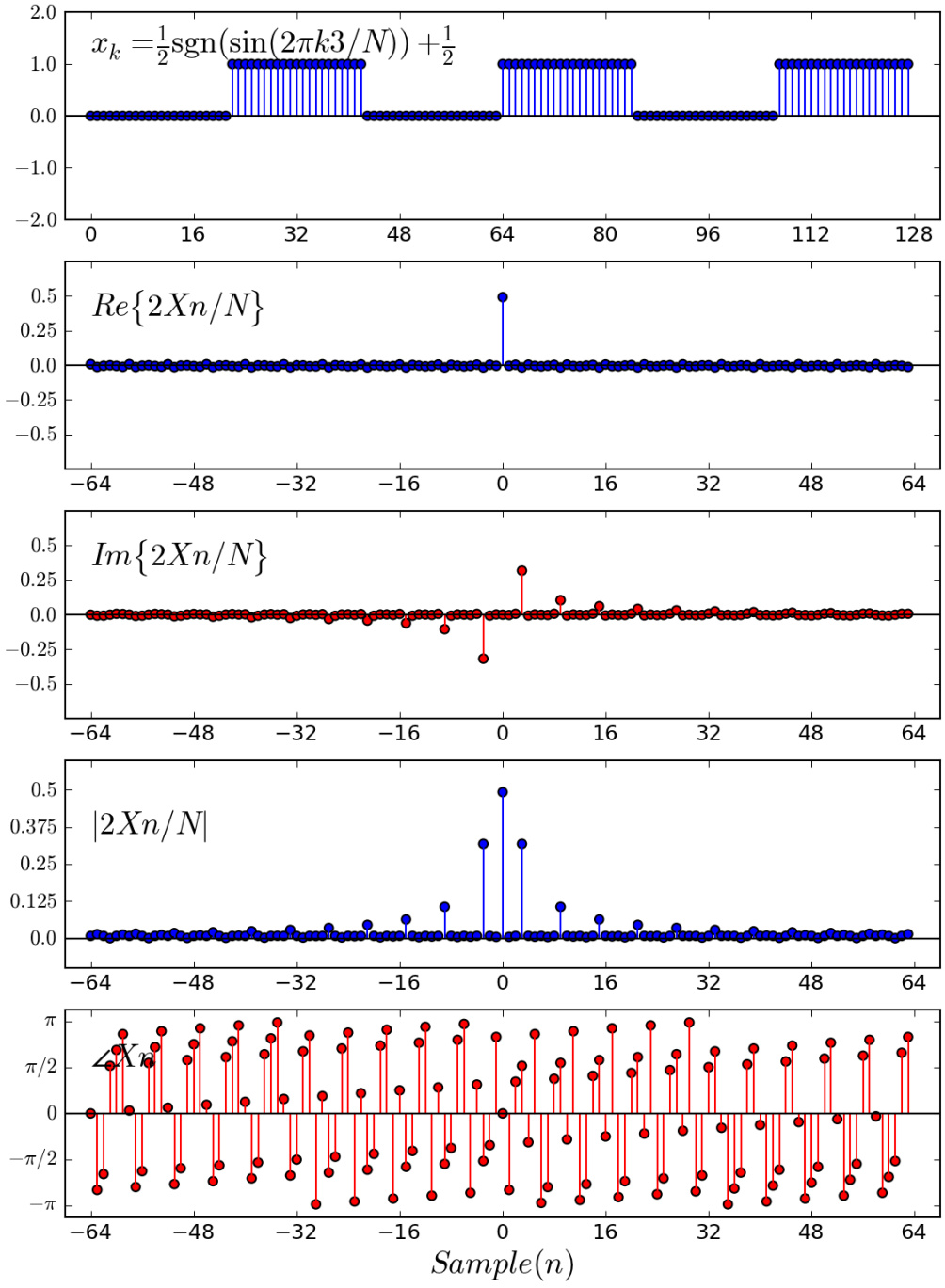


Figure 3.4.18 *Spectral leakage. When the input sequence has energy at frequencies between bins the energy leaks into adjacent bins. This phenomena is called spectral leakage.*

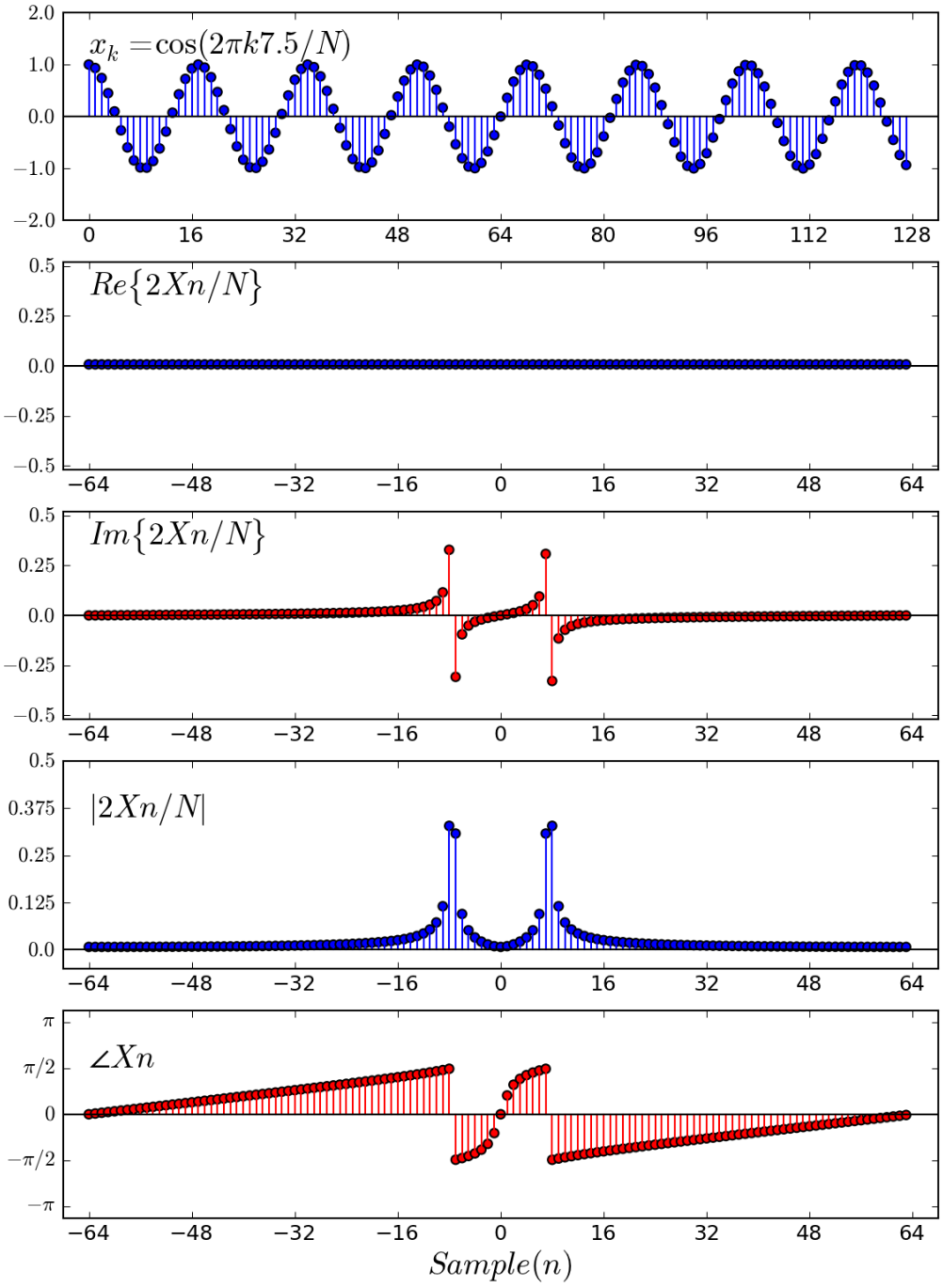


Figure 3.4.19 *The DFT treats input signal as if it were periodic. Discontinuities between the end and beginning of the input sequence cause the spectral leakage. This plot shows that when the discontinuity is placed in the middle of the sequence the magnitude from the discrete Fourier transform remain identical.*

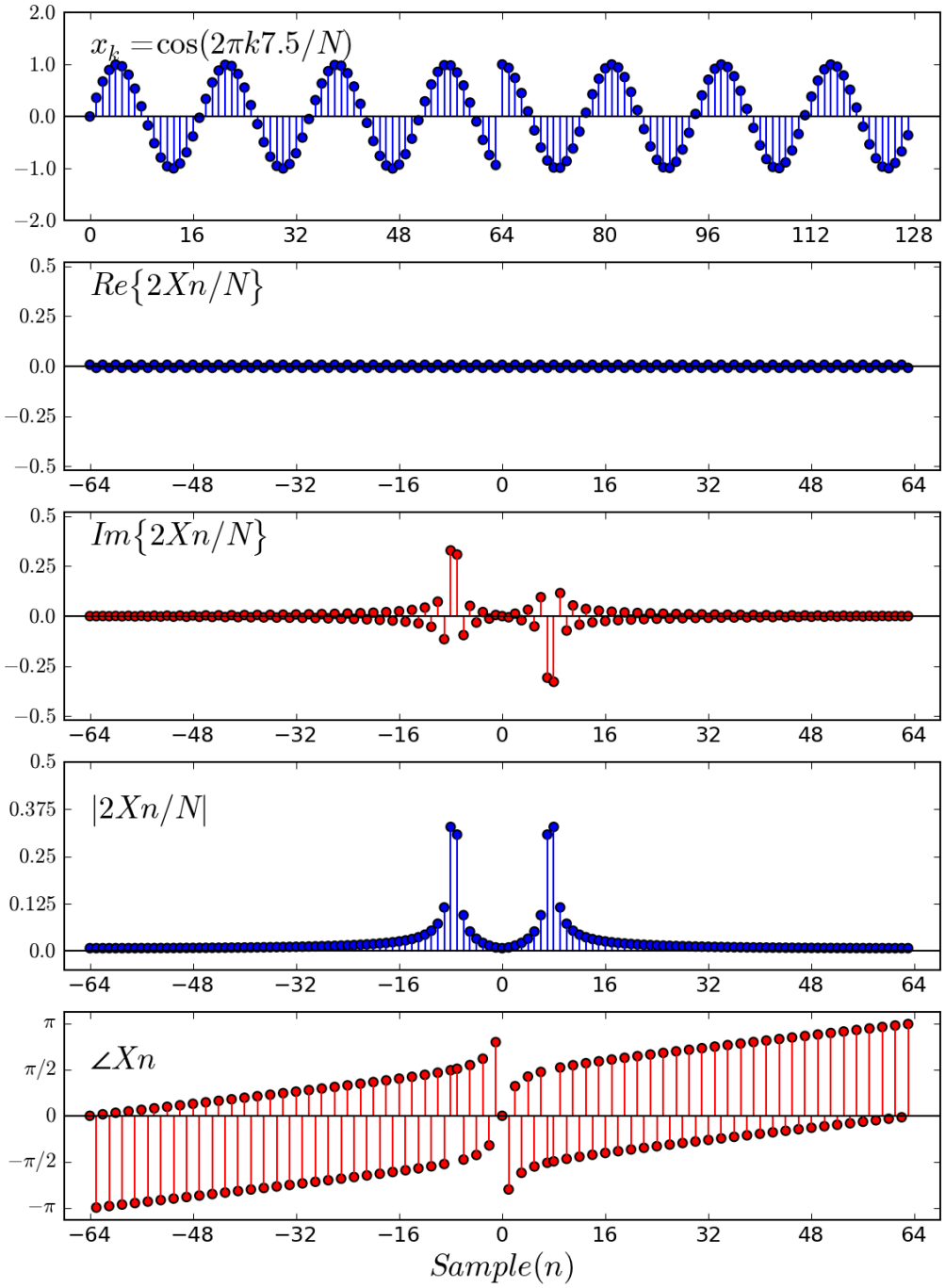
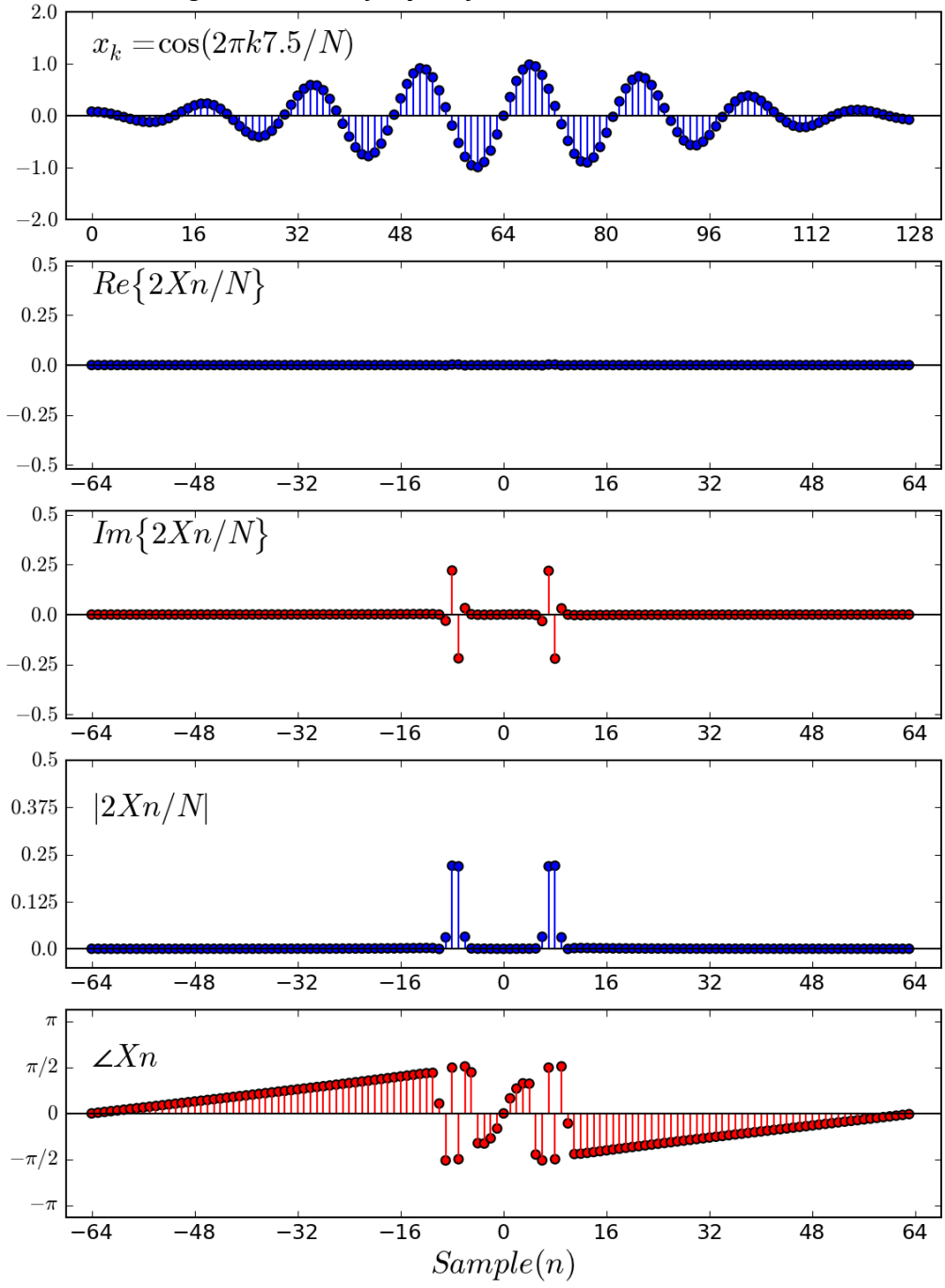


Figure 3.4.20 *Windowing and spectral leakage.*
Windowing the input sequence can control the spectral leakage but it trades off with reduced gain and lower frequency resolution.



Beyond these simple heuristics we must look to windowing. Window functions are what make the signals finite by controlling the interval of the signal that is being multiplied with the complex sinusoids. Although, we did not explicitly mention windowing the previous DFT examples have all employed rectangular windows defined as windows that are 1 inside the interval and zero outside the interval. Given a window sequence $\mathbf{w}_k$ the discrete Fourier transform can be defined as

$$\mathbf{X}_n \stackrel{\text{def}}{=} \sum_{\mathrm{k}=0}^{\mathrm{N}-1} \mathbf{w}_k \mathbf{x}_k e^{-j2\pi kn/\mathrm{N}}, \text{for } n = 0, \ldots, \mathrm{N}-1.$$

This equation makes the point that the DFT not only captures the spectrum of the input sequence but also the window. To understand spectral leakage we need to understand how these interact.

To try and control the spectral leakage we can apply a *window* that ramps the signal up at the beginning and ramps the signal down at the end thereby reducing the discontinuity. Figure 3.4.20 applies a Hamming window before performing the DFT (Hogwei, 2009). As we can see from the magnitude panel the spectral leakage is greatly reduced (at least with a linear magnitude axis). The consequence, as we can also see from the Figure; is that amplitudes in bins 7 and 8 have less energy (46% less energy to be precise). In this particular example this is not problematic because the signal is virtually noise free. However, in circumstances with low signal to noise ratios this might be problematic. Window function choices are about compromise.

3.4.3.7 Spectral Interpolation. To get a better look at the spectral leakage we can *zero pad* the original sequence and take the FFT to interpolate the frequency spectrum (the *zero padding theorem*). The result does not affect the spectral leakage, it merely gives us a better picture of the leakage (Smith, 2007). Zero padding is adding zeros to increase the length of the original sequence. Zero padding in this manner is equivalent to the discrete time Fourier transform (DTFT) of a finite sequence. The use of FFT is to make the interpolation computationally efficient. To avoid confusion let's denote the length of the original sequence L instead of N. Then let's denote the length after padding M. The DTFT is then defined as,

$$\mathbf{X}_n \triangleq \sum_{k=0}^{N-1} \mathbf{x}_k e^{-j2\pi kn/M}, \text{for } n = 0, \ldots, M-1.$$

As before the k variable indexes our sequence and n variable relates to the frequency. If the frequency was $\omega_n = e^{-j2kn/N}$ with DFT it becomes $\omega_{n'} = e^{-j2knL/M}$ after zero padding. The DTFT, used in this manner, samples frequencies in between the basis functions of $\mathbf{x}_k$ but it does not increase the frequency resolution. Figure 3.4.21 depicts how the basis functions are interpolated when $L = 9$ and $M = 15$. The following section uses DTFT (zero padding and FFT) to examine the spectral leakage of a rectangular window, a Hamming window, and a Blackman-Harris4 window.

3.4.3.8 Window Functions. Here we take our discrete $\cos(7.5\pi k/N)$ waveform where N=128 and applied windowing. After windowing we then applied zero padding so that the total length of the sequence was 4096. Then we applied FFT to interpolate the magnitude response. The top panel of Figure 3.4.22 depicts the original input cosine sequence. The 2nd panel shows the interpolated frequency response magnitudes in decibels (dB).[18] The 3rd panel shows input sequence with a Hamming window applied. The discrete Hamming window with indexes of n and a length of N is defined as

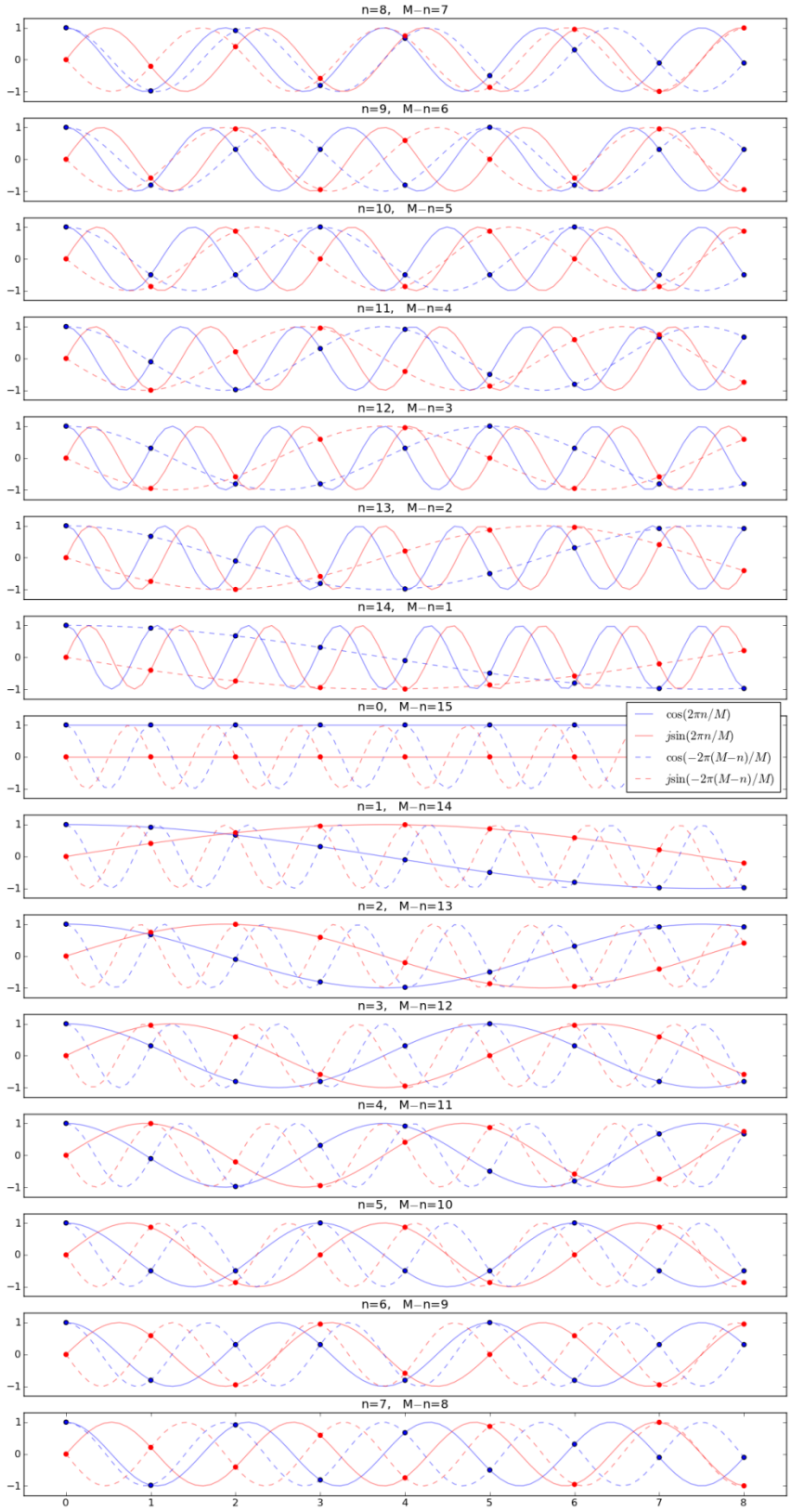
$$\mathbf{w}_k \triangleq 0.54 + .46\cos\left(\frac{2\pi k}{N-1}\right).$$

The 4th panel depicts the corresponding interpolated frequency magnitudes. The 5th and 6th panels show the same results with a discrete Blackman-Harris 4 window, a high dynamic range window, defined as,

$$\mathbf{w}_k \triangleq 0.3232153788877343 - 0.4714921439576260\cos\left(\frac{2\pi k}{N-1}\right)$$
$$+0.175534129960197\cos\left(\frac{4\pi k}{N-1}\right) - 0.0284969901061499\cos\left(\frac{4\pi k}{N-1}\right)$$
$$-0.0284969901061499\cos\left(\frac{4\pi k}{N-1}\right) + 0.0012613570882926\cos\left(\frac{4\pi k}{N-1}\right).$$

[18] Decibels is a base-10 logarithmic scale used to express the ratio between two values of a physical quantity.

Figure 3.4.21 *DTFT basis functions with L = 9 and M = 15.*



In the magnitude plots the peaks are called *lobes.* In each, the highest lobe corresponding to the cosine's frequency is called the *main lobe.* All the other lobes are called *side lobes.* When frequency components fall between the lobes the magnitude estimate becomes biased downwards. This is called *scalloping loss.* When we compare across the three different window magnitudes we can see that the main lobe is the narrowest with the rectangular window, and is wider for the Hamming, and even wider for the Blackman-Harris 4. The width of the main lobe relates to the window's frequency resolution. The rectangular window has the best frequency resolution, but it also has the poorest *dynamic range.* Take note that the range on the y-axis differs across the three magnitude plots.

The dynamic range refers to the ability of the window to suppress noise. It is quantified by assessing the difference between the peak of the main lobe and the peak of adjacent nodes in decibels as well as by measuring the rate at which the side lobes roll-off with respect to frequency (dB/decade). The rectangular window might be better in situations where signal components have roughly equivalent amplitudes and are close in frequency. Figure 3.4.23 depicts the discrete Fourier transforms of a signal composed of two sinusoids separated by only 1.6 bins. The higher frequency component also has a .54 radian phase shift. From the plots we can see the rectangular window can distinguish between the peaks, but the Hamming and Blackman-Harris4 cannot.

However, windows with high dynamic range are better when spectral components have disparate amplitudes and are not as close in frequency. Figure 3.4.24 we also have a signal composed of two sinusoids and random noise. One of the sinusoids has an amplitude of 1, and the second has an amplitude of .06 just slightly above the noise floor. In this the noise completely masks the second amplitude with the rectangular window. The Hamming and Blackman-Harris4 are able to detect it. Keep in mind the y-axes are scaled to reflect the dynamic range of the windows. The Blackman-Harris and Hamming have comparable performance when this is taken into consideration.

The Hamming and Hann windows are popular choices due to having "Goldilocks" characteristics: they have average dynamic range and average frequency resolution. The Gauss window is also popular due to having a shape parameter that can be tailored to the application. The Kaiser-Bessell window is designed to give an optimal mathematical tradeoff between frequency resolvability and main lobe width. There are literally dozens of windowing functions to choose from. Hogwei (2009) has published measurements for a collection of 55 windowing functions. Unfortunately, no single window function is clearly superior for every application. In this section we have seen that the resulting spectra from discrete Fourier transform is not just dependent the complex sinusoidal basis functions. The choice of window, or more precisely, how the window is multiplied with the interval affects the resulting spectral estimates. We are "deep down the rabbit hole" and perhaps stepping back to gain some perspective will help. Our goal with DFT is to estimate the spectra of a signal. The problem is that whenever we examine signals of finite length we must take the transform of the signal multiplied by a window. With the DFT there is no way to separate the true spectra from the leakage. However, we can separate them analytically using the continuous Fourier transform.

3.4.3.9 Convolution and the Convolution Theorem. Before we get to looking at analyzing the windowing using the continuous Fourier transform we first need to discuss convolution. Convolution is an operation between of two functions. One function is reflected (reversed) and slide as a function of the independent variable across the second function. While sliding the functions are multiplied and integrated with one another. The result reflects the degree of overlap between the two functions similar to how the inner product of the Fourier transform is a measure of similarity between the $x(t)$ with the complex sinusoid $e^{-j\omega t}$. The convolution operator is usually the asterisk symbol. For two functions $g(t)$ and $h(t)$ the convolution operation is defined as

$$(g * h)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} g(t-\tau)h(\tau)\,\mathrm{d}\tau = \int_{-\infty}^{\infty} h(t-\tau)g(\tau)\,\mathrm{d}\tau.$$

Figure 3.4.22 *Windowing and zero-padding.Windowed and zero padded input sequences reveal the interaction between the input sequence and windowing functions. y-axis units for 2nd ,4th, and 6th panels are dB.*

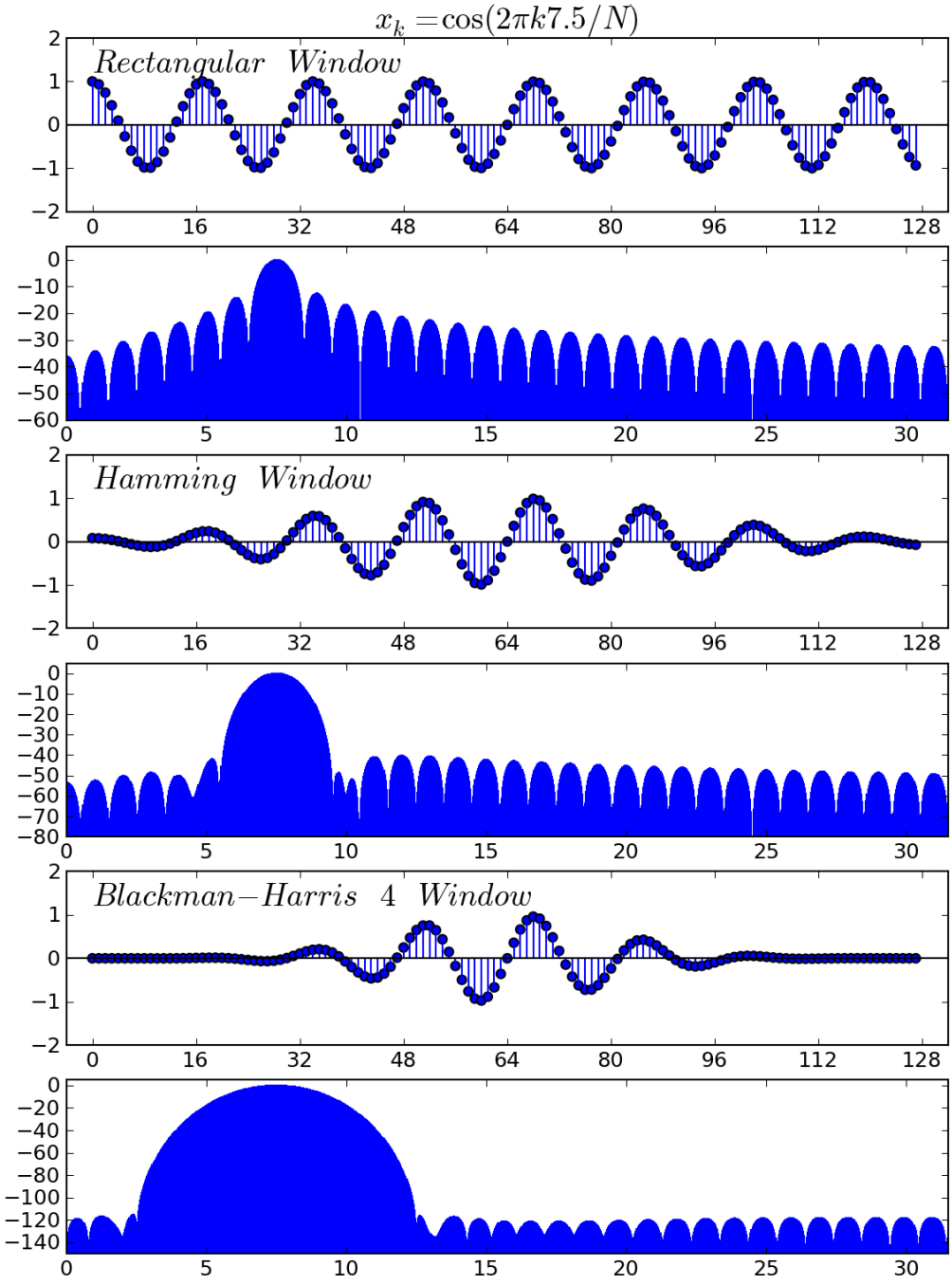


Figure 3.4.23 *Low dynamic range windowing.When signals have components that are roughly equivalent amplitude and close in frequency windows with low dynamic range and high frequency resolution, like the rectangular window, are better at distinguishing spectral components. y-axis units for 2nd ,4th, and 6th panels are dB.*

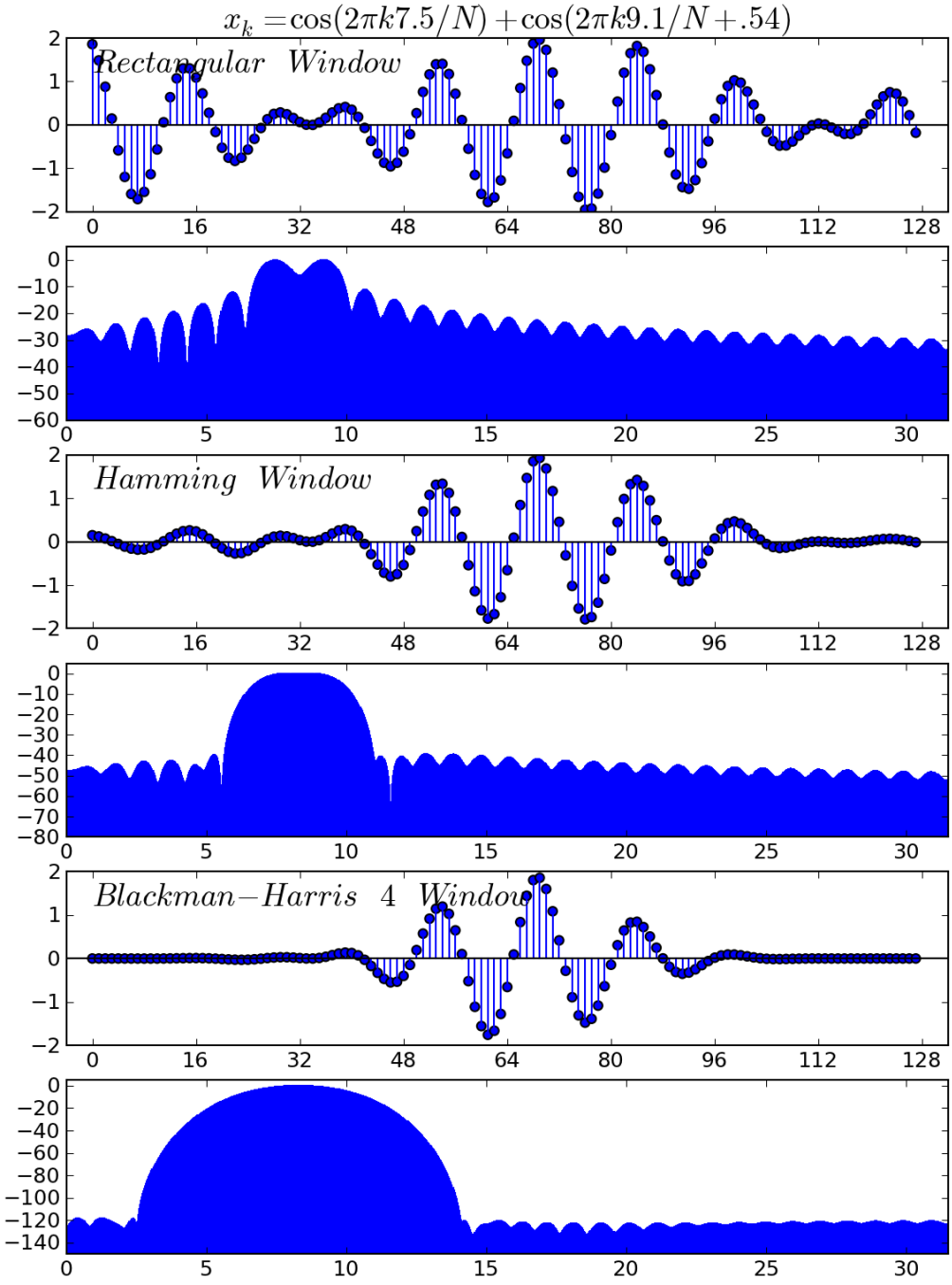
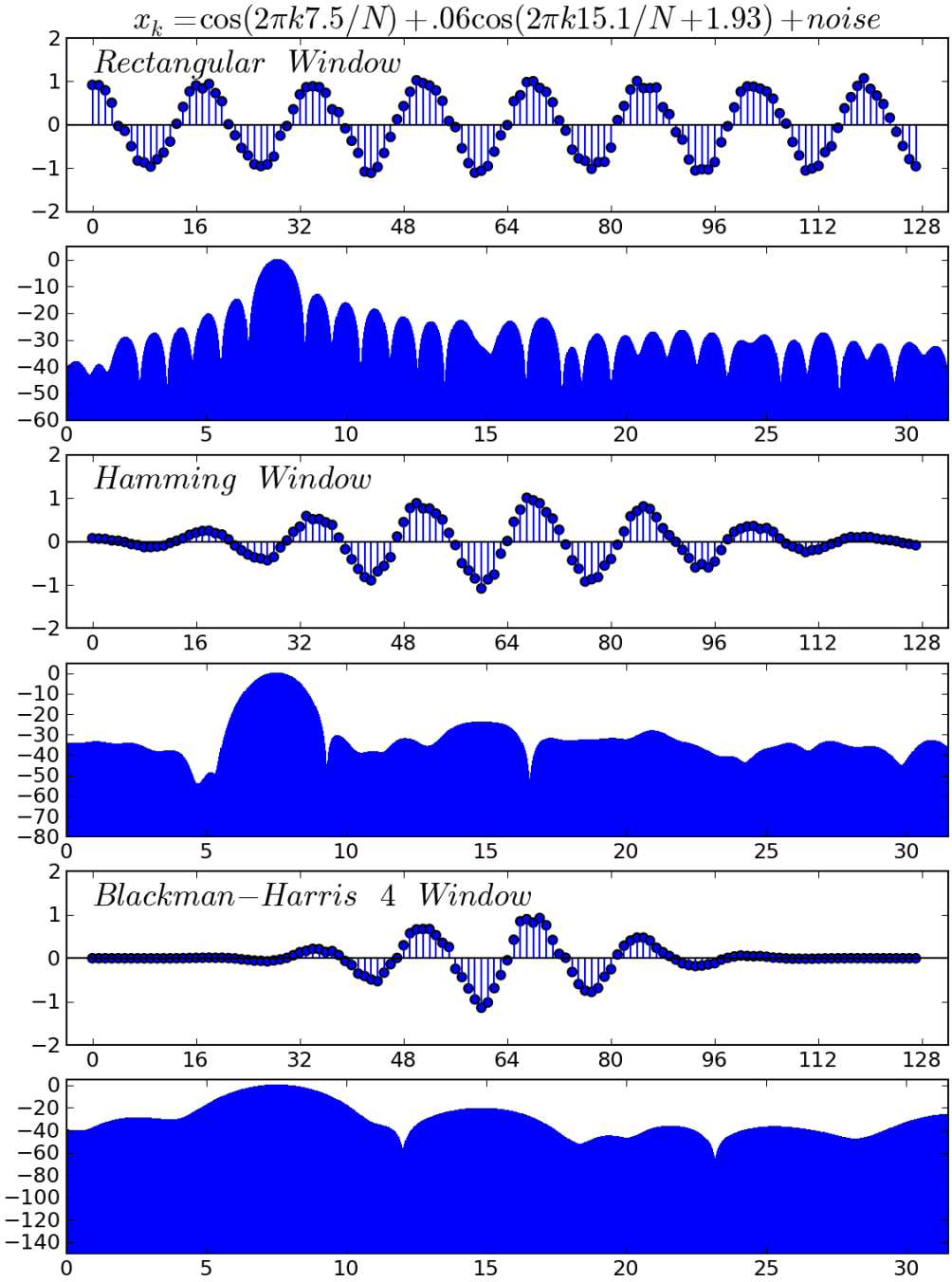


Figure 3.4.24 *High dynamic range windowing. When signals have components that are disparate in amplitude and not extremely close in frequency windows with higher dynamic range and lower frequency resolution are better at distinguishing spectral components. y-axis units for 2nd ,4th, and 6th panels are dB.*



Convolution is related to the Fourier transform through the convolution theorem (Weisstein, Autocorrelation, 2011). The convolution theorem for two functions $g(t)$ and $h(t)$ and their respective transforms $\mathcal{F}\{g(t)\}$ and $\mathcal{F}\{h(t)\}$ states that the Fourier transform of $g(t) * h(t)$ is equivalent to the products of their Fourier transforms,

$$\mathcal{F}\{(g * h)(t)\} = \mathcal{F}\{g(t)\}\,\mathcal{F}\{h(t)\}.$$

The meaningful implication of this result is that the convolution between of $g(t)$ and $h(t)$ can exploit the efficiency of FFT algorithms:

$$(g * h)(t) = \mathcal{F}^{-1}\{\mathcal{F}\{g(t)\}\,\mathcal{F}\{h(t)\}\}.$$

The corollary is that the Fourier transform of two functions which have been multiplied in the time domain is equivalent to the convolution of their Fourier transforms in the frequency domain,

$$\mathcal{F}\{g(t)h(t)\} = \mathcal{F}\{g(t)\} * \mathcal{F}\{h(t)\}.$$

This expression is of importance to our current goal.

3.4.3.10 Analytic View of Spectral Leakage. Now we can get back to the task at hand by first defining the rectangular window function as

$$\Pi(\tau) = \begin{cases} 0 & \text{if } |\tau| > \frac{1}{2}, \\ \frac{1}{2} & \text{if } |\tau| = \frac{1}{2}, \\ 1 & \text{if } |\tau| < \frac{1}{2}. \end{cases}$$

Using the convolution theorem we can represent the integration of $\cos(\omega_0 t)$ over a finite interval as

$$X(\omega) = \int \cos(\omega_0 t)\,\Pi(L^{-1}t)e^{-j\omega t}\,\mathrm{d}t$$

$$= \mathcal{F}\{\cos(\omega_0 t)\} * \mathcal{F}\{\Pi(L^{-1}t)\}$$

where the ω_0 parameter is the frequency of the cosine function and the L parameter controls the size of the rectangular window. When ω_0 is an integer multiple of 2π we can set L to be integer multiples of π and the domain will be restricted to full periods of the $\cos(\omega_0 t)$. Now we need to find the transforms for $\cos(\omega_0 t)$ and $\Pi(L^{-1}t)$.

Earlier we said that a function must be square integratable to be transformable. Now we are going to be make an exception to that rule so we can find the Fourier transform of $\cos(\omega_0 t)$.

$$\mathcal{F}\{\cos(\omega_0 t)\} = \int \cos(\omega_0 t)\, e^{-j\omega t}\, \mathrm{d}t$$

$$= \int \frac{e^{j2\pi\omega_0 t}}{2} + \frac{e^{-j2\pi\omega_0 t}}{2} e^{-j\omega t}\, \mathrm{d}t$$

$$= \int \frac{e^{-j(\omega+\omega_0)t}}{2} + \frac{e^{-j(\omega-\omega_0)t}}{2}\, \mathrm{d}t$$

[To go further we have to use one of the identities for the Dirac delta function $\delta(t)$,

$$\delta(\xi_1 - \xi_2) = \int e^{-j(\xi_1-\xi_2)t}\, \mathrm{d}t.$$

Recall from section 2.3.3 that the Dirac delta is defined as $\delta(t) \stackrel{\text{def}}{=} \begin{cases} +\infty, & t = 0 \\ 0, & t \neq 0 \end{cases}$ and has a defined integral of $\int \delta(t)\, \mathrm{d}t \stackrel{\text{def}}{=} 1$.]

$$\mathcal{F}\{\cos(\omega_0 t)\} = \frac{1}{2}[\delta(\omega + \omega_0) + \delta(\omega - \omega_0)]$$

From this we can see that the Fourier transform of a cosine with a frequency ω_0 is an infinite impulse at $-\omega_0$ and a second impulse at ω_0. Now to the rectangle function.

$$\mathcal{F}\{\Pi(L^{-1}t)\} = \int \Pi(L^{-1}t) e^{-j\omega t}\, \mathrm{d}t$$

$$= \int_{-L}^{L} e^{-j\omega t}\, \mathrm{d}t$$

$$= \frac{e^{-j\omega t}}{-j\omega}\Bigg|_{-L}^{L}$$

$$= \frac{e^{j\omega L} - e^{-j\omega L}}{j\omega}$$

$$= \frac{2\sin(\omega L)}{\omega}$$

Some readers may recognize this result as the sinc function. The Fourier transform of a window is called a Fourier kernel or a window kernel. The top panel of Figure 3.4.25 plots the window kernel

function where $L = \pi$. The middle panel plots the absolute value of the function, and the bottom panel plots the absolute value in decibels. See anything familiar? The decibel magnitudes have the exact same scallop structure we observed earlier using zero padding. The scallop losses occur where the sinc function has zero crossings. To provide further insight consider the *sifting property* of Dirac delta function $\delta(t-\tau)$ when it is convolved with a function $f(t)$

$$f(t) * \delta(t-\tau) = \int f(t')\delta(t-\tau-t')\,\mathrm{d}\tau$$

$$= \int f(t')\delta(t'-(t-\tau))\,\mathrm{d}\tau \quad [\delta \text{ function is symmetric}]$$

$$= \boldsymbol{f(t-\tau)}$$

In our case this means that we can find the Fourier transform of

$$\mathcal{F}\{\cos(\omega_0 t)\,\Pi(L^{-1}t)\} = \mathcal{F}\{\cos(\omega_0 t)\} * \mathcal{F}\{\Pi(L^{-1}t)\}$$

$$= \frac{1}{2}[\delta(\omega+\omega_0)+\delta(\omega-\omega_0)] * \frac{2\sin(L\omega)}{\omega}$$

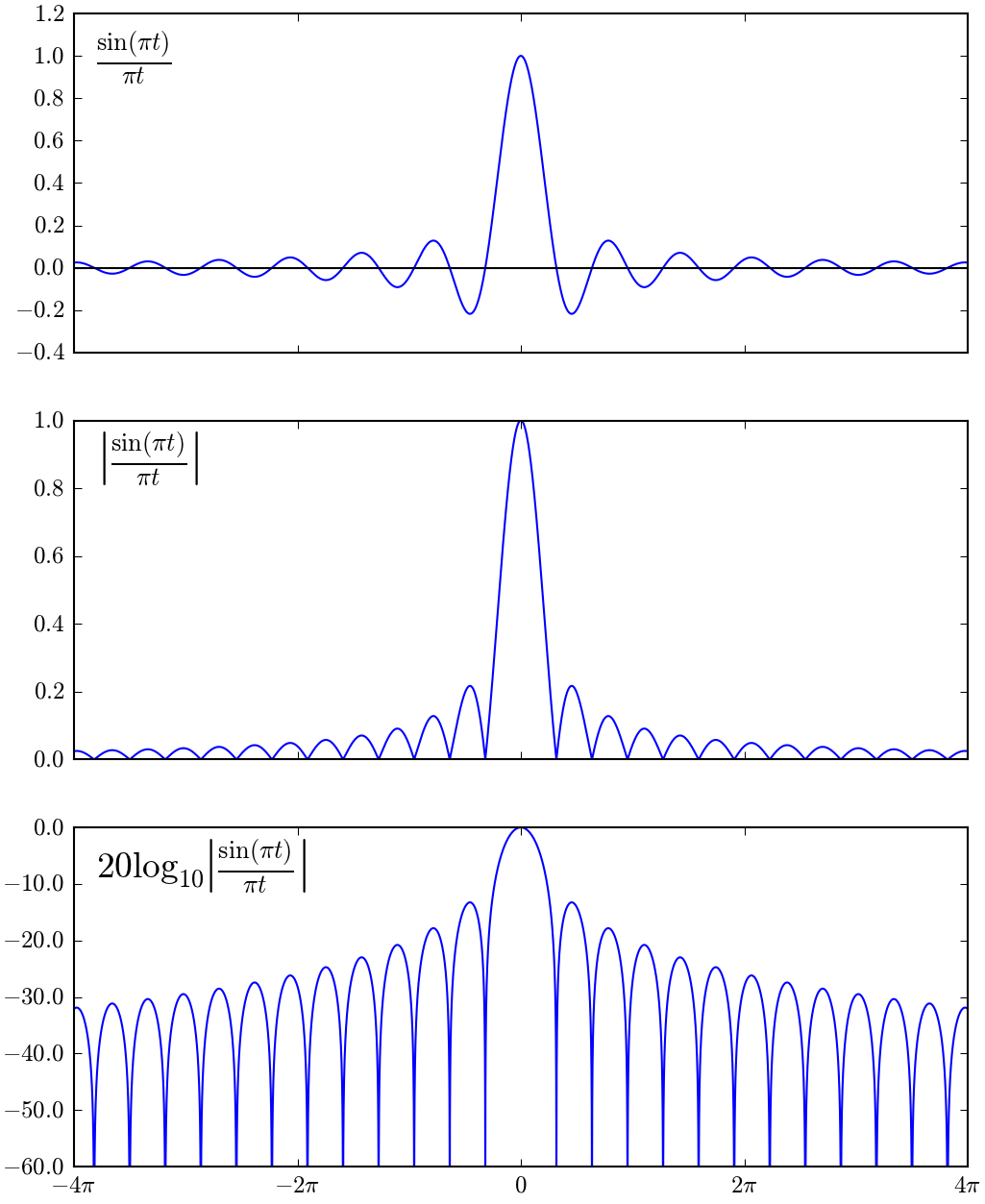
$$= \frac{1}{2}\delta(\omega+\omega_0) * \frac{2\sin(L\omega)}{\omega} + \frac{1}{2}\delta(\omega-\omega_0) * \frac{2\sin(L\omega)}{\omega}$$

$$= \frac{\sin(L(\omega+\omega_0))}{(\omega+\omega_0)} + \frac{\sin(L(\omega+\omega_0))}{(\omega+\omega_0)}$$

The interpretation of this result is that the Fourier transform of a bounded cosine function of frequency ω_0 is a sinc function that is dependent on the bounds centered about $\pm\omega_0$. We can see what this looks like by setting $\omega_0 = 2\pi$ and setting L equal to integer multiples of 2π so the cosine function will always complete full periodic cycles over L. In Figure 3.4.26 we can see the results of these assumptions. From the plots we can see that as L increases the spectral leakage decreases and the results get closer to the true spectra of $\cos(\omega_0 t)$.

The result above only applies precisely to a pure cosine with a frequency of $\omega_0$and a rectangular window centered about zero, and extending by L in both positive and negative directions, but similar treatments could be performed with other signals and other windows. The general implications of such exercises would be that the window acts as a *band pass filter*.

Figure 3.4.25 *Fourier transform of a rectangular window.The top panel plots the Fourier transform of a rectangular window: the sinc function. The middle panel shows the absolute value of a sinc function. The bottom panel plots the absolute value in decibels.*



The main lobe corresponds to the *pass band* of the filter. Frequencies in the pass band are let through while frequencies outside of pass band are attenuated. When we discuss using different windows we are actually discussing using different filters. The rectangular window has the narrowest passband but rolls off the slowest. The Blackman-Harris4 has a wider pass band but rolls off much faster. The DFT approximates the true spectra of the input signal convolved with the window kernel. Further analysis also shows that the passband in the frequency domain decreases as the size of the window in the time domain increases. This generalizes to other windows. Figure 3.4.27depicts how the passband of the Blackman-Harris4 decreases as the size of the window size increases. The consequence is that we cannot simply increase the sampling rate to increase the frequency resolution (assuming the sampling rate is sufficient for the bandwidth of the signal in the first place). The frequency resolution can only be increased from observing more periodic repetitions in the signal. Instead of merely talking about the theory we can show the theory.

3.4.3.1 From Theory to Application. We previously demonstrated in Figure 3.3.3.9 that when two spectral components are close to one another spectral leakage make them difficult if not impossible to distinguish. In the time domain the sinusoidal components had 7.5 and 9.1 cycles in 128 samples or 0.056 and 0.071 cycles/sample respectively. In the DFT the samples were separated by just 1.6 frequency bins. With the rectangular window two distinct peaks were barely visible with interpolation. With the Hamming and Blackman-Harris4 windows the components were not distinguishable. If we let extend the time signal so it is four times the length (512 samples) and maintain the signals at their original frequencies (0.056 and 0.071 cycles/sample respectively) we see that the increased frequency resolution makes the components much easier to distinguish (Figure 3.4.28)

In Figure 3.4.24 we observed how the spectral leakage from windows with low dynamic range can completely bury a low amplitude component in the presence of noise. We can revisit this example and see how extending the signals duration improves the delectability of periodic low

Figure 3.4.26 *Role of sample duration. As the number of cycles increase the spectral leakage decreases and the resulting transform more closely approximates the true spectra of the signal.*

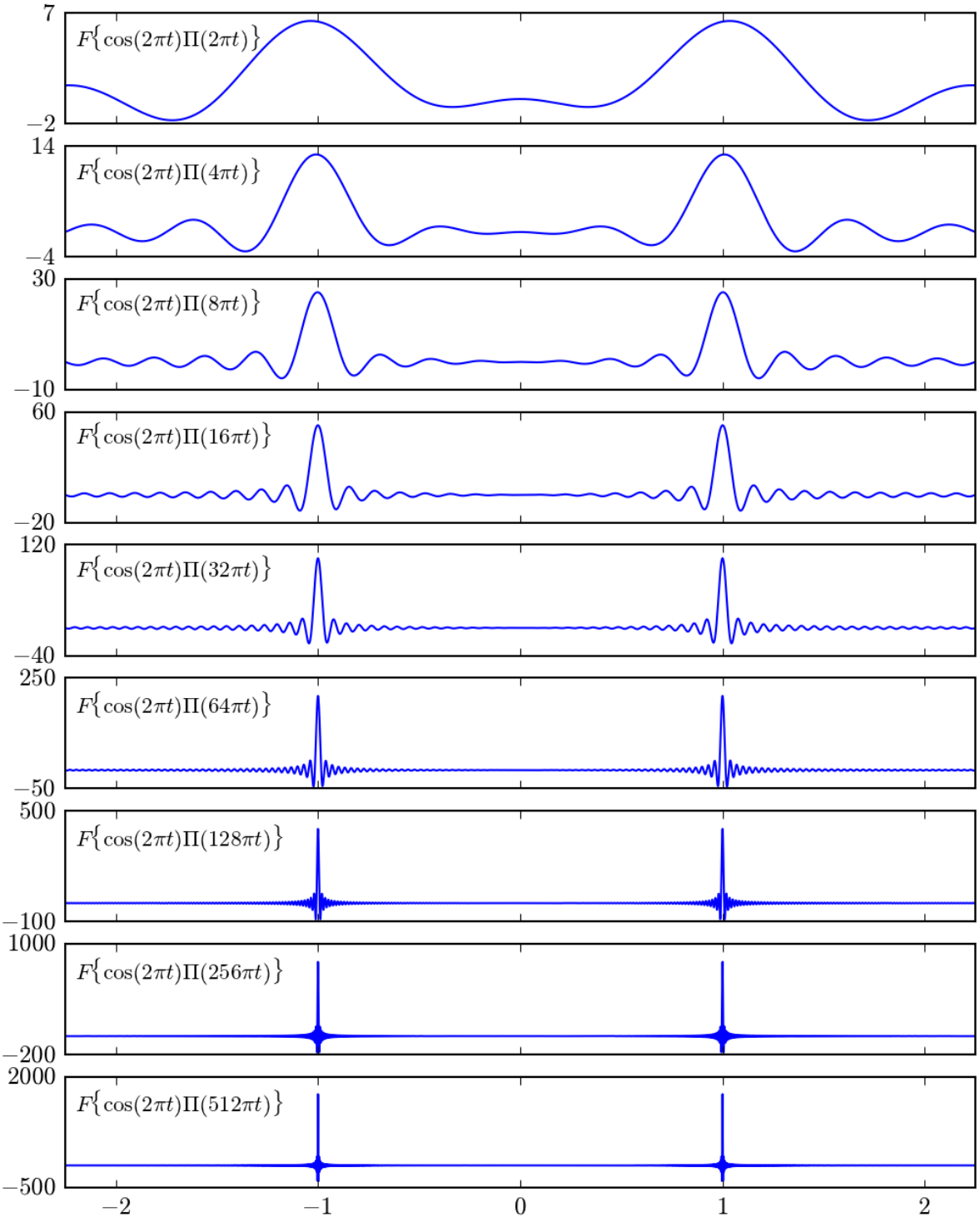
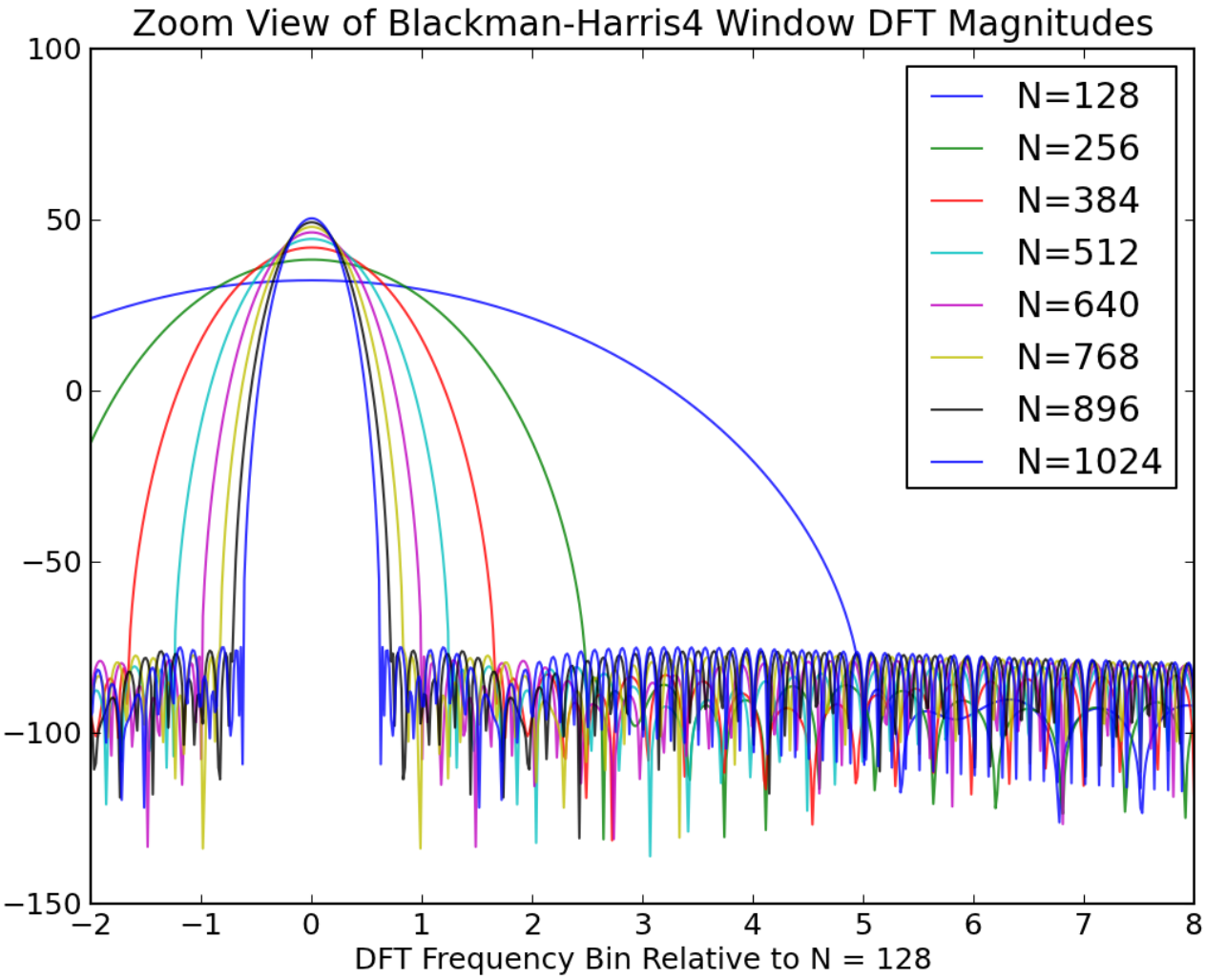


Figure 3.4.27 *Frequency resolution and signal duration. Frequency resolution is linked to the signals duration, not the sampling rate of the signal. As a signals duration increases the frequency resolution of the window functions increases. [Collecting data at a higher sampling rate or interpolating data does not increase frequency resolution.]*



amplitude components (see Figure 3.4.29). In circumstances where it is not possible, or not simple, to increase the duration of the signal under analysis the advantages shown here can still be exploited by windowing several signals and appending so they overlap one another. This technique is known as Welch's method.

3.4.4 ***Introduction to short-time Fourier transform.*** All the Fourier methods have presented so far have all assigned a single spectrum to the time interval under examination. In many instances the spectral characteristics of a signal may change over time. In these situations it can be informative to examine how the spectral characteristics changes over time. The short-time Fourier transformations (STFT) do just this by using a moving window to control the portion of the time signal that is transformed. Figure 3.4.30 presents a visual example known as a cumulative spectral decay (CSD) plot or waterfall plot, which uses discrete short-time Fourier transformation, to examine the resonance characteristics of a loudspeaker by taking windowed transformations of the loudspeaker's impulse response. The DFT of the impulse response yields the frequency response of the loudspeaker.[19] The plot frequency is on the x-axis, magnitude is on the y-axis, and the z-axis depicts time as the window moves forward over the impulse response. The *fall* in the back reflects the entire impulse response. As the falls move forward the impulse response is truncated and the loudspeaker's resonances show up as ridges along the z-axis.

3.4.4.1 Continuous STFT. Short-time Fourier transformations come in continuous and discrete varieties. The continuous short-time Fourier transformation is a transform of two parameters ω and τ,

$$X(\tau, \omega) \overset{\text{def}}{=} \int_{-\infty}^{\infty} x(t) w(t-\tau) e^{-j\omega t}\, \mathrm{d}t.$$

[19] The loudspeaker under examination was designed, fabricated, and measured by the author.

Figure 3.4.28 *High dynamic range windowing. When signals have components that are disparate in amplitude and not extremely close in frequency windows with higher dynamic range and lower frequency resolution are better at distinguishing spectral components. y-axis units for 2nd ,4th, and 6th panels are dB.*

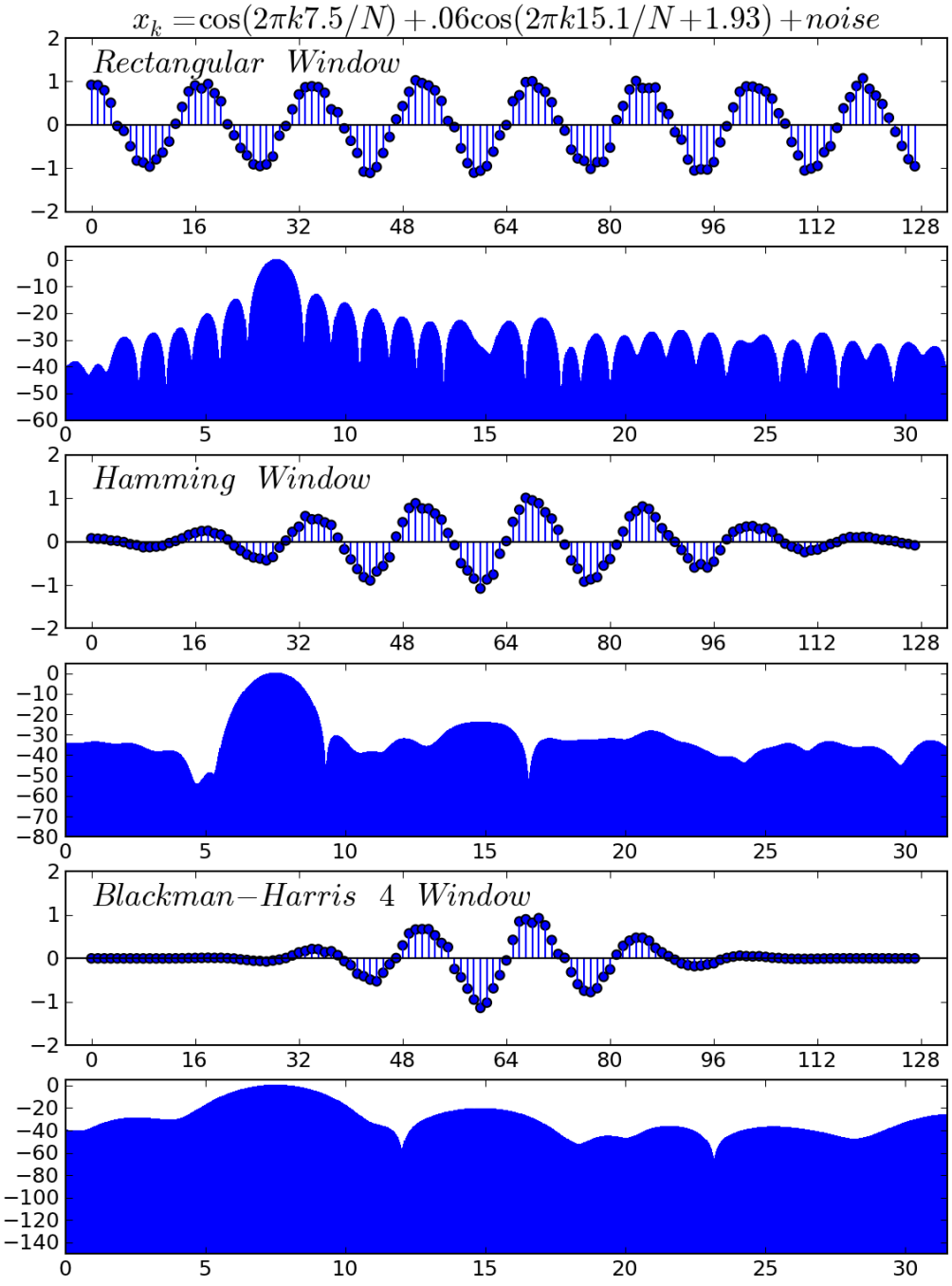


Figure 3.4.29 *Revisiting the high dynamic range example.Processing 4 times as much data also makes it much easier to detect a low amplitude signal in the presence of white noise.*

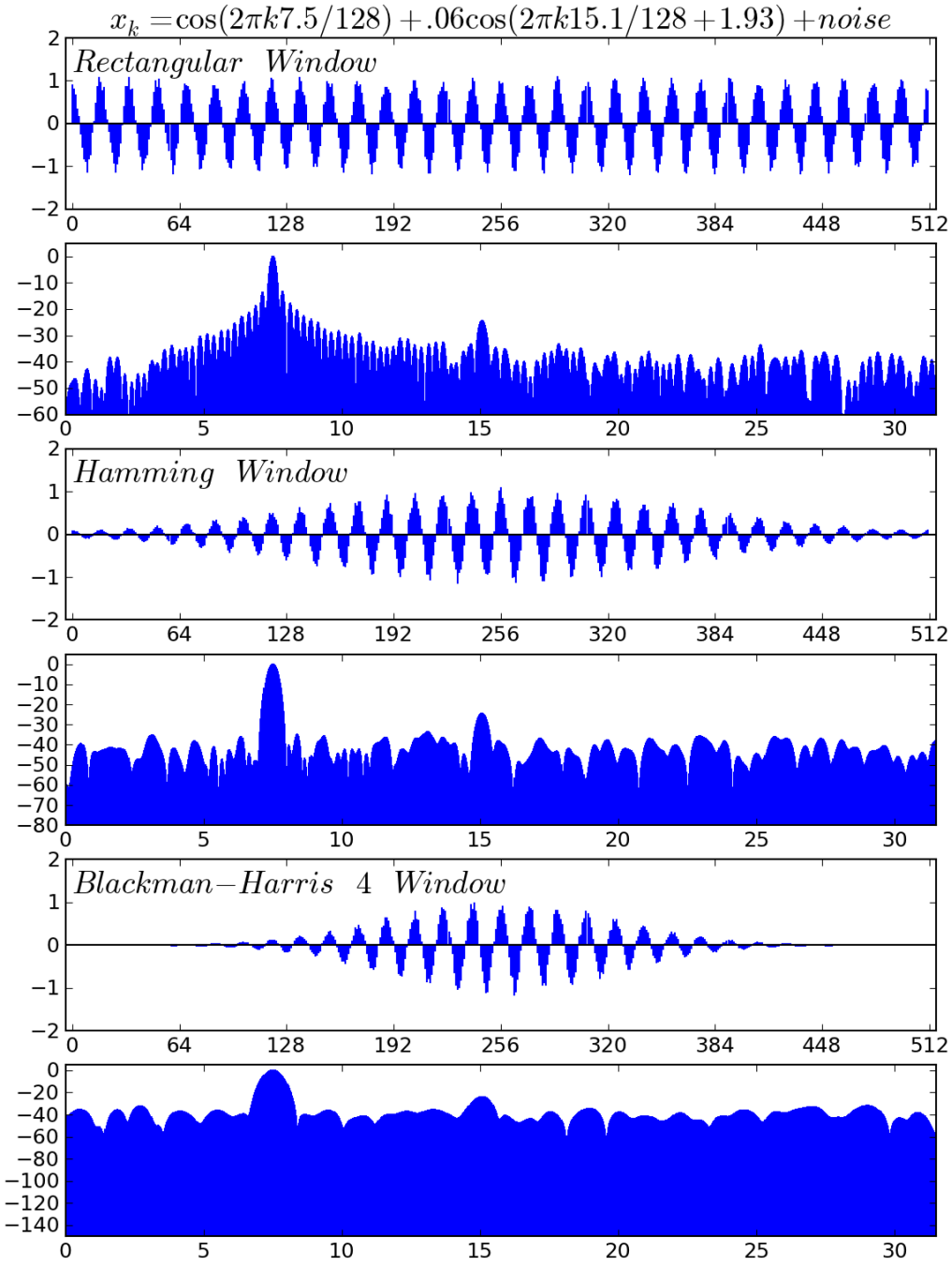
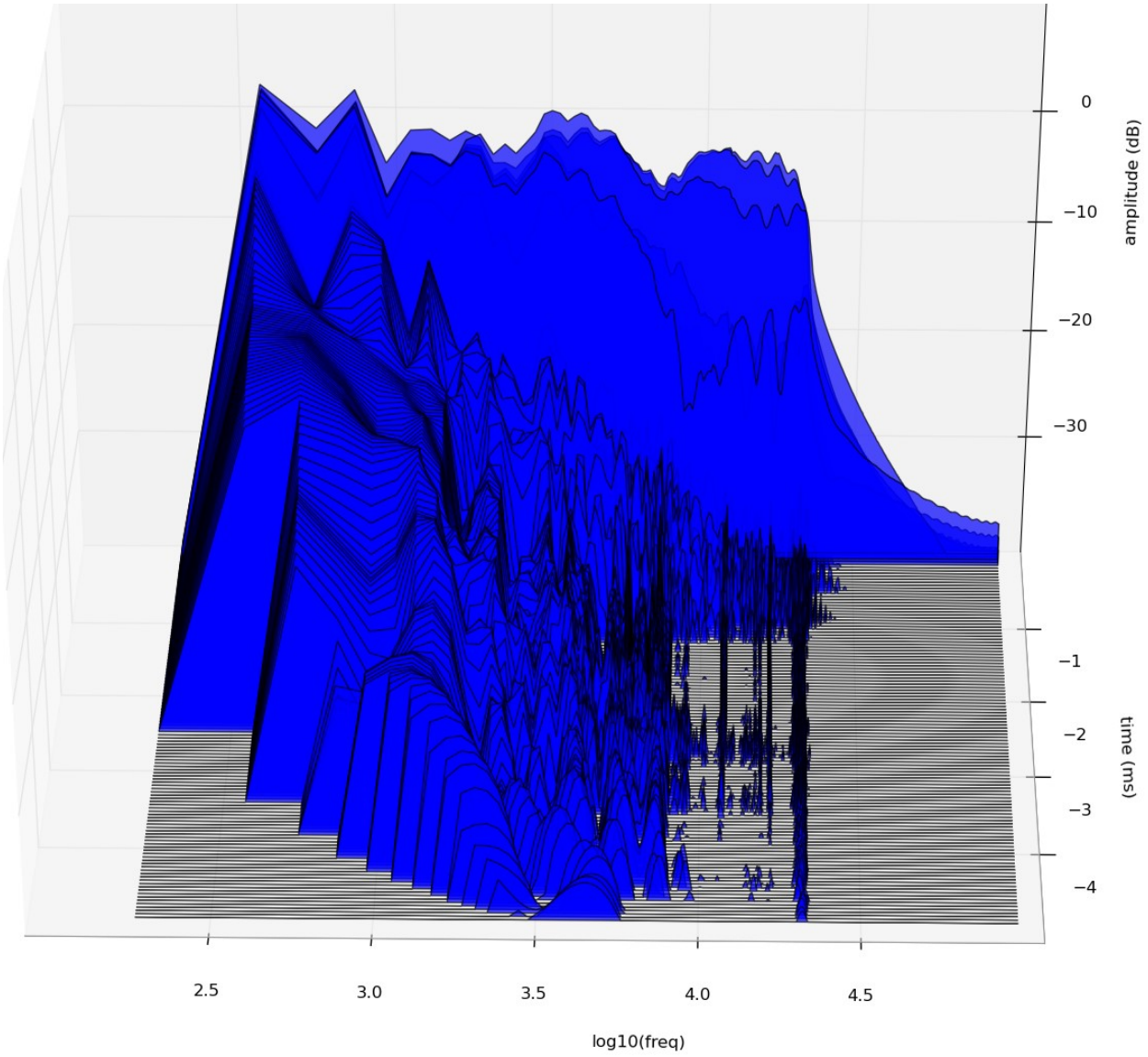


Figure 3.4.30 *A cumulative spectral decay (CSD) plot.By applying a moving windowed Fourier transform to the impulse response of a loudspeaker, the resonance characteristics can be examined.*



As with the previous transformations the ω variable specifies frequency. The function $x(t)$ is still the function of time that is being transformed. The $w(t-\tau)$ function is a window function. For the sake of completeness the inverse continuous short time Fourier transform is

$$x(t)w(t-\tau) \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\tau,\omega)e^{j\omega t}\,\mathrm{d}\omega.$$

Continuous window functions have the same utility as the discrete windows we presented in section 2.3.3.8. They are intended to limit the time signal and reduce the discontinuity between the beginning and end of the signal. The primary difference is that the discrete windows are symmetric about $N/2$ while the continuous windows are symmetric about 0. The τ variable is called the translation variable because it specifies how much the window should be translated. Figure 3.4.31 depicts the how the window function translates over time and the function $x(t)$ is multiplied by the window function. In our earlier discussion of windowing we discovered longer windows yield better frequency resolution. With STFT this fact remains.

Windows of longer duration will have less spectral leakage and better frequency resolution, but the increased frequency resolution is at the cost of time resolution. Likewise, short windows will have good time resolution, that is better precision in distinguishing changes in time, but poorer frequency resolution. Because the window duration sets the frequency and time resolution STFT is said to have *fixed resolution*. Figure 3.4.32 illustrates the tradeoff between time and frequency resolution as the window durations increase. The top diagram depicts the resulting time and frequency resolution from having short windows. The time resolution is good, but the frequency resolution is poor. In contrast, when the windows are long the time resolution is poor, but the frequency resolution is good.

In Figure 3.4.33 we applied STFTs with varying window sizes to a log sweep from 1 Hz to 10 Hz over 100 seconds (2^{14} samples) so we can see the effects of fixed resolution. Hamming window lengths were varied at .39, .78, 1.56, 3.12, and 6.25 seconds (2^6, 2^7, 2^8, 2^9, and 2^{10} samples

Figure 3.4.31 *Short time Fourier transformation. By using a moving window the time interval exposed to transformation can be limited.*

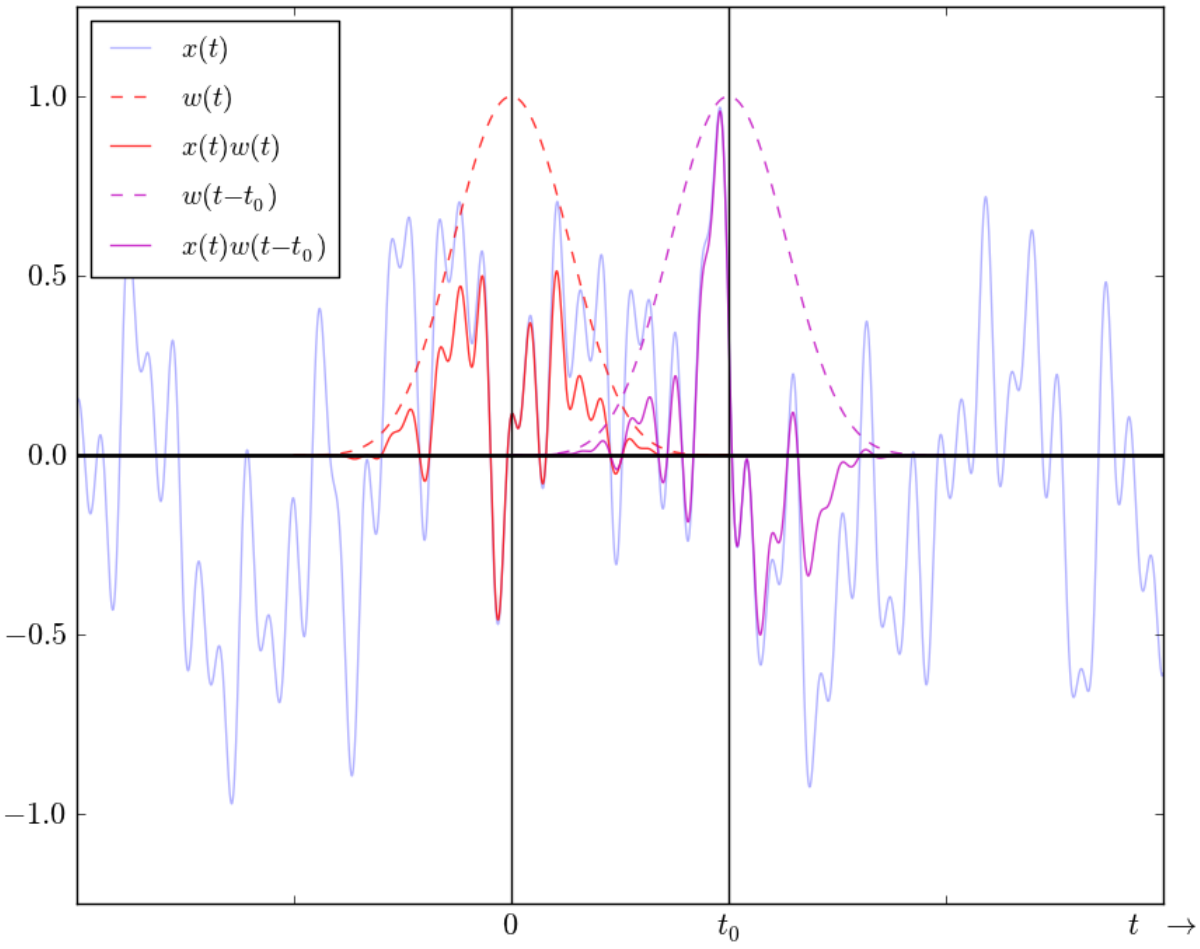


Figure 3.4.32 *Trade-off between time and frequency resolution. Short time Fourier transforms have fixed time and frequency resolution dependent on the duration of the window. Short windows have good time resolution but poor frequency resolution. Long windows have poor time resolution but good frequency resolution.*

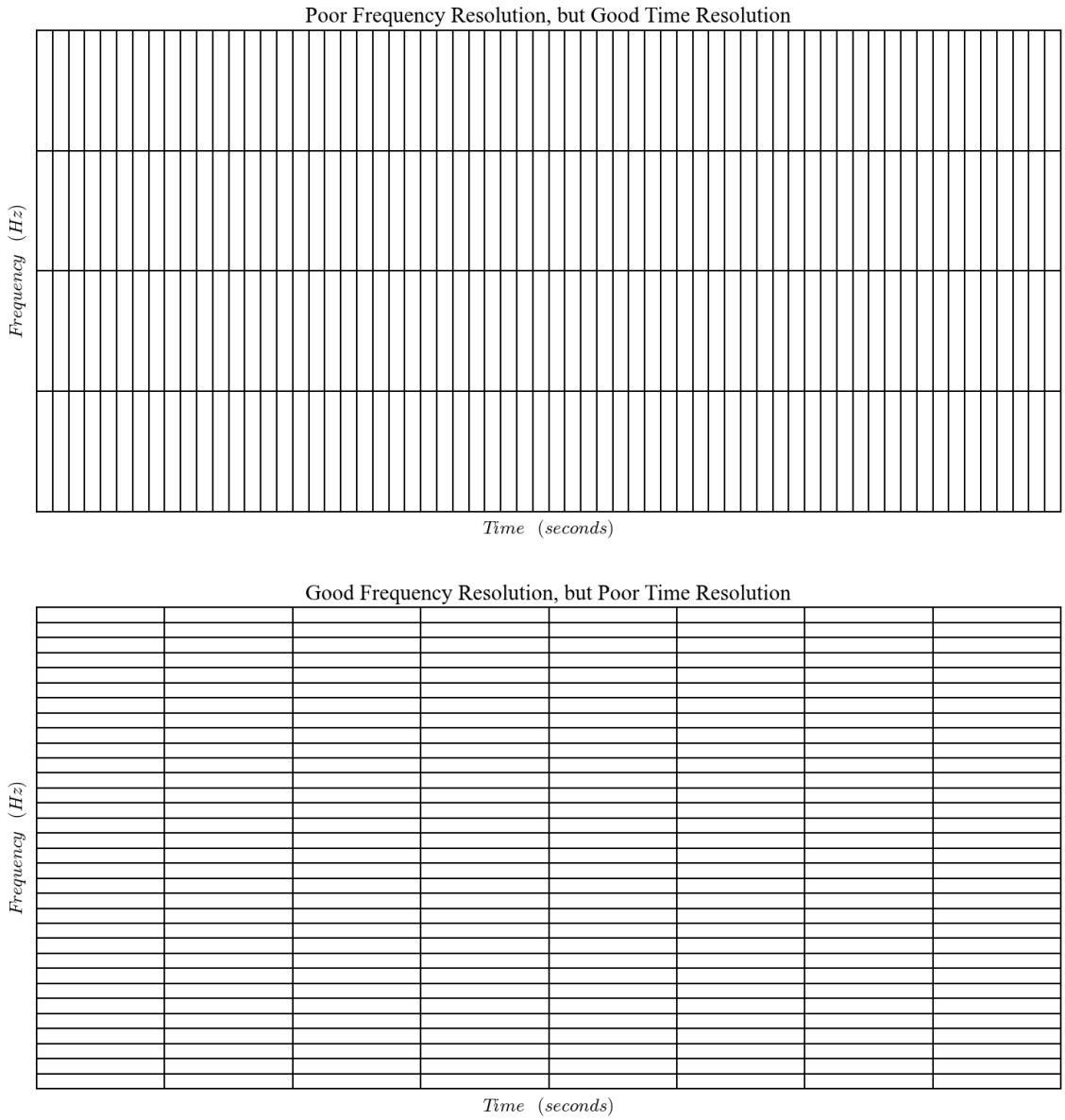
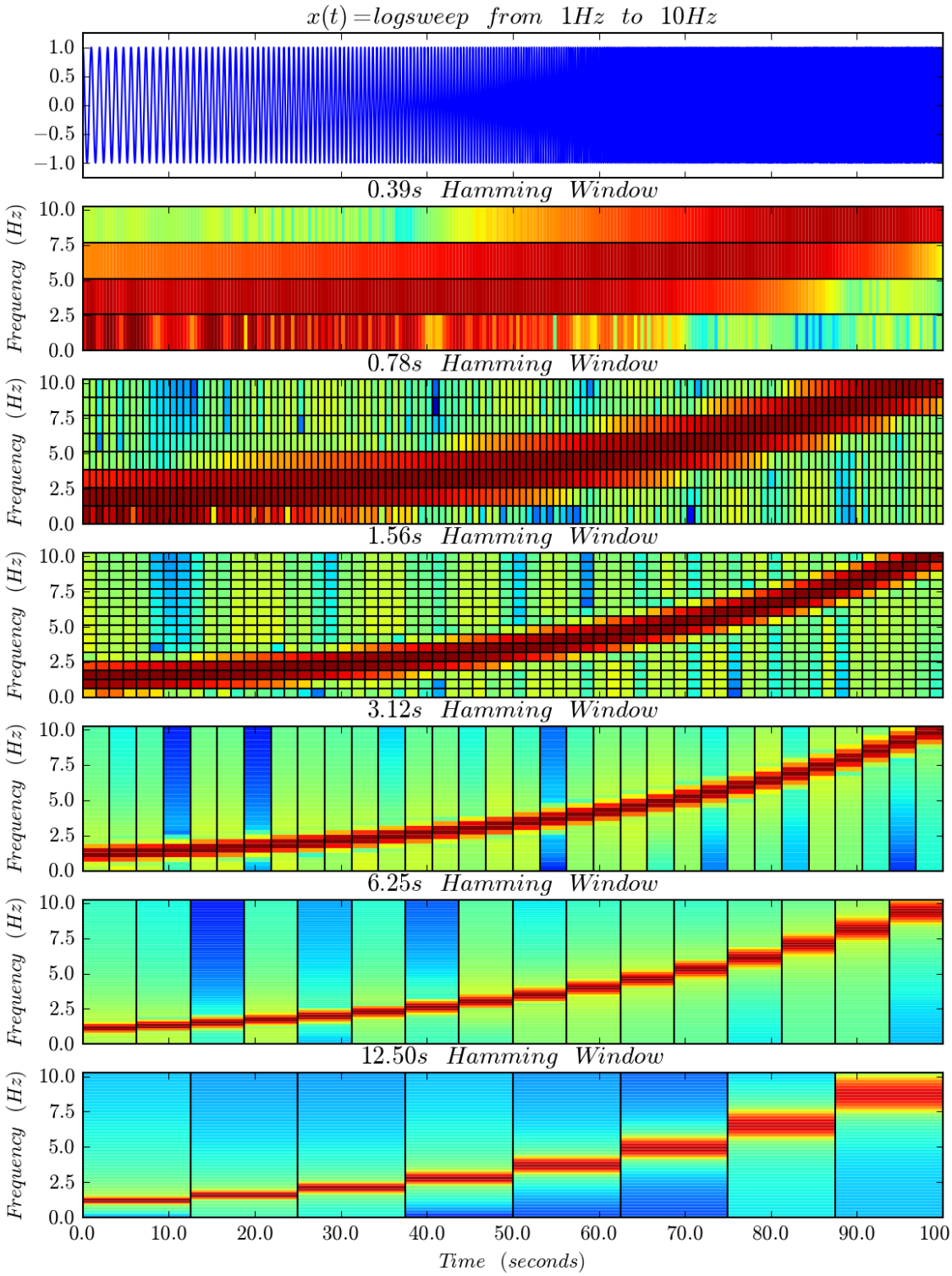


Figure 3.4.33 *STFTs of a logarithmic chirp at logarithmically increasing window sizes.*



respectively). In the color plots time is on the x-axis and frequency is on the y-axis and color indicates the decibel magnitude in time and frequency. Red values indicate higher magnitudes while blue indicates lower magnitudes. The In each subplot the grid or lattice structure is indicative of fixed resolution windowed Fourier transform techniques. [Take note that not all the grid lines between time and frequency bins are plotted in some of the subplots. When the bins become small the grid lines make them difficult to interpret. The color gradients should be adequate to infer the presence of multiple bins.] With the shortest window of 0.39 seconds we can see that the spectral leakage obscures the frequency information. In contrast the longest window of 12.5 seconds obscures changes in the time domain.

3.4.4.2 The Uncertainty Principle. The tradeoff between frequency and time resolution reflects Heisenberg's uncertainty principle. The uncertainty principle is usually discussed in the context of quantum mechanics and expresses the inability to simultaneously determine pairs of physical properties with arbitrarily high precision. The classic example is resolving a particle's position and its momentum. The more certainty we have in regard to one of those entities the less certainty we have in the other. In the above definition the term *inability* was used but the uncertainty is not related to the technical capabilities of the measurement devices; the uncertainty is related to the system under observation. Applied to spectral analysis the theorem states (Folland & Sitaram, 1997):

A nonzero function and its Fourier transform cannot both be sharply localized.

In Figure 3.4.33 we can see this uncertainty at work as precise time resolution yields poor time resolution and vice-versa. The medium window sizes of 1.56 and 3.12 give the visually appearance of being a quality compromise but they could be better. From our experience with spectral leakage we know the *raw window size* isn't as important as the *window size relative to the frequencies under examination*. The inherent shortcoming of STFT is that it doesn't take this into account. A more optimized approach to this problem is to scale the size of the windows to match the frequencies

under examination. This would be called a *multi-resolution* analysis. The uncertainty principle will never let us have arbitrarily good time and good frequency resolution, but Wavelets will at least provide a more optimized means of attaining both time and frequency resolution..

In the previous section we presented the short comings of fixed resolution analysis. A clunky approach to multi-resolution analysis would be to perform multiple STFTs with various window sizes and take the high frequency information across time from the 0.39 window, the medium frequency information across time from the medium windows, and so on.
The result of such an effort is depicted in Figure 3.4.34. In this figure we also included window sizes of 25 seconds and 50 seconds. Here we see a very different lattice structure. The low frequencies have long time windows with many frequency divisions, and as we move up in frequency the time windows become shorter and the frequency bins become wider. In this example the scaling in both the time and frequency domains is dyadic (based on doubling/ halving). Figure 3.4.35 depicts the dyadic time frequency lattice structure. In this figure we can see how the time windows halve as we move up in frequency and how the frequency windows halve as we move down in frequency. This in essence is multi-resolution wavelet transformations.

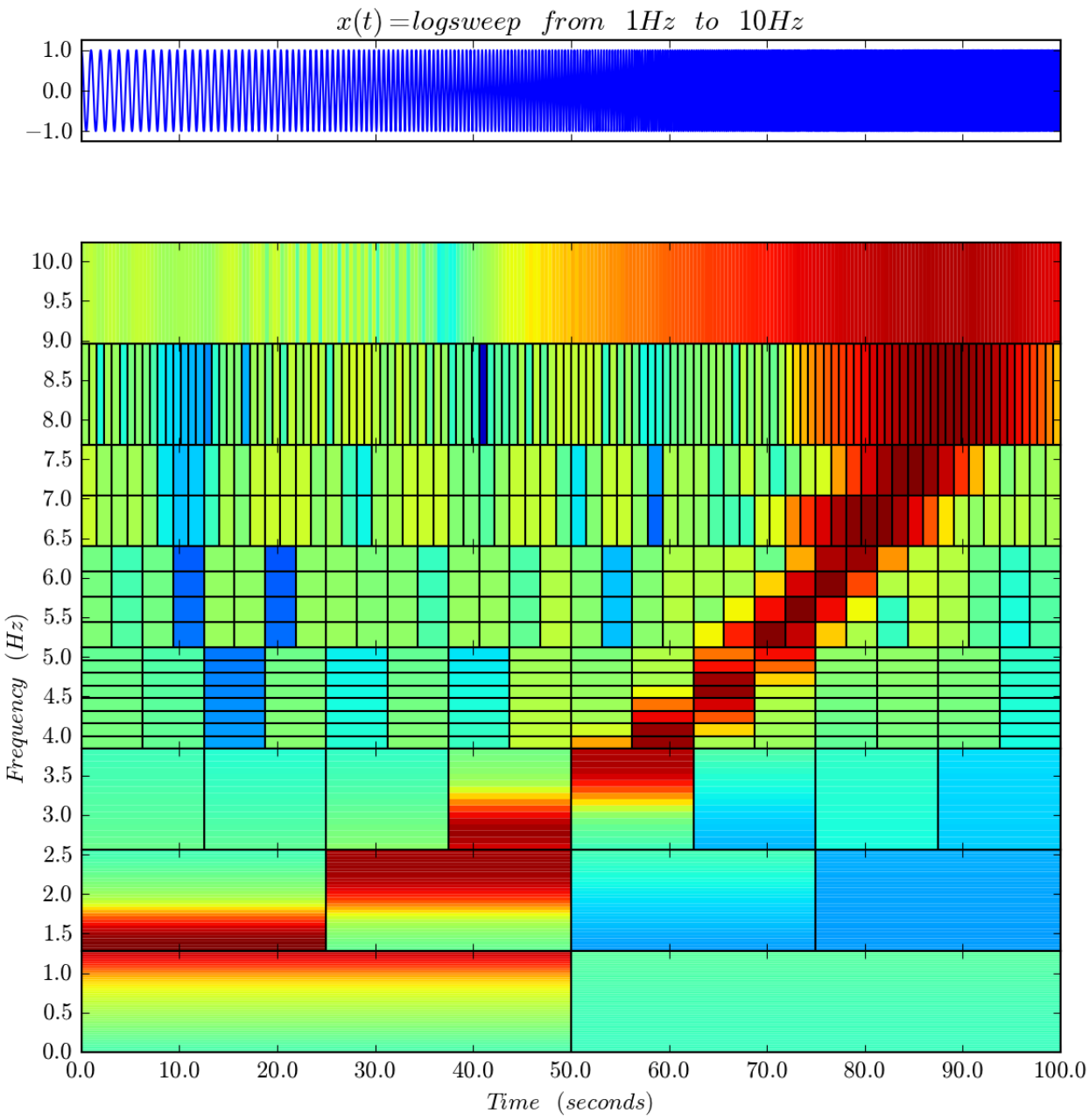
3.4.4.1 STFT and Cross Correlation. We can get a better idea of how this works by tweaking the continuous short-time Fourier transform,

$$X(\tau, \omega) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}\,\mathrm{d}t.$$

With our current "mental model" of the continuous short time Fourier transform the window function constrains $x(t)$, and only the constrained portion is multiplied and integrated with respect to time. Now instead of thinking of the window function constraining $x(t)$, think of the window constraining $e^{-j\omega t}$. We can define this function as

$$g_{\tau,\omega}(t) \stackrel{\text{def}}{=} w(t-\tau)e^{j\omega t}.$$

Figure 3.4.34 *Multi resolution analysis with dyadic time frequency resolution.*


$x(t) = logsweep \quad from \quad 1Hz \quad to \quad 10Hz$
1.0
0.0
−1.0
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Frequency (Hz)
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
100.0
Time (seconds)

This function is called an envelope function because it contains the complex sinusoidal components of $e^{j\omega t}$ within the envelope of the $w(t-\tau)$. Figure 3.4.36 depicts $g_{\tau,\omega}(t)$ with a fixed window width of 1 second and ω of 5, 10, and 20 Hz. These strange looking functions act like filters when they are shifted and multiplied across time varying functions. The filters will let frequencies close to ω pass through will excluding other frequency components. With our newly defined function we can express the continuous short-time Fourier transform as

$$X(\tau,\omega) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} x(t)\overline{g_{\tau,\omega}(t)}\,\mathrm{d}t.$$

In this context we can view the continuous STFT as the *cross-correlation* between $g_{\tau,\omega}(t)$ and $x(t)$. Because some reader's may not be familiar with cross correlation we will briefly review the concept before elaborating on this idea.

The cross-correlation is also known as the sliding dot-product or sliding inner product. The operation reflects the amount of similiarity between two functions as one is slide across the other. For two functions $f(\tau)$ and $h(\tau)$ the cross-correlation operation is defined as

$$(f\star h)(\tau) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \overline{f(t)}h(\tau+t)\,\mathrm{d}t.$$

It can also be expressed in terms of convolution as

$$(f\star h)(\tau) \stackrel{\text{def}}{=} \overline{f(-\tau)} * h(\tau)$$

From the definitions we can see the distinction between cross-correlation and convolution is that in cross-convolution none of the functions are mirrored, instead the conjugate is taken of the first argument. The conjugation is needed for the same reason we discussed in regard to the Hermitian inner product. When either function $f(\tau)$ and $h(\tau)$are Hermitian,

$$(f\star h)(\tau) = (f * h)(\tau).$$

In regard to continuous STFT the convolution and cross correlation between $x(t)$ and $g_{\tau,\omega}(t)$ yields identical results because the window functions have even symmetry about zero, the

resulting $g_{\tau,\omega}$ functions will be Hermitian because the window envelope are filled by $e^{j\omega t}$ which is Hermitian (the real part is a cosine wave, the imaginary part is a sine wave). We can visualize this graphically in the $g_{\tau,\omega}$ functions in Figure 3.4.36. Taking the complex conjugate inverts leaves the real part alone and inverts the imaginary part. Reversing $g_{\tau,\omega}$ results in the same result as complex conjugation.

To transition back to our task at hand Figure 3.4.37 depicts the cross correlation between a linear sweep between 1 and 10 *Hz* over 10 seconds and a Gaussian envelope with a duration of 1.25 seconds filled with a complex sinusoid at 3 *Hz*. The peak of the cross correlation corresponds to the point in time that has the most amount of overlap similarity between the sweep and enveloped sinusoid. Keep in mind we are still under the umbrella of continuous STFT. We still have the same limitations associated with fixed window sizes. We are just taking a different approach to visualizing what the continuous STFT is doing.

Figure 3.4.35 *Dyadic grid used for multi-resolution analysis.*

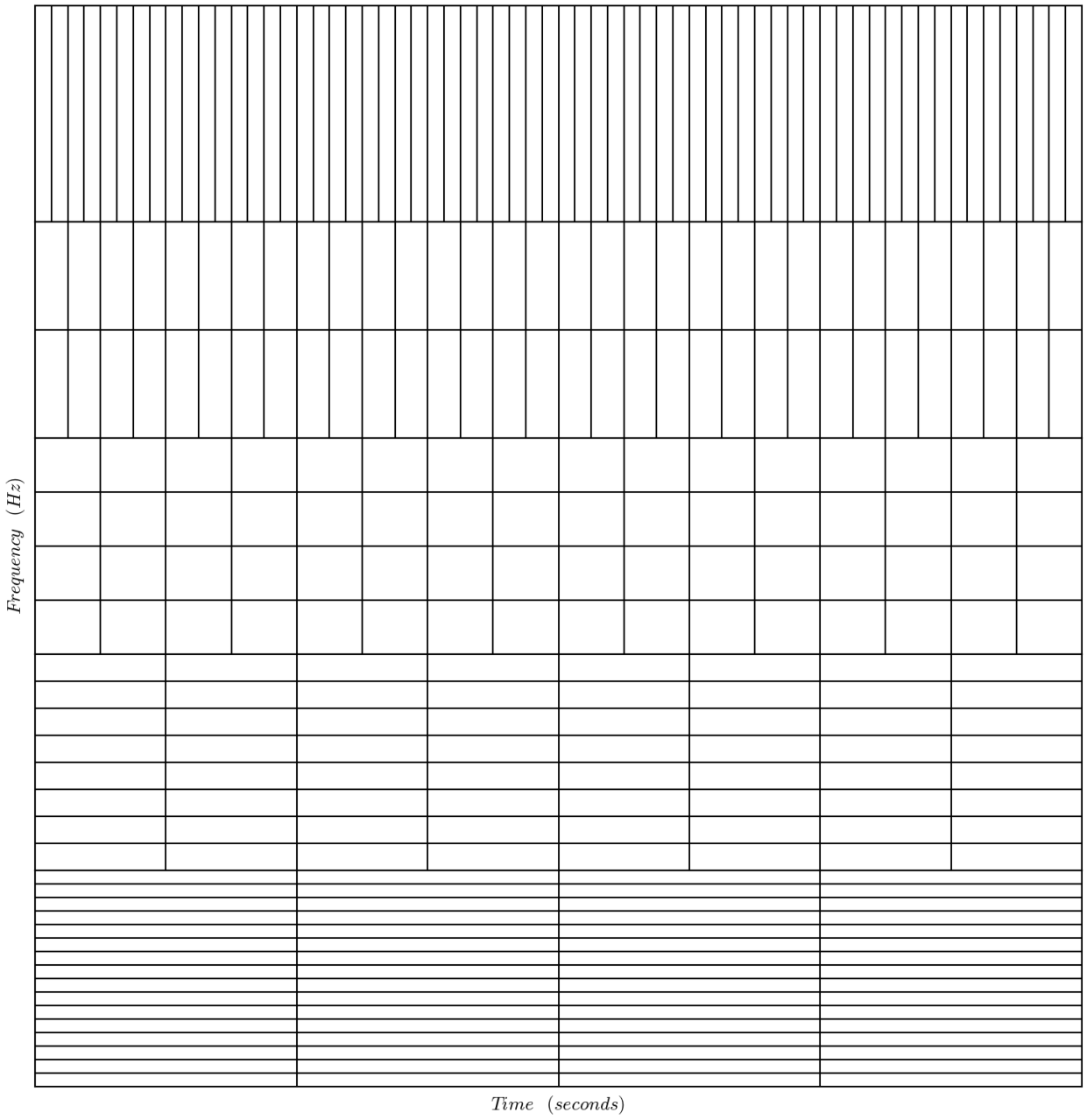


Figure 3.4.36 *Envelope functions of a fixed width filled with complex sinusoids.Sinusoids are at frequencies of 5, 10, and 20 Hz.*

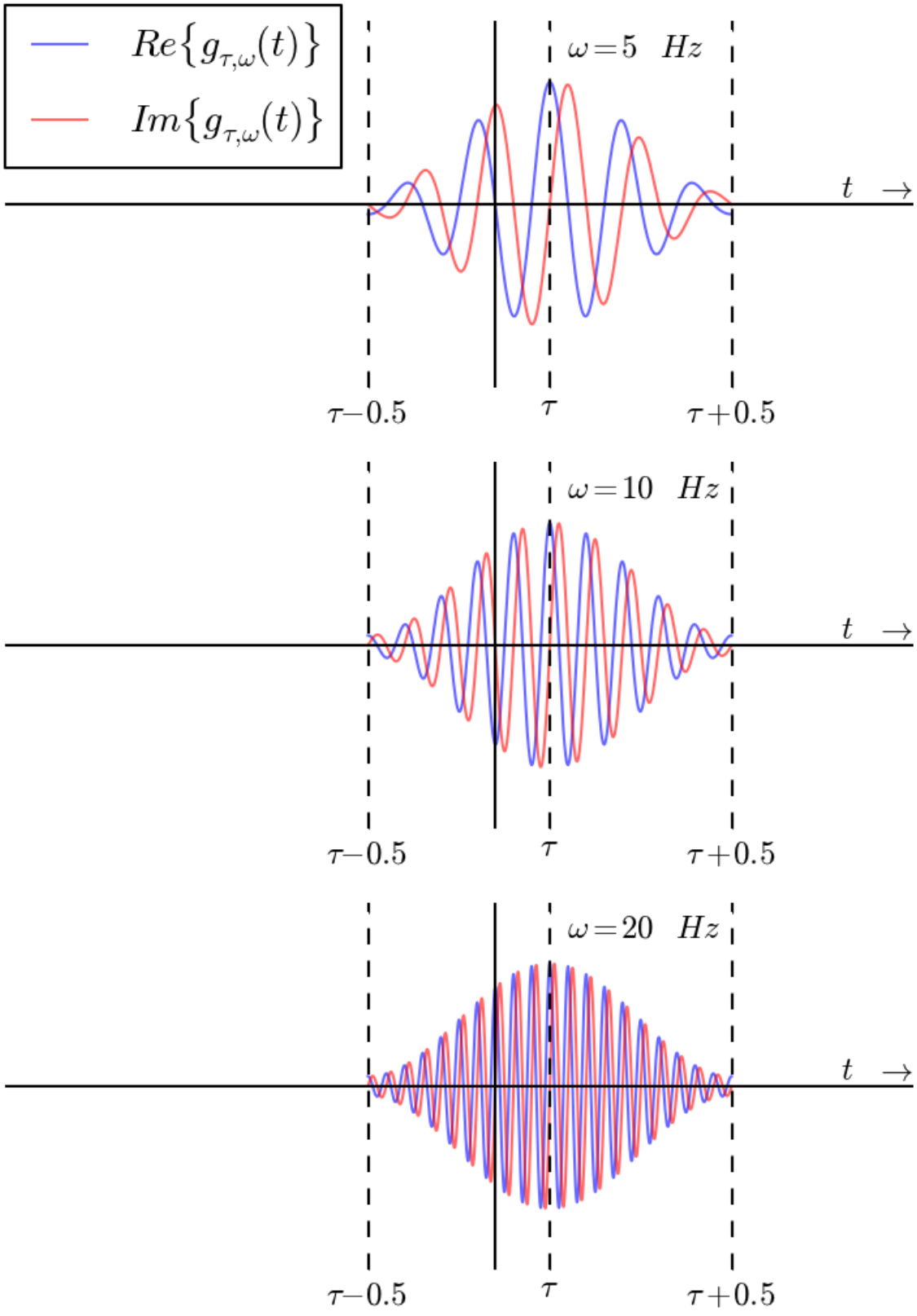
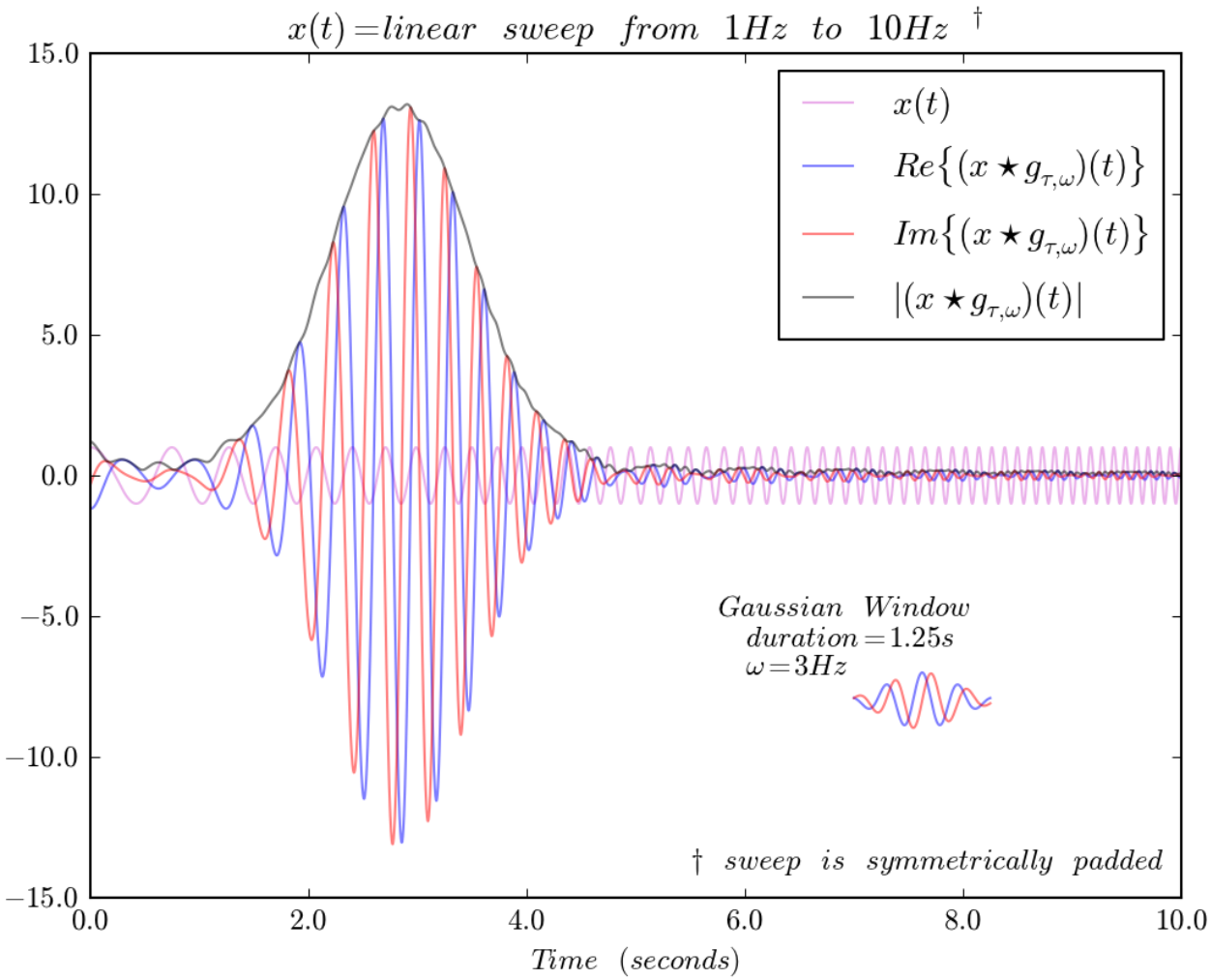


Figure 3.4.37 *Cross correlation of a windowed complex sinusoid with a linear sweep.*



3.5 Wavelet Analysis

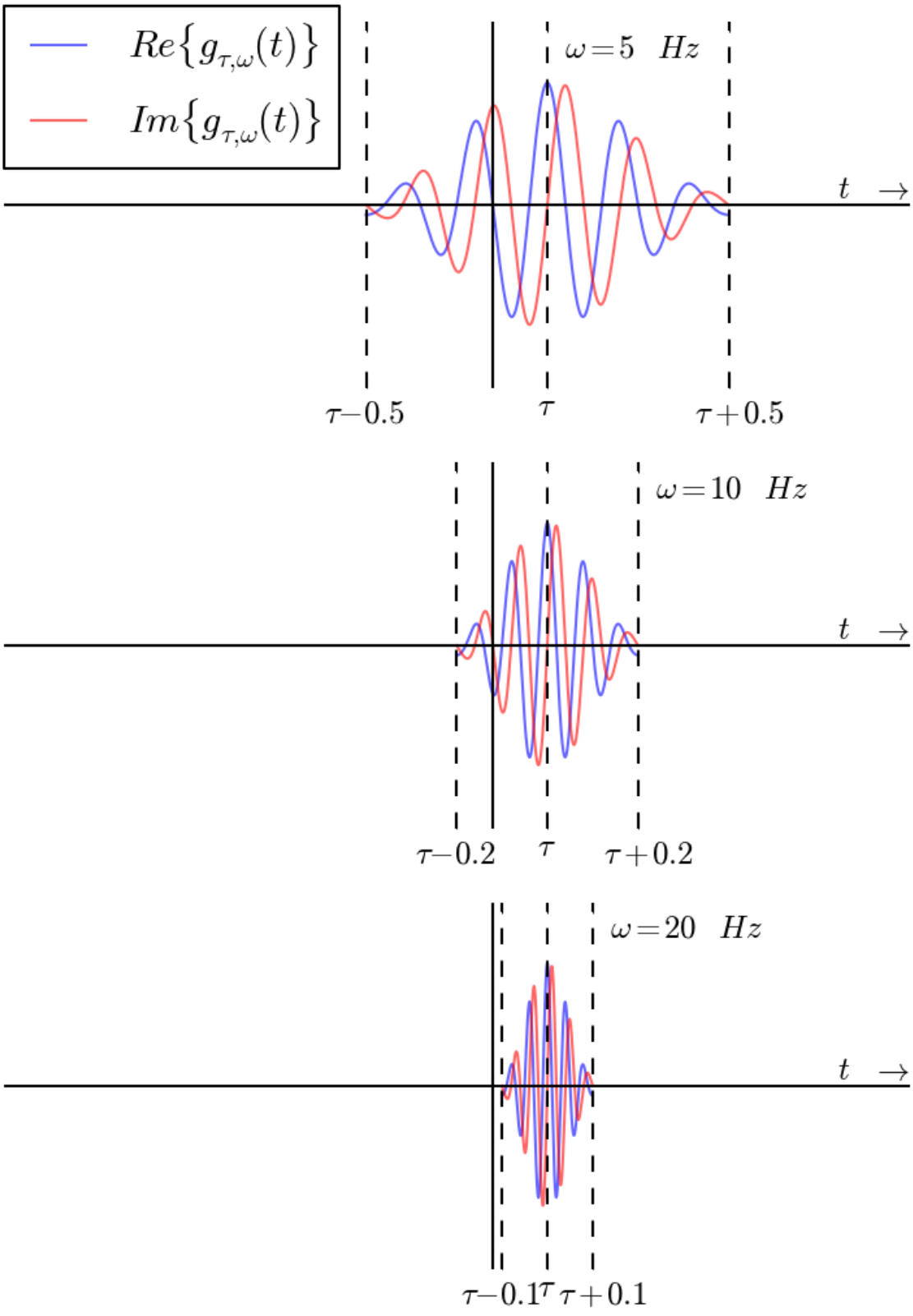
In the previous section it was demonstrated that STFT can be performed by windowing complex sinusoids and performing cross correlations. Here we show how wavelet analysis is a generalization of STFT.

3.5.1 ***What is a wavelet?*** The Gaussian enveloped complex sinusoids we have been examining are actually a special class of wavelets termed Morlet wavelets. Our goal with wavelet analysis is to optimize the size of the window relative to the frequencies under examination. The key insight of wavelet analysis is recognizing that this goal can be achieved by scaling the wavelet to set the pass band of the filter. The enveloped complex sinusoids in Figure 3.5.1 filter the same frequencies as their counterparts in Figure 3.4.36 but the scaling provides better time resolution at 10 and 20 Hz. Instead of setting the duration of the window and manipulating ω we manipulate both the duration and ω by scaling the wavelet. Controlling the window and ω insures that the window size relative to the frequencies under examination stays fixed.

Morlet wavelets provide a natural transition from discussing Fourier transformations but there are in fact several dozen different types of wavelets and a literal infinite number waiting to be discovered. The Morlet wavelet is actually somewhat atypical in that it is a complex wavelet. Most of the wavelets we will encounter are real. Wavelet analysis has a variety of applications beyond spectral analysis from image compression to transmission of data over a bandlimited channel. Some wavelets are optimized for time or frequency resolution or dynamic range, while some are optimized for computational efficiency, while others are optimized for particular applications (Daubechies, 1992).

A wavelet is a filter with some special characteristics. Wavelets are designated as $\psi(t)$. They are considered to be wavelets if they satisfy the admissibility criterion.

Figure 3.5.1 *Envelope functions of varied width filled with complex sinusoids. Sinusoids are at frequencies of 5, 10, and 20 Hz.*



The admissibility criterion is given by,

$$\int \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < +\infty$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(t)$

For the admissibility condition to hold $|\Psi(\omega)|^2$ must decay faster then $1/\omega$. This ensures the wavelet does not have infinite bandwidth. Secondly, for the admissibility condition to hold $|\Psi(\omega)|^2\Big|_{\omega=0} = 0$. In the time domain this essentially means the average value of the wavelet is zero or that the positive area of the wavelet equals the negative area of the wavelet or $\int \psi(t)dt = 0$. This ensures that the wavelet's bandwidth does not extend to zero, or that it is unbiased (no DC component, AC coupled).

The admissibility condition ensures that the wavelet acts as a band-pass filter when it is convolved with a time-varying single. The frequencies which are allowed through the filter are referred to as the as the pass band, frequency band, or subspace. Unlike the Fourier Transform the basis functions can be non-sinusoidal. Non-sinusoidal kernels are optimized to provide better time/frequency localization and reduce redundancy. In spectral analysis the advantages of sinusoidal basis functions are mostly analytic rather than practical. As previously demonstrated the analysis of sines and cosines is often aided by the fact that they can be represented and manipulated as complex exponentials through Euler's formula. From a practical standpoint any wavelet kernel (whether sinusoidal or not, and by definition satisfying the admissibility criterion) can represent any signal with arbitrarily small precision. This is accomplished by scaling and translating the kernel to obtain an infinite number of subspaces or daughter wavelets:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right)$$

where:

s is the scale factor or dilation factor (replaces the ω)

τ is the translation factor .

Without the normalization factor $\frac{1}{\sqrt{s}}$ low frequency wavelets would become amplified. The normalization factor scales the amplitudes to reflect the power in the input signal.

As you have probably noticed by now transformations usually come in continuous and discrete varieties. Wavelet analysis is no different. We will begin by discussing the continuous transformation and move on the discrete transformations. The discrete transforms come in two varieties: one is redundant (similar to the DTFT), while the second is non-redundant (like the DFT).

3.5.2 ***The continuous wavelet transform (CWT).*** The CWT transforms a single valued function of continuous time into a function of both continuous frequency and continuous time given by the vector space of $\psi_{s,\tau}(t)$. The continuous wavelet transform is defined as:

$$\gamma(s,\tau) = \int x(t)\overline{\psi_{s,\tau}}(t)dt$$

The transform is commonly represented in Hilbert space shorthand by $\langle x, \psi_{s,\tau}\rangle$. As with the continous Fourier transform the CWT yields a lot of redundant information, but can be a powerful tool for time-frequency analysis. In Figure 3.5.2 magnitude estimates obtained from CWT with Morlet wavelets of our familiar logsweep between 1 to 10Hz. We can see from the figure that using scaling wavelets much better time-frequency compared to windowed Fourier transformations. In the digital age the CWT is of limited utility. With digital signals we would like to be able to calculate wavelet transformations more efficiently or to obtain coefficients that are orthogonal to one another. Discrete wavelet transformations fill these needs. They come in both redundant and non-redundant varieties.

3.5.3 ***The redundant discrete wavelet transform (DWT).*** With the discrete transformation the dilation and translation parameters only take discrete values. When the dilation is small the wavelet is scaled by small steps to cover high frequencies, and when the dilation is larger the translation increases to cover low frequencies. The wavelets are scaled according to,

$$\psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}}\psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j}\right).$$

A second disctinction worth mentioning is that with discrete wavelets the kernels are almost always real-valued. The tails of the Gaussian envelopes of complex-valued Morlet wavelets take a while to converge to zero. The consequence of this is that signals need to be fairly long relative to the windows to be reconstructable.

3.5.4 ***The non-redundant discrete wavelet transform (DWT).*** By specially choosing the ψ , s_0^j and τ_0 parameters it is possible to form an orthonomal basis for $L^2\mathbb{R}$. To state this more simply, this means that any function with finite energy can be represented with arbitrarily good precision and zero redundancy. To see how this is accomplished we first apply a dyadic sampling scheme by scaling a mother wavelet by $2^j$and translating it by $k2^j$:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t - k2^j}{2^j}\right)$$

The dyadic sampling is optimal because the Nyquist rate always cooresponds with the time variable for every frequency. Non-dyadic sampling schemes can also be used but dyadic sampling is by far the most common. Graphically this is depicted in Figure 3.5.3. Note that the translation factor scales with the dilation. In essence, the dyadic sampling assures that the time sampling is non-redundant. Next we turn our attention to the choosing a wavelet. The wavelet must be orthogonal to itself when it is scaled and translated. This condition can be stated as:

$$\langle \psi_{j,k}, \psi_{m,n} \rangle = \begin{cases} 1 & \text{if } j = m \text{ and } k = n \\ 0 & \text{otherwise} \end{cases}$$

Secondly we need the frame of wavelets to be a tight frame,

$$A\|f\|^2 \leq \sum_{j,k} \left| \langle x, \psi_{j,k} \rangle \right|^2 \leq B\|f\|^2 \text{ where } 0 < A \leq B < \infty.$$

Recall a tight frame is when the lower and upper frame bounds are equal to one another. A tight frame behaves as if it were orthonormal. If the frame is also a Parseval frame ($A = B = 1$) then orthonormality is satisfied.

To demonstrate how any arbitrary continuous function can be approximated by a discrete wavelet let's take a look at one of the simplest wavelets known as the Haar sequence or Haar wavelet. The sequence was discovered in 1909 before a formal definition of wavelets existed. The Haar wavelet is a very simple piecewise function,

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2, \\ -1 & 1/2 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3.5.4 depicts the Haar mother wavelet and an orthogonal scaled and translated daughter wavelet. From the figure the independence of the mother and daughter wavelets can be visually confirmed by observing how the daughter wavelet only varies only where the mother wavelet is fixed. Figure 5.5.4 depicts an arbitrary function $f(x)$ and the Haar wavelet approximations of increasing precision. The Figure only depicts 7 levels of precision, but is hopefully enough to at least give the basic idea of how the decomposition can represent any function with arbitrarily good precision if enough levels are used in the approximation.

Figure 3.5.2 *Continuous Wavelet Transform of a logsweep from 1 to 10 Hz.*

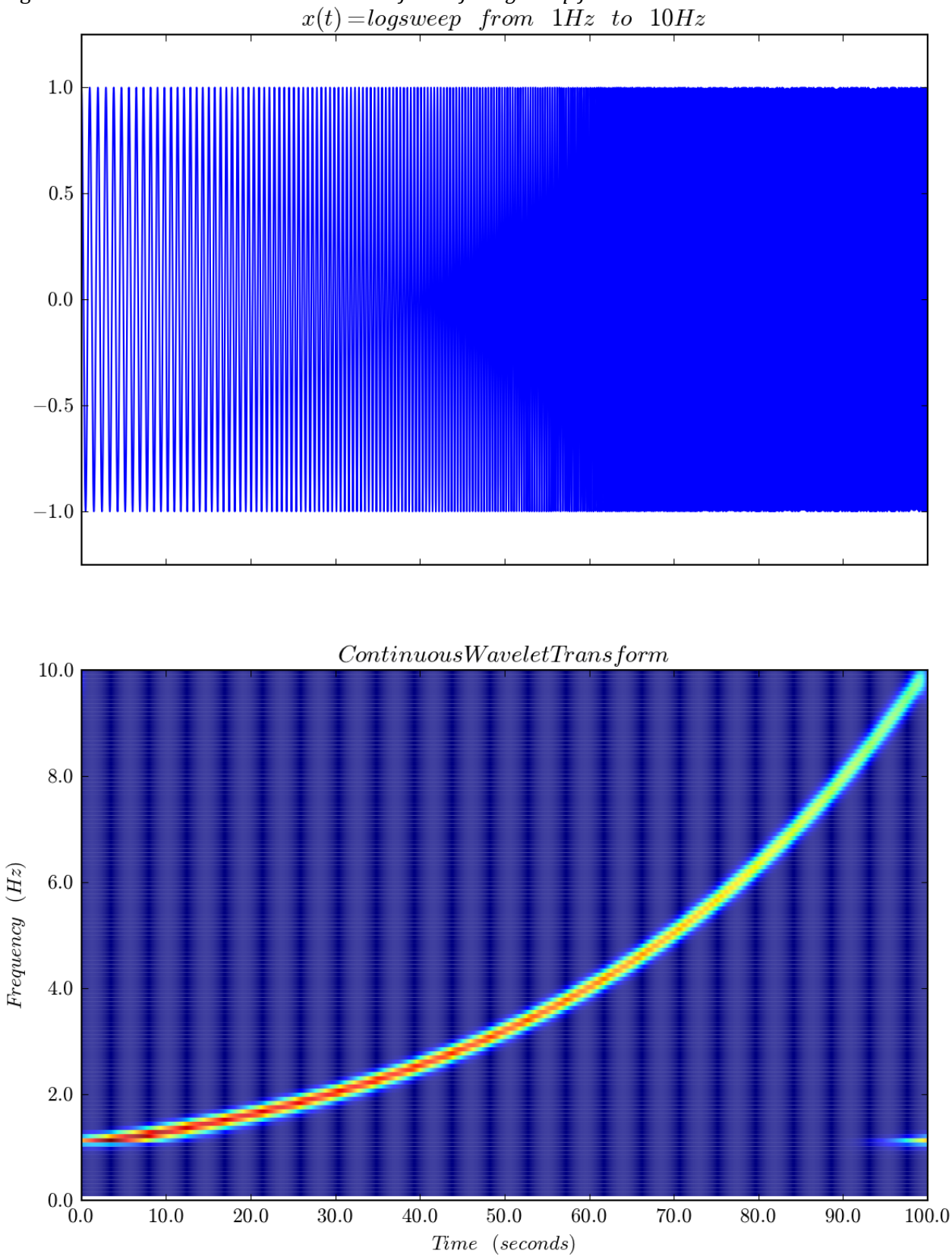


Figure 3.5.3 *Cascade filter bank. Wavelet decomposition can be conceptualized and implemented as a cascaded filter bank yielding orthogonally bandpassed coefficients.*

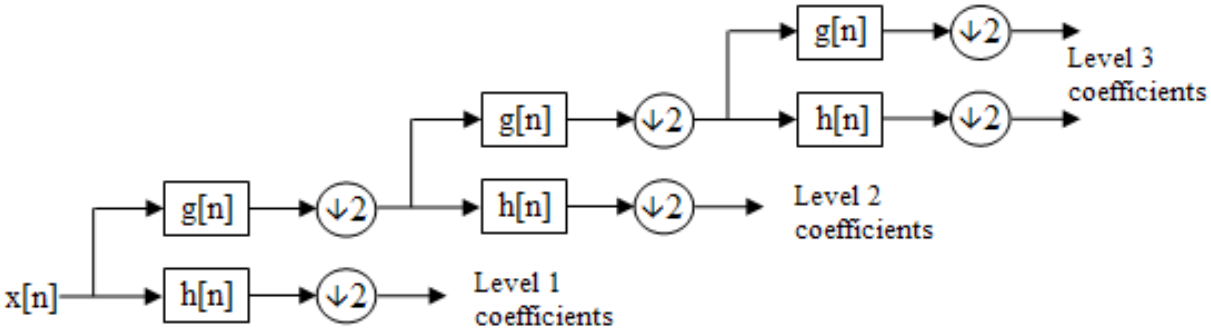


Figure 3.5.4 *Haar Wavelet.Top panel depicts a mother Haar Wavelet. Even this simple piecewise function is capable of representing any continuous function with arbitrary precision. The bottom panel depicts a scaled and translated daughter wavelet that is orthogonal to the mother.*

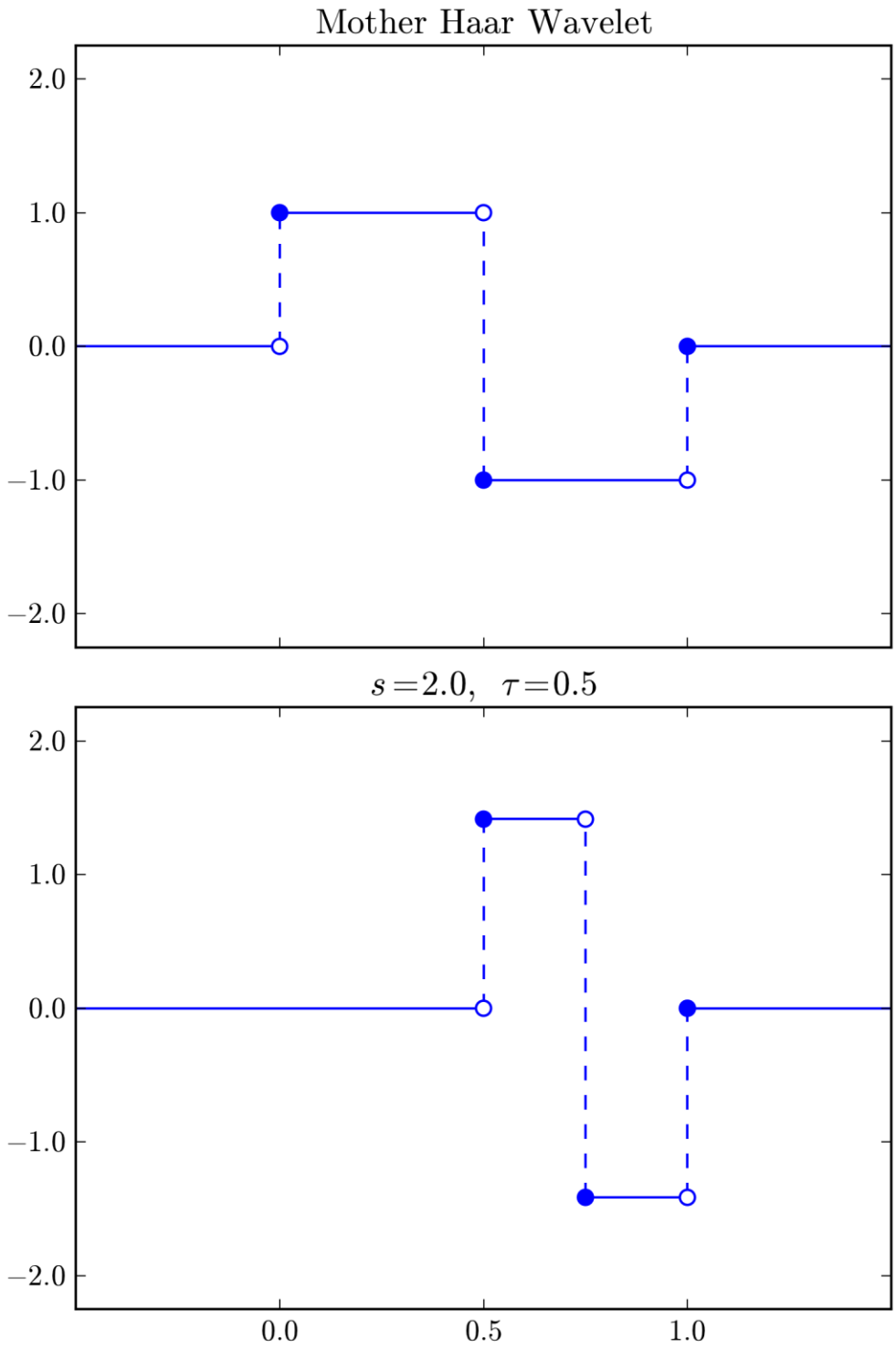
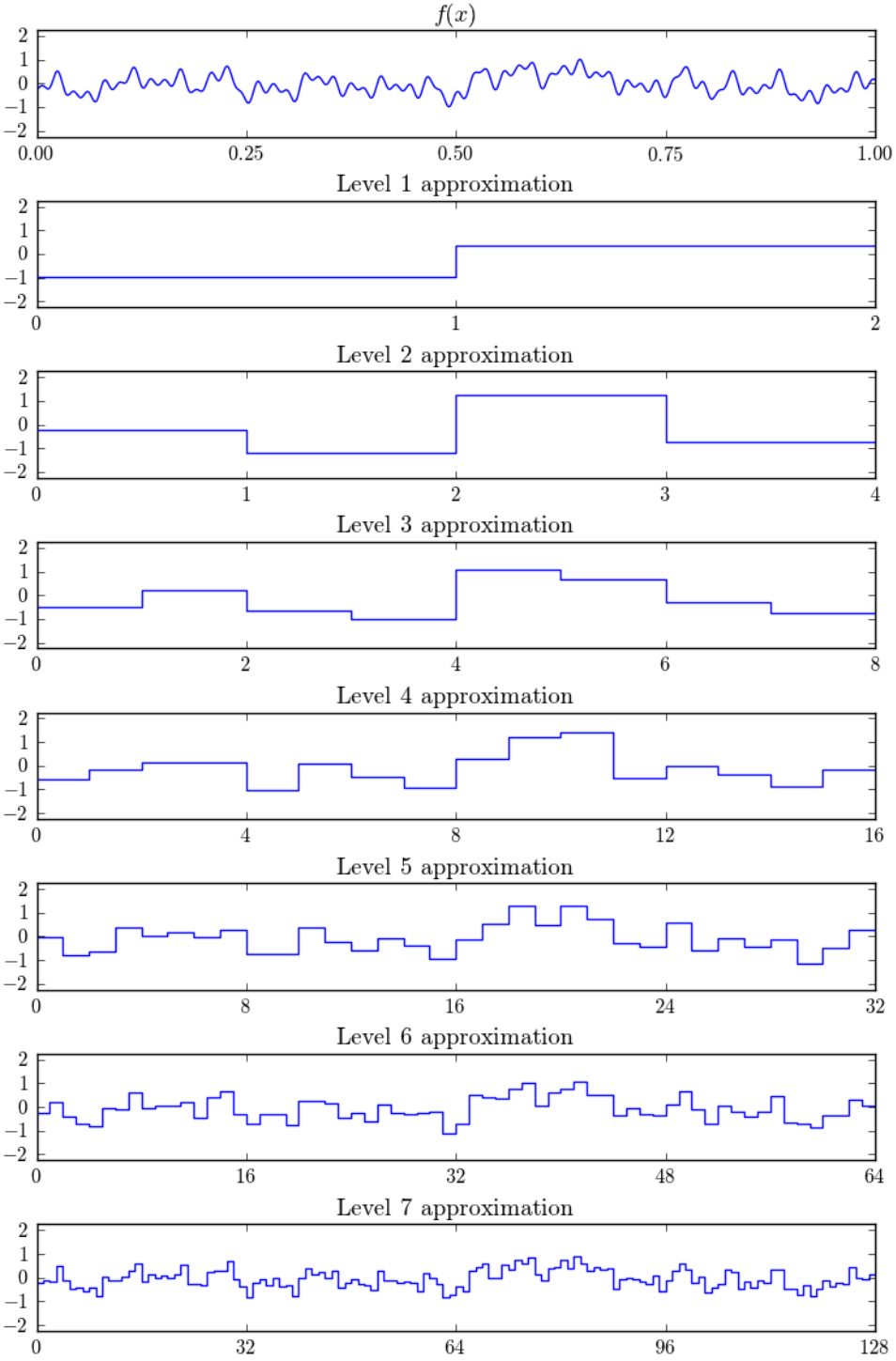


Figure 3.5.5 *Discrete redundant wavelet decomposition with Haar wavelets.*



Appendix 3.A Source Code for Chapter 3 plots

3.A.1 Fourier series approximations of a square wave (Figure 2.3.1)

```
"""
Plots Fourier Series approximations of a square wave.

Square Wave Function
f(x) = sgn(sin(x))

Fourier Series Coefficients
a0 = .5
an = 0
bn = (2./(pi*(n*2.-1.)))*sin((n*2.-1.)*x)
"""

import pylab
from numpy import array,pi,linspace,mod,cos,sin,ones

def if_else(a,c,d):
    if a: return c
    else: return d

def sgn(x):
    return [if_else(k<0.,0.,1.) for k in x]

def square(x):
    return sgn(sin(x))

N=16 # Specify the number of approximations
x=linspace(-3*pi,3*pi,512)
pylab.figure(figsize=(9,12))

for n in xrange(N):
    if n==0 : y=ones(512)*.5
    else: y+=(2./(pi*(n*2.-1.)))*sin((n*2.-1.)*x)
    pylab.subplot(N,1,n+1)
    pylab.plot(x,y)
    pylab.plot(x,square(x),'k--')
    pylab.text(-3*pi,.8,r'$n=%i$'%(if_else((n*2-1)<0,0,(n*2-1))))
    pylab.ylim([-.5,1.5])
    pylab.yticks([-0.,.5,1.],(r'$0.0$',r'$0.5$',r'$1.0$'))
    pylab.xticks([])

pylab.xticks(linspace(-3*pi,3*pi,7),(r'$-3\pi$',r'$-2\pi$',
                                     r'$-1\pi$', r'$0\pi$',
                                     r'$1\pi$',r'$2\pi$',
                                     r'$3\pi$'))

pylab.savefig('square_fourier_series.png',dpi=300)
```

3.A.2 Phasor representation of 2cos(2πt) (Figure 2.3.1.1)

```
"""3d plot of the phasor representation of cos(2pi t)"""

import matplotlib as mpl
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import matplotlib.pyplot as plt

from numpy import pi

fig = plt.figure(figsize=(12,9))
ax = fig.gca(projection='3d')

t = np.linspace(0., 2., 100)
re1 = np.cos(t*2.*pi)
im1 = np.sin(t*2.*pi)

re2 = np.cos(t*-2.*pi)
im2 = np.sin(t*-2.*pi)

# unit circle
ax.plot(np.ones(50)*0.,re1[:50],im1[:50], c=(0.8, 0.8, 0.8))

# real plane shadows
ax.plot(t, re1, np.ones(100)*-2.,'b', alpha=0.45)
ax.plot(t, re2, np.ones(100)*-2.,'g', alpha=0.45)
ax.plot(t, re1+re2, np.ones(100)*-2.,'k', alpha=0.6,
        label=r'$Re\{e^{2\pi t} + e^{-j2\pi t}\}$')

# imaginary plane shadows
ax.plot(t, np.ones(100)*2., im1,'b--', alpha=0.45)
ax.plot(t, np.ones(100)*2., im2,'g--', alpha=0.45)
ax.plot(t, np.ones(100)*2., im1+im2, 'k--', alpha=0.6,
        label=r'$Im\{e^{j2\pi t} + e^{-j2\pi t}\}$')

# phasors
ax.plot(t, re1, im1,'b', linewidth=2, label=r'$e^{j2\pi t}$')
ax.plot(t, re2, im2,'g', linewidth=2, label=r'$e^{-j2\pi t}$')
##ax.plot(t, re1+re2, im1+im2, c=(0.3, 0.3, 0.3), linewidth=2, label=r'$2 \cos(2\pi t)$')
ax.legend()

ax.set_ylim3d([-2.0,2.0])
ax.set_zlim3d([-2.0,2.0])

ax.set_title(r'$2 \cos(2\pi t) =  e^{j2\pi t} + e^{-j2\pi t}$')
ax.set_xlabel('Time (s)')
ax.set_ylabel('Real Axis')
ax.set_zlabel('Imaginary Axis (j)')

fig.savefig('cos(2pit).png',dpi=300)
plt.close()
```

3.A.3 Phasor representation of 2sin(2πt) (Figure 2.3.1.3)

```
"""3d plot of the phasor representation of sin(2pi t)"""

import matplotlib as mpl
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import matplotlib.pyplot as plt

from numpy import pi

fig = plt.figure(figsize=(12,9))
ax = fig.gca(projection='3d')

t = np.linspace(0., 2., 100)
re1 = np.cos(t*2.*pi)
im1 = np.sin(t*2.*pi)

re2 = np.cos(t*-2.*pi)
im2 = np.sin(t*-2.*pi)

# unit circle
ax.plot(np.ones(50)*0.,re1[:50],im1[:50], c=(0.8, 0.8, 0.8))

# real plane shadows
ax.plot(t, re1, np.ones(100)*-2.,'b', alpha=0.45)
ax.plot(t, re2, np.ones(100)*-2.,'g', alpha=0.45)
ax.plot(t, re1-re2, np.ones(100)*-2.,'k', alpha=0.6,
        label=r'$Re\{je^{j2\pi t} - je^{-j2\pi t}\}$')

# imaginary plane shadows
ax.plot(t, np.ones(100)*2., im1,'b--', alpha=0.45)
ax.plot(t, np.ones(100)*2., im2,'g--', alpha=0.45)
ax.plot(t, np.ones(100)*2., im1-im2, 'k--', alpha=0.6,
        label=r'$Im\{je^{j2\pi t} - je^{-j2\pi t}\}$')

# phasors
ax.plot(t, re1, im1,'b', linewidth=2, label=r'$je^{j2\pi t}$')
ax.plot(t, re2, im2,'g', linewidth=2, label=r'$je^{-j2\pi t}$')
##ax.plot(t, re1+re2, im1+im2, c=(0.3, 0.3, 0.3), linewidth=2, label=r'$2 \cos(2\pi t)$')
ax.legend()

ax.set_ylim3d([-2.0,2.0])
ax.set_zlim3d([-2.0,2.0])

ax.set_title(r'$2 \sin(2\pi t) =  je^{j2\pi t} - je^{-j2\pi t}$')
ax.set_xlabel('Time (s)')
ax.set_ylabel('Real Axis')
ax.set_zlabel('Imaginary Axis (j)')

fig.savefig('sin(2pit).png',dpi=300)
plt.close()
```

3.A.4 Orthogonality of phasors over infinite bounds (Figure 2.3.2.1.1)

```
"""plot depicting the orthogonality of cosines over time"""

import pylab
from numpy import cos, pi, linspace
x=linspace(-50,50,500)
y1=cos(x)
y2=cos(x*1.03)

pylab.figure()
pylab.plot(x,y1)
pylab.plot(x,y2)
pylab.ylim([-1.25,1.25])
pylab.yticks(linspace(-1,1,5))
pylab.xlim([-50.,50.])
pylab.xticks([-50.,0.,50.],
             (r'$-\infty \leftarrow$',r'$0$',
              r'$\rightarrow \infty ;$'))
pylab.savefig('orth_phasors.png',dpi=300)
```

3.A.5 Inner product line and fill plots (Figure 2.3.2.4.1-2)

```
import pylab
from numpy import pi,cos,sin,exp,linspace
from scipy.integrate import quad

f = lambda t: cos(6.*pi*t)*exp(-pi*t**2)
t=linspace(-2.,2.,500)
y=f(t)

ws=[6.,1.3]
for w in ws:
    print 'making plot for %.1f'%w
    re=y*cos(w*pi*t)
    im=y*sin(w*pi*t)

    re_int,re_err=quad(lambda x:f(x)*cos(w*pi*x),-2.,2.)
    im_int,im_err=quad(lambda x:f(x)*sin(w*pi*x),-2.,2.)

    pylab.figure(figsize=(16,12))

    pylab.subplot(311)
    pylab.plot(t,y)
    pylab.ylabel('x(t)')
    pylab.title(r'$x(t) = \cos(6\pi t) e^{-\pi t^2}$')
    pylab.xticks([])
    pylab.ylim([-1,1])

    pylab.subplot(312)
    pylab.plot(t,y,label='$x(t)$')
    pylab.plot(t,cos(w*pi*t),'r',
               label='$\cos(%.1f\pi t)$'%w)
    pylab.fill(t,re,'m',linewidth=0.,alpha=.6,
               label='$x(t) \cos(%.1f\pi t)$'%w)
    pylab.xticks([])
    pylab.ylim([-1,1])
    pylab.title(r'$Re\{X(%.1f\pi)\} = \int \,  x(t) \cos(%.1f\pi t) \,'
                '\mathrm{d}t = %.2f$'%(w,w,re_int))
    pylab.legend()
    pylab.ylabel('Real')

    pylab.subplot(313)
    pylab.plot(t,y,label='$x(t)$')
    pylab.plot(t,sin(w*pi*t),'r',
               label='$\sin(%.1f\pi t)$'%w)
    pylab.fill(t,im,'m',linewidth=0.,alpha=.6,
               label='$x(t) \sin(%.1f\pi t)$'%w)
    pylab.ylim([-1,1])
    pylab.title(r'$Im\{X(%.1f\pi)\} = \int \, x(t) \sin(%.1f\pi t) \,'
                '\mathrm{d}t = %.2f$'%(w,w,im_int))
    pylab.legend()
    pylab.ylabel('Imaginary (j)')
    pylab.xlabel('Time (s)')

    pylab.savefig('cos(6pit)e^(t^2),w=%.1f.png'%w,dpi=150)
```

3.A.6 Inner product 3d plots (Figure 2.3.2.5.2-3, 2.3.2.7.1)

```
import matplotlib as mpl
from mpl_toolkits.mplot3d import Axes3D
from numpy import pi, complex,exp,sin,cos,ones,real,imag,linspace
import matplotlib.pyplot as plt

j=complex(1j)
pi=pi

plots=[{'func':lambda t: exp(6.*j*pi*t)*exp(-pi*t**2),
        'title': r'$x(t) = e^{j6\pi t} e^{-\pi t^2}$',
        'w':'6',
        'fname':'e(j6pit)e^(-pit^2),w=6.png'},
       {'func':lambda t: sin(6.*pi*t)*exp(-pi*t**2),
        'title': r'$x(t) = \sin(6\pi t) e^{-\pi t^2}$',
        'w':'6',
        'fname':'sin(j6pit)e^(-pit^2),w=6.png'},
       {'func':lambda t: cos(6.*pi*t)*exp(-pi*t**2),
        'title': r'$x(t) = \cos(6\pi t) e^{-\pi t^2}$',
        'w':'6',
        'fname': r'cos(j6pit)e^(-pit^2),w=6.png'}]

for plot in plots:
    f=plot['func']
    w=plot['w']
    wi='-%s'%w

    t=linspace(-2.,2.,500)
    y=f(t)

    Fw=y*exp(-j*pi*float(w)*t)
    Fwi=y*exp(-j*pi*float(wi)*t)

    re=real(Fw)
    im=imag(Fw)
    rei=real(Fwi)
    imi=imag(Fwi)

    fig = plt.figure(figsize=(12,9))
    ax = fig.gca(projection='3d')

    ax.plot(t, re, im,'b', linewidth=2, alpha=0.8, label=r'$x(t) e^{-j%s\pi t}$' %w)
    ax.plot(t, re, ones(500)*-1.,'b', alpha=0.45)#, label=r'$re \{x(t, %s\pi)\}$'%w)
    ax.plot(t, ones(500)*1., im,'b', alpha=0.45)#, label=r'$Im \{x(t, %s\pi)\}$'%w)

    ax.plot(t, rei, imi,'g', linewidth=2, alpha=0.8, label=r'$x(t) e^{j%s\pi t}$' %w)
    ax.plot(t, rei, ones(500)*-1.,'g', alpha=0.45)#, label=r'$Re \{x(t, %s\pi)\}$'%wi)
    ax.plot(t, ones(500)*1., imi,'g', alpha=0.45)#, label=r'$Im \{x(t, %s\pi)\}$'%wi)

    ax.legend()
    ax.set_ylim3d([-1.,1.])
    ax.set_zlim3d([-1.,1.])

    ax.set_title(plot['title'])
    ax.set_xlabel('Time (s)')
    ax.set_ylabel('Real Axis')
    ax.set_zlabel('Imaginary Axis (j)')

    plt.savefig(plot['fname'],dpi=300)
    plt.close()
```

3.A.7 Even odd function decomposition plot (Figure 2.3.2.6.1)

```
"""Plots f(x)=x^4 - 5x^3 + x^2 + 10x - 10,
   f_e(x), and f_o(x) of f(x)"""

import pylab
from numpy import linspace

def f(x):
    return x**4 - 5*x**3 + x**2 + 10*x - 10

x=linspace(-5,5,100)

pylab.figure(figsize=(9,12))
pylab.subplot(3,1,1)
pylab.plot(x,f(x))
pylab.text(-4.5,-20,r'$f(x) = x^4 - 5x^3 + x^2 + 10x - 10$',fontsize=16)
pylab.axvline(color='k')
pylab.axhline(color='k')
pylab.xlim([-5,5])
pylab.ylim([-30,30])

pylab.subplot(3,1,2)
pylab.plot(x,f(x)+f(-x))
pylab.text(-4.5,-20,r'$f_e(x) = x^4 + x^2 - 10$',fontsize=16)
pylab.axvline(color='k')
pylab.axhline(color='k')
pylab.xlim([-5,5])
pylab.ylim([-30,30])

pylab.subplot(3,1,3)
pylab.plot(x,f(x)-f(-x))
pylab.text(-4.5,-20,r'$f_o(x) = 5x^3 + 10x$',fontsize=16)
pylab.axvline(color='k')
pylab.axhline(color='k')
pylab.xlim([-5,5])
pylab.ylim([-30,30])

pylab.savefig('even_odd_poly.png',dpi=300)
```

3.A.8 Discrete Fourier basis functions plot (Figure 2.3.3.2.1, .7.1)

```
"""Demonstrates how the aliasing of positive phasors > N/2
reflect negaive basis function in the discrete Fourier
transform"""

import pylab
from numpy import cos,sin,linspace,pi

N=9.

k=linspace(0,N-1,N)
x=linspace(0,N-1,N*10)

pylab.figure(figsize=(18,22))
for n in xrange(int(N)):

    pylab.subplot(N,1,(n+N/2+1)%N)
    pylab.title('n=%i,   N$-$n=%i'%(n,int(N-n)))
    pylab.plot(x,cos(2*pi*x*n/N),alpha=.5,
               label=r'$cos(2pi n/N)$')
    pylab.scatter(k,cos(2*pi*k*n/N),linewidth=1.)


    pylab.plot(x,sin(2*pi*x*n/N),'r',alpha=.5,
               label=r'$j sin(2pi n/N)$')
    pylab.scatter(k,sin(2*pi*k*n/N),color='r',marker='o',linewidth=1.)

    pylab.plot(x,cos(2*pi*x*-(N-n)/N),'b--',alpha=.5,
               label=r'$cos(-2pi (N-n)/N)$')
    pylab.plot(x,sin(2*pi*x*-(N-n)/N),'r--',alpha=.5,
               label=r'$j sin(-2pi (N-n)/N)$')

    pylab.ylim([-1.3,1.3])
    pylab.yticks([-1.,0.,1.])
    pylab.xlim([-.5,N-.5])
    pylab.xticks([])

    if n==0: pylab.legend(loc=4)
    if n==int(N/2): pylab.xticks(k)

pylab.savefig('folding.png')

M=15.
L=9.

k=linspace(0,L-1,L)
x=linspace(0,L-1,L*10)

pylab.figure(figsize=(16,30))
for n in xrange(int(M)):

    pylab.subplot(M,1,(n+M/2+1)%M)
##    pylab.subplot(N,1,n)
    pylab.title('n=%i,   M$-$n=%i'%(n,int(M-n)))
    pylab.plot(x,cos(2*pi*x*n/M),alpha=.5,
               label=r'$\cos(2\pi n/M)$')
    pylab.scatter(k,cos(2*pi*k*n/M),linewidth=1.)

    pylab.plot(x,sin(2*pi*x*n/M),'r',alpha=.5,
               label=r'$j \sin(2\pi n/M)$')
```

```
    pylab.scatter(k,sin(2*pi*k*n/M),color='r',marker='o',linewidth=1.)

    pylab.plot(x,cos(2*pi*x*-(M-n)/M),'b--',alpha=.5,
               label=r'$\cos(-2\pi (M-n)/M)$')
    pylab.plot(x,sin(2*pi*x*-(M-n)/M),'r--',alpha=.5,
               label=r'$j \sin(-2\pi (M-n)/M)$')

    pylab.ylim([-1.3,1.3])
    pylab.yticks([-1.,0.,1.])
    pylab.xlim([-.5,L-.5])
    pylab.xticks([])

    if n==0: pylab.legend(loc=4)
    if n==int(M/2): pylab.xticks(k)

pylab.savefig('DTFT_folding.png')
```

3.A.9 Discrete Fourier transform examples (Figure 2.3.3.2.1)

```
from numpy import abs,angle,array,concatenate,pi,exp,\
                  linspace,log10,mod,cos,sin,ones,nonzero,\
                  zeros,real,imag,iscomplex,isreal
from scipy import fft
import numpy

import matplotlib as mpl
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import matplotlib.pyplot as plt
from numpy.random import random
import pylab
from math import floor

j=complex(1j)

def square(X):
    def f(x):
        x=mod(x+pi,2.*pi)-pi
        if x < 0. : return -1.
        else      : return 1.
    return [f(k) for k in x]

def if_else(a,c,d):
    if a: return c
    else: return d

def step(x):
    y=zeros(N)
    y[N/2-8:N/2+8]=ones(16)

    return y

def ticks2textuple(yticks):
    return tuple([r'$%s$'%str(round(y,2)) for y in yticks])

N=128

funcs=[{'x':linspace(-1.,1.,N),
        'f':lambda x:exp(j*10*pi*x)*exp(-pi*x**2),
        'title': r'$x_k = e^{-j10\pi k/N}e^{-\pi k^2/N}$',
        'ylim':[-1.25,1.25],
        'yticks':[-1.,-.5,0.,.5,1.],
        'fname':'DFT_exp10pi.png'},
       {'x':linspace(-3*pi,3*pi-.1,N),
        'f':square,
        'title': r'$x_k = \mathrm{sgn}(\sin(6\pi k / N))$'
        'ylim':[-1.5,1.5],
        'yticks':[-1.5,-.75,-.5,0.,.75,1.5],
        'fname':'DFT_square.png'},
       {'x':'',
        'f':step,
        'title': r'$x_k = \mathrm{rect} ( ( k - N/2)/ 16 )$',
        'ylim':[-.3,.3],
        'yticks':[-.3,-.15,0.,.15,.3],
        'fname':'DFT_step.png'},
       {'x':linspace(-3*pi,3*pi-.1,N),
        'f':lambda x:sin(9.*x),
        'title':r'$x_k = \sin(54\pi k / N)$',
```

```
         'ylim':[-1.25,1.25],
         'yticks':[-1.,-.5,0.,.5,1.],
         'fname':'DFT_54sin.png'},
        {'x':linspace(-3*pi,3*pi-.1,N),
         'f':lambda x:cos(72./6.*x),
         'title':r'$x_k = \cos(72\pi k / N)$',
         'ylim':[-1.25,1.25],
         'yticks':[-1.,-.5,0.,.5,1.],
         'fname':'DFT_72cos.png'},
        {'x':linspace(-3*pi,3*pi-.1,N),
         'f':lambda x:cos(116./6.*x),
         'title':r'$x_k = \cos(116\pi k / N)$',
         'ylim':[-1.25,1.25],
         'yticks':[-1.,-.5,0.,.5,1.],
         'fname':'DFT_116cos.png'},
        {'x':linspace(-3*pi,3*pi-.1,N),
         'f':lambda x:cos(144./6.*x),
         'title': r'$x_k = \cos(144\pi k / N)$'
         'ylim':[-1.25,1.25],
         'yticks':[-1.,-.5,0.,.5,1.],
         'fname':'DFT_144cos.png'},
        {'x':linspace(-3*pi,3*pi-.1,N),
         'f':lambda x:cos(144./6.*x)+cos(116./6.*x),
         'title': r'$x_k = \cos(144\pi k / N)+\cos(116\pi k / N)$',
         'ylim':[-1.25,1.25],
         'yticks':[-1.,-.5,0.,.5,1.],
         'fname':'DFT_144_and116cos.png'},
        {'x':linspace(-3*pi,3*pi-.1,N),
         'f':lambda x:cos(x),
         'title': r'$x_k = \cos(6\pi k / N)$',
         'ylim':[-1.25,1.25],
         'yticks':[-1.,-.5,0.,.5,1.],
         'fname':'DFT_cos.png'},
        {'x':linspace(-3*pi,3*pi-.1,N),
         'f':lambda x:cos(x)*j,
         'title': r'$x_k = j \cos(6\pi k / N)$',
         'ylim':[-1.25,1.25],
         'yticks':[-1.,-.5,0.,.5,1.],
         'fname':'DFT_jcos.png'}]

for func in funcs:
    print func['title']
    x,f,ylim,yticks=func['x'],func['f'],func['ylim'],func['yticks']

    print ticks2textuple(yticks)
    y=f(x)
    c=fft(y)
    c=concatenate((c[N/2:],c[:N/2]))
    c*=(2./N)
    p=abs(c)
    a=angle(c)

    fig = plt.figure(figsize=(9,12))
    ax = fig.gca()

    pylab.subplot(511)
    if any(iscomplex(y)):
        for n in xrange(N):
            pylab.plot([n,n], [0.,imag(y[n])],'r')
        pylab.scatter(array(range(N)),imag(y),color='r',edgecolor='k')
    if len(nonzero(real(y).round(decimals=2))[0])>0:
```

```
    for n in xrange(N):
        pylab.plot([n,n], [0.,y[n]],'b')
    pylab.scatter(array(range(N)),real(y))
pylab.ylim([-2.,2.])
pylab.yticks([-2.,-1.,0.,1.,2.],ticks2textuple([-2.,-1.,0.,1.,2.]))
pylab.xlim([0-4,N+4])
pylab.text(0.,1.2,func['title'],fontsize=18)
pylab.xticks(linspace(0,N,9))
pylab.axhline(color='k')

pylab.subplot(512)
for n in xrange(N):
    pylab.plot([n-N/2,n-N/2], [0.,real(c[n])],'b')

pylab.scatter(array(range(N))-float(N)/2.,real(c))
pylab.ylim(ylim)
pylab.yticks(yticks,ticks2textuple(yticks))
pylab.xlim([-N/2-4,N/2+4])
pylab.text(-N/2,ylim[1]-(ylim[1]-ylim[0])*.25,r'$Re\{2 Xn / N\}$',fontsize=18)
pylab.xticks(linspace(-N/2,N/2,9))
pylab.axhline(color='k')

pylab.subplot(513)
for n in xrange(N):
    pylab.plot([n-N/2,n-N/2], [0.,imag(c[n])],'r')

pylab.scatter(array(range(N))-float(N)/2.,imag(c),color='r',edgecolor='k')
pylab.ylim(ylim)
pylab.yticks(yticks,ticks2textuple(yticks))
pylab.xlim([-N/2-4,N/2+4])
pylab.text(-N/2,ylim[1]-(ylim[1]-ylim[0])*.25,r'$Im\{2 Xn / N\}$',fontsize=18)
pylab.xticks(linspace(-N/2,N/2,9))
pylab.axhline(color='k')

pylab.subplot(514)
for n in xrange(N):
    pylab.plot([n-N/2,n-N/2], [0.,p[n]],'b')
pylab.scatter(array(range(N))-float(N)/2.,p)
pylab.ylim([-.075,ylim[1]-(ylim[1]-0.)*.1])
halfticks=linspace(0.,yticks[-1],len(yticks))
pylab.yticks(halfticks,ticks2textuple(halfticks))
pylab.xlim([-N/2-4,N/2+4])
pylab.text(-N/2,ylim[1]-(ylim[1]-0.)*.25,r'$|2 Xn / N|$',fontsize=18)
pylab.xticks(linspace(-N/2,N/2,9))
pylab.axhline(color='k')

pylab.subplot(515)
for n in xrange(N):
    pylab.plot([n-N/2,n-N/2], [0.,a[n]],'r')

pylab.scatter(array(range(N))-float(N)/2.,a,color='r',edgecolor='k')
pylab.ylim([-pi-pi/8,pi+pi/8])
pylab.yticks([-pi,-pi/2,0,pi/2,pi],
             (r'$-\pi$', r'$-\pi/2$', r'$0$', r'$\pi/2$' , r'$\pi$'))
pylab.xlim([-N/2-4,N/2+4])
pylab.text(-N/2,pi-pi/2,r'$ \angle Xn$',fontsize=18)
pylab.xticks(linspace(-N/2,N/2,9))
pylab.axhline(color='k')

fig.savefig(func['fname'],dpi=150)
```

3.A.10 DTFT Interpolated Mags (Figure 2.3.3.8.1-3, 2.3.3.11.1-2)

```
from numpy import array,cos,exp,linspace,complex,\
                 pi,zeros,log10,ones,concatenate,correlate
from numpy.random import random
from scipy import fft

import pylab

j=complex(1j)

Ns=[128,512] # size of input sequence
M=4096       # size after pad

windows=[
         {
          'w':lambda N:ones(N),
          'title':r'$Rectangular \; Window$',
          'dyn_range':60
         },
         {
          'w':lambda N:0.54 - 0.46*cos((2*pi*array(range(N)))/(N-1)),
          'title':r'$Hamming \; Window$',
          'dyn_range':80
         },
         {
          'w':lambda N:0.3232153788877343 \
              -0.4714921439576260*cos(2*pi/(N-1)*array(range(N))) \
              +0.1755341299601972*cos(4*pi/(N-1)*array(range(N))) \
              -2.849699010614994e-2*cos(6*pi/(N-1)*array(range(N))) \
              +1.261357088292677e-3*cos(8*pi/(N-1)*array(range(N))),
          'title':r'$Blackman-Harris \; 4 \; Window$',
          'dyn_range':150
         }
        ]

functions=[
           {
            'f':lambda x:cos(2.*pi*7.5*x/128),
            'title':r'$x_k = \cos(2\pi k 7.5 / 128)$',
            'fname':'7.5cos_windows.png'
           },
           {
            'f':lambda x:cos(2.*pi*7.5*x/128)+cos(2.*pi*9.1*x/128+.54),
            'title':r'$x_k = \cos(2\pi k 7.5 / 128)+ '
                     '\cos(2\pi k 9.1 / 128+.54)$',
            'fname':'7.5cos_windows_eq_mag.png'
           },
           {
            'f':lambda x:cos(2.*pi*7.5*x/128) \
                    +.06*cos(2.*pi*15.1*x/128+1.93) \
                    +(random(len(x))-0.5)*.28,
            'title':r'$x_k = \cos(2\pi k 7.5 / 128)+ '
                     '.06\cos(2\pi k 15.1 / 128+1.93)+noise$',
            'fname':'7.5cos_windows_dis_mag.png'
           },
           {
            'f':lambda x:cos(2.*pi*7.5*x/128) \
                    +.06*cos(2.*pi*15.1*x/128+1.93),
            'title':r'$x_k = \cos(2\pi k 7.5 / 128)+ '
                    '.06\cos(2\pi k 15.1 / 128+1.93)$',
```

```
            'fname':'7.5cos_windows_dis_mag_nonoise.png'
           }
          ]

for N in Ns:
    pad=M-N # size of pad

    for function in functions:
        print function['title']
        x=linspace(0.,N-1,N)
        f=function['f']

        fig = pylab.figure(figsize=(9,12))
        num_plots=len(windows)*2
        onplot=1

        for window in windows:
            print '\t',window['title']
            w=window['w']
            y=w(N)*f(x)

            c=fft(concatenate((y,zeros(pad))))
            p=20.*log10(abs(c))
            p-=max(p)

            pylab.subplot(num_plots,1,onplot)
            if onplot==1:
                pylab.title(function['title'],fontsize=18)

            for n in xrange(N):
                pylab.plot([n,n], [0.,y[n]],'b')

            if N<=128:
                pylab.scatter(array(range(N)),y)
            pylab.ylim([-2.,2.])
            pylab.yticks([-2.,-1.,0.,1.,2.])
            pylab.xlim([0-4,N+4])
            pylab.text(0.,1.2,window['title'],fontsize=18)
            pylab.xticks(linspace(0,N,9))
            pylab.axhline(color='k')

            onplot+=1

            pylab.subplot(num_plots,1,onplot)
            for n in xrange(32*(M/128)):
                pylab.plot([128*n/float(M),128*n/float(M)],
                           [-window['dyn_range'],p[n]],'b')
            pylab.ylim([-window['dyn_range'],5.])
            pylab.xlim([0,31.5])

            onplot+=1

        fig.savefig('DFT_N=%i,'%N+function['fname'],dpi=150)
    ##    fig.savefig('PSD'+function['fname'],dpi=150)

        pylab.clf(); pylab.cla(); pylab.close(); del fig

        print
```

3.A.11 Sinc function, abs(sinc), 20log10(sinc) (Figure 2.3.3.10.1)

```
from numpy import array,cos,sin,linspace,pi,complex,logspace,log10

import pylab

def ticks2textuple(yticks):
    return tuple([r'$%s$'%str(round(y,2)) for y in yticks])

sinc=lambda w,N :sin(pi*N*w)/(pi*w)

x=linspace(-pi*4,pi*4,2**14)
y=sinc(x,1)

pylab.figure(figsize=(8,10))
pylab.subplot(3,1,1)
pylab.plot(x,y)
pylab.axhline(color='k')
ymin,ymax=pylab.ylim()
pylab.text(-3.8*pi,ymax-(ymax-ymin)*.05,
           r'$\frac{\sin(\pi t)}{\pi t}$',
           fontsize=18)
pylab.xlim([-4*pi,4*pi])
pylab.xticks(linspace(-4*pi,4*pi,5),('','','','',''))
yticks=pylab.yticks()[0]
pylab.yticks(yticks,ticks2textuple(yticks))

pylab.subplot(3,1,2)
pylab.plot(x,abs(y))
ymin,ymax=pylab.ylim()
pylab.text(-3.8*pi,ymax-(ymax-ymin)*.15,
           r'$\left| \frac{\sin(\pi t)}{\pi t} \right|$',
           fontsize=18)
pylab.xlim([-3.8*pi,4*pi])
pylab.xticks(linspace(-4*pi,4*pi,5),('','','','',''))
yticks=pylab.yticks()[0]
pylab.yticks(yticks,ticks2textuple(yticks))

pylab.subplot(3,1,3)
pylab.xlim([-4*pi,4*pi])
pylab.plot(x,20*log10(abs(y)))
pylab.ylim(ymin=-60.)
ymin,ymax=pylab.ylim()
pylab.text(-3.8*pi,ymax-(ymax-ymin)*.15,
           r'$\mathrm{20log_{10}} \left| \frac{\sin(\pi t)}{\pi t} \right|$',
           fontsize=18)
pylab.xticks(linspace(-4*pi,4*pi,5),
             (r'$-4\pi$',r'$-2\pi$',r'$0$',r'$2\pi$',r'$4\pi$'))
yticks=pylab.yticks()[0]
pylab.yticks(yticks,ticks2textuple(yticks))

pylab.savefig('sinc_function.png',dpi=150)
```

3.A.12 Continuous Fourier Transform of windowed Cosine (2.3.3.10.2)

```
from numpy import array,cos,sin,linspace,pi,complex,logspace,log10

import pylab

x=linspace(-2.25,2.25,2**14)

f=lambda w,N : sin(pi*N*(w-1))/(w-1)+sin(pi*N*(w+1))/(w+1)
onplot=1
ylim=array([-2.,10.])

pylab.figure(figsize=(8,10))
for i in [2,4,8,16,32,64,128,256,512]:
    print i
    pylab.subplot(9,1,onplot)
    pylab.plot(x,f(x,i))
    pylab.xlim([-2.25,2.25])
    ymin,ymax=pylab.ylim()
    pylab.yticks([ymin,ymax],(r'$%i$'%int(ymin),r'$%i$'%int(ymax)))
    pylab.xticks([-2,-1,0,1,2],('','','','',''))
    pylab.text(-2.2,ymax*.6,
               '$F\{ \cos(2 \pi t) \Pi (%i \pi t) \}$'%i,fontsize=11)

    onplot+=1
    ylim*=2

pylab.xticks([-2,-1,0,1,2],(r'$-2$',r'$-1$',r'$0$',r'$1$',r'$2$'))
pylab.savefig('continuous_cosine_by_period.png',dpi=150)
```

3.A.13 Blackman-Harris4 Frequency Resolution (Figure 2.3.3.10.3)

```
import pylab
from numpy import pi,ones,zeros,linspace,log10,imag,real,concatenate,cos,array
from scipy import fft

M=8192

w=lambda N:0.3232153788877343
          -0.4714921439576260*cos(2*pi/(N-1)*array(range(N)))
          +0.1755341299601972*cos(4*pi/(N-1)*array(range(N)))
          -2.849699010614994e-2*cos(6*pi/(N-1)*array(range(N)))
          +1.261357088292677e-3*cos(8*pi/(N-1)*array(range(N)))

pylab.figure()
onplot=1
for i in range(1,9):
    print i

    N=128*i
    pad=M-N
    c=fft(concatenate((w(N),zeros(pad))))
    c=concatenate((c[-2*(M/128):],c[:-2*(M/128)]))
    pylab.plot(linspace(-2.,8.,10*(M/128)),
               20*log10(abs(c)[:10*(M/128)]),
               alpha=.8,label='N=%i'%N)

    onplot+=1

pylab.xticks(linspace(-2,8,11))
pylab.legend()
pylab.title('Zoom View of Blackman-Harris4 Window DFT Magnitudes')
pylab.xlabel('Magnitude (dB)')
pylab.xlabel('DFT Frequency Bin Relative to N = 128')
pylab.savefig('blackman_harris4_mags.png',dpi=150)
```

3.A.14 Example of window translation (Figure 2.3.4.1)

```
import pylab
from numpy.random import random
from numpy import linspace, sin,pi,cos,array

def ticks2textuple(ticks,precision=2):
    return tuple([r'$%s$'%str(round(tk,precision)) for tk in ticks])

blackman_harris4 = lambda N: 0.3232153788877343 \
              -0.4714921439576260*cos(2*pi/(N-1)*array(range(N))) \
              +0.1755341299601972*cos(4*pi/(N-1)*array(range(N))) \
              -2.849699010614994e-2*cos(6*pi/(N-1)*array(range(N))) \
              +1.261357088292677e-3*cos(8*pi/(N-1)*array(range(N)))

x=linspace(-6*pi,6*pi,12000)

y=.2*sin(.245*x+random()*2*pi) + \
  .2*sin(.434*x+random()*2*pi) + \
  .2*sin(.745*x+random()*2*pi) + \
  .2*sin(1.53*x+random()*2*pi) + \
  .2*sin(2.73*x+random()*2*pi) + \
  .2*sin(3.30*x+random()*2*pi) + \
  .2*sin(5.60*x+random()*2*pi)+ \
  .2*sin(9.10*x+random()*2*pi)

w=blackman_harris4(4000)

pylab.figure(figsize=(10,8))
pylab.plot(x,y,'b', alpha=.3,label=r'$x(t)$')

pylab.plot(linspace(-2*pi,2*pi,4000),
           w,'r--',alpha=.8,label=r'$w(t)$')
pylab.plot(linspace(-2*pi,2*pi,4000),
           w*y[4000:8000],'r',alpha=.8,label=r'$x(t)w(t)$')
pylab.axvline(0., color='k')

pylab.plot(linspace(0,4*pi,4000),
           w,'m--',alpha=.8,label=r'$w(t-t_0)$')
pylab.plot(linspace(0,4*pi,4000),
           w*y[6000:10000],'m',alpha=.8,label=r'$x(t)w(t-t_0)$')
pylab.axvline(2*pi,color='k')

pylab.axhline(color='k',linewidth=2.)

pylab.xlim([-4*pi,6*pi])
ticks=linspace(-6*pi,6*pi,7)
pylab.xticks(linspace(-4*pi,6*pi,6),
             (r'',r'',r'$0$',r'$t_0$',r'',r'$t \; \rightarrow$'),
             fontsize=16)

pylab.ylim([-1.25,1.25])
ticks=linspace(-1,1,5)
pylab.yticks(ticks,ticks2textuple(ticks,precision=2), fontsize=16)

pylab.legend(loc='upper left')
pylab.savefig('STFT_windowing.png')
pylab.close()
```

3.A.15 Fixed-resolution lattice structure (Figure 2.3.4.1)

```
import pylab
from numpy import array,ones,linspace

N=2**14
ns=[128,1024]
titles={1024:'Good Frequency Resolution, but Poor Time Resolution',
         128:'Poor Frequency Resolution, but Good Time Resolution'}

pylab.figure(figsize=(12,12))
numplots=len(ns)

for i,n in enumerate(ns):
    pylab.subplot(numplots,1,i+1)

    step=n
    s0=n/2
    X,T,F=[],[],[]
    j=0
    while s0 <= N-(n/2):
        X.append(ones(n/2))
        T.append([j*(100./(N/n)) for i in xrange(n/2)])
        F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))
        s0+=step
        j+=1

    # Work around to very very annoying pylab idiosyncracy
    # (I would call it a bug, but it probably makes sense to someone.)
    # These don't actually get plotted. The pcolor plots are one column
    # short if we don't append these.
    X.append(ones(n/2))
    T.append([j*(100./(N/n)) for i in xrange(n/2)])
    F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))

    T,F,X=array(T),array(F),array(X)
    pylab.pcolor(T,F,X,alpha=0.)

    for t in T[:,0]:
        pylab.axvline(t,color='k')

    for f in F[0,:]:
        pylab.axhline(f,color='k')

    pylab.xticks([])
    pylab.ylim([0.,10.24/2.])
    pylab.xlim([0.,50.])
    pylab.yticks([])

    pylab.title(titles[n],fontname='Times New Roman')
    pylab.ylabel(r'$Frequency \; (Hz)$')
    pylab.xlabel(r'$Time \; (seconds)$')

pylab.savefig('STFT_of_sweep_grid.png', dpi=150)
```

3.A.16 Fixed-resolution log chirp decomposition (Figure 2.3.4.2)

```
import pylab
from numpy import array,logspace, linspace, \
     log10, log, pi, sin,abs,zeros,ones,cos,exp
from scipy import fft
from scipy.signal.waveforms import chirp
from matplotlib.colors import LogNorm

from numpy.random import normal

def if_else(a,b,c):
    if a: return b
    return c

def ticks2textuple(ticks,precision=2):
    return tuple([r'$%s$'%str(round(tk,precision)) for tk in ticks])

rect = lambda N : ones(N)
hamming = lambda N : 0.54 - 0.46*cos((2*pi*array(range(N)))/(N-1))
blackman_harris4 = lambda N: 0.3232153788877343 \
              -0.4714921439576260*cos(2*pi/(N-1)*array(range(N))) \
              +0.1755341299601972*cos(4*pi/(N-1)*array(range(N))) \
              -2.849699010614994e-2*cos(6*pi/(N-1)*array(range(N))) \
              +1.261357088292677e-3*cos(8*pi/(N-1)*array(range(N)))

N=2**14
ns=[64,128,256,512,1024,2048]
numplots=len(ns)+1

t = linspace(0,100, N)
x=chirp(t, f0 = 1., t1 = 100., f1 = 10., method = 'log')

pylab.figure(figsize=(9,12))
pylab.subplot(numplots,1,1)
pylab.plot(t,x)
xticks=linspace(0.,100.,11)
pylab.xticks(xticks,['' for i in xrange(12)])
pylab.ylim([-1.25,1.25])
pylab.yticks([-1,-.5,0,.5,1.],
             ticks2textuple([-1,-.5,0,.5,1.]))
pylab.title(r'$x(t) = logsweep \; from \; 1Hz \; to \; 10Hz$')

for i,n in enumerate(ns):
    w=hamming(n)
    pylab.subplot(numplots,1,i+2)

    step=n
    s0=n/2
    X,T,F=[],[],[]
    j=0
    while s0 <= N-(n/2):
        X.append(abs(fft(w*x[s0-n/2:s0+n/2]))[:n/2])
        X[-1]=[if_else(v<1e-5,1e-5,v)for v in X[-1]]

        T.append([j*(100./(N/n)) for i in xrange(n/2)])

        print s0+n/2,j*(100./(N/n))
        F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))
        s0+=step
```

```
        j+=1

    # Work around to very very annoying pylab idiosyncracy
    # (I would call it a bug, but it probably makes sense to someone.)
    # These don't actually get plotted. The pcolor plots are one column
    # short if we don't append these.
    X.append(ones(n/2))
    T.append([j*(100./(N/n)) for i in xrange(n/2)])
    F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))

    X,T,F=array(X),array(T),array(F)
    pylab.pcolor(T,F,X, norm=LogNorm(vmin=X.min(), vmax=X.max()))


    for t in T[:,0]:
        if len(T[:,0])<=256:
            pylab.axvline(t,color='k')

    for f in F[0,:]:
        if len(F[0,:])<=128:
            pylab.axhline(f,color='k')

    pylab.xticks(xticks,['' for i in xrange(12)])
    pylab.ylim([0.,10.24])
    pylab.yticks([0,2.5,5.,7.5,10.],
             ticks2textuple([0,2.5,5.,7.5,10.]))

    pylab.title(r'$%.2f s \; Hamming \; Window$'%(100.*n/float(N)))
    pylab.ylabel(r'$Frequency \; (Hz)$')


pylab.xticks(xticks,ticks2textuple(xticks))
pylab.xlabel(r'$Time \; (seconds)$')
pylab.savefig('STFT_of_sweep.png', dpi=150)
```

3.A.17 Multi-resolution log chirp decomposition (Figure 2.3.4.3)

```
import pylab
from numpy import array,logspace, linspace,log10, log, pi, sin,abs,zeros,ones,cos,exp
from scipy import fft
from scipy.signal.waveforms import chirp
from matplotlib.colors import LogNorm

from numpy.random import normal

def if_else(a,b,c):
    if a: return b
    return c

def ticks2textuple(ticks,precision=2):
    return tuple([r'$%s$'%str(round(tk,precision)) for tk in ticks])

rect = lambda N : ones(N)
hamming = lambda N : 0.54 - 0.46*cos((2*pi*array(range(N)))/(N-1))
blackman_harris4 = lambda N: 0.32321537888877343 \
              -0.4714921439576260*cos(2*pi/(N-1)*array(range(N))) \
              +0.1755341299601972*cos(4*pi/(N-1)*array(range(N))) \
              -2.849699010614994e-2*cos(6*pi/(N-1)*array(range(N))) \
              +1.261357088292677e-3*cos(8*pi/(N-1)*array(range(N)))

N=2**14
ns=[64,128,256,512,1024,2048,4096,8192]
numplots=len(ns)+2

t = linspace(0,100, N)

x=chirp(t, f0 = 1., t1 = 100., f1 = 10., method = 'log')

pylab.figure(figsize=(9,9))
pylab.subplot(numplots,1,1)
pylab.plot(t,x)
xticks=linspace(0.,100.,11)
pylab.xticks(xticks,['' for i in xrange(12)])
pylab.ylim([-1.25,1.25])
pylab.yticks([-1,0,1],ticks2textuple([-1,0,1]))
pylab.title(r'$x(t) = logsweep \; from \; 1Hz \; to \; 10Hz$')

pylab.subplot(numplots,1,2,frame_on=False,aspect='equal')
pylab.xticks([])
pylab.yticks([])

ylim=array([10.24*(float(len(ns)-1)/float(len(ns))),10.24])
yticks=linspace(0.,10.,21)
print yticks

pylab.subplots_adjust(hspace=0.0)
for i,n in enumerate(ns):
    w=hamming(n)
    pylab.subplot(numplots,1,i+3)

    step=n
    s0=n/2
    X,T,F=[],[],[]
    j=0
    while s0 <= N-(n/2):
```

```
            X.append(abs(fft(w*x[s0-n/2:s0+n/2]))[:n/2])
            X[-1]=[if_else(v<1e-5,1e-5,v)for v in X[-1]]
            T.append([j*(100./(N/n)) for i in xrange(n/2)])
            F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))
            s0+=step
            j+=1

        # Work around to very VERY annoying pylab idiosyncracy
        # (I would call it a bug, but it probably makes sense to someone.)
        # These don't actually get plotted. The pcolor plots are one column
        # short if we don't append these.
        X.append(ones(n/2))
        T.append([j*(100./(N/n)) for i in xrange(n/2)])
        F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))

        X,T,F=array(X),array(T),array(F)

        pylab.pcolor(T,F,X, norm=LogNorm(vmin=X.min(), vmax=X.max()))

        for t in T[:,0]:
            if len(T[:,0])<=256:
                pylab.axvline(t,color='k')

        for f in F[0,:]:
            if len(F[0,:])<=512:
                pylab.axhline(f,color='k')

        print F[0,:]
        pylab.xticks(xticks,['' for i in xrange(12)])
        pylab.ylim(ylim)
        ticks=[t for t in yticks if t<ylim[-1] and t>ylim[0]]
        pylab.yticks(ticks,ticks2textuple(ticks))

        if n==ns[int(len(ns)/2.)]:
            pylab.ylabel(r'$Frequency \; (Hz)$')

        ylim-=10.24*(1./float(len(ns)))

    pylab.xticks(xticks,ticks2textuple(xticks))
    pylab.xlabel(r'$Time \; (seconds)$')
    pylab.savefig('STFT_of_sweep_wavelet.png', dpi=150)
```

3.A.18 Multi-resolution lattice structure (Figure 2.3.4.4)

```
import pylab
from numpy import array,ones,linspace

N=2**14
ns=[128,256,512,1024,2048]
numplots=len(ns)

pylab.figure(figsize=(12,12))

ylim=array([10.24*6./8.,10.24*7./8.])

pylab.subplots_adjust(hspace=0.0)
for i,n in enumerate(ns):
    pylab.subplot(numplots,1,i+1)

    step=n
    s0=n/2
    X,T,F=[],[],[]
    j=0
    while s0 <= N-(n/2):
        X.append(ones(n/2))
        T.append([j*(100./(N/n)) for i in xrange(n/2)])
        F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))
        s0+=step
        j+=1

    # Work around to very VERY annoying pylab idiosyncracy
    # (I would call it a bug, but it probably makes sense to someone.)
    # These don't actually get plotted. The pcolor plots are one column
    # short if we don't append these.
    X.append(ones(n/2))
    T.append([j*(100./(N/n)) for i in xrange(n/2)])
    F.append(array((N/100.*linspace(0,n/2-1,n/2)/n)))

    X,T,F=array(X),array(T),array(F)

    for t in T[:,0]:
        pylab.axvline(t,color='k')

    for f in F[0,:]:
        pylab.axhline(f,color='k')

    if n==512:
        pylab.ylabel(r'$Frequency \; (Hz)$')

    pylab.ylim(ylim)
    pylab.yticks([])
    pylab.xlim([0.,50.])
    pylab.xticks([])

    ylim-=10.24*1./8.

pylab.xlabel(r'$Time \; (seconds)$')
pylab.savefig('FourierWavelet_of_sweep_wavelet_grid.png', dpi=150)
```

3.A.19 Example Window Envelopes (Figure 2.3.1.1)

```
import pylab

from numpy import array,pi,exp,linspace,real,imag,cos

j=1j

N=4096
t_0=-.5
t_end=.5

hamming = lambda N : 0.54 - 0.46*cos((2*pi*array(range(N)))/(N-1))

def g(n,w, sr=N/(t_end-t_0), window=hamming):
    m=w*(float(n)/sr)
    return window(n)*exp(j*2.*pi*(array(range(n))-n/2.)*m/n)

nws=[(N,5.),(N,10.),(N,20.)]
numplots=len(nws)
tau=.15

pylab.figure(figsize=(6,8))

for i,(n,w) in enumerate(nws):
    pylab.subplot(numplots,1,i+1, frameon=False)
    pylab.plot(linspace(-.5+tau,.5+tau,N),real(g(n,w)),
               'b',label=r'$Re\{g_{\tau,\omega}(t)\}$')
    pylab.plot(linspace(-.5+tau,.5+tau,N),imag(g(n,w)),
               'r',label=r'$Im\{g_{\tau,\omega}(t)\}$')
    pylab.text(tau+.05,1.2,r'$\omega=%.0f\;Hz$'%w)

    pylab.axvline(0,color='k')
    pylab.axhline(0,color='k')


    pylab.axvline(tau,color='k', linestyle='--')
    pylab.axvline(-.5+tau,color='k', linestyle='--')
    pylab.axvline(.5+tau,color='k', linestyle='--')
    pylab.ylim([-1.5,1.5])
    pylab.yticks([])
    pylab.xlim([tau-1.5,tau+1.])
    pylab.xticks([-.5+tau,tau,.5+tau],
                 (r'$\tau-0.5$',r'$\tau$',r'$\tau+0.5$'))
    pylab.text(tau+.8,0.05,
               r'$t \; \rightarrow$')

    if i==0: pylab.legend(loc=(0.,.6))

pylab.savefig('fixed_width_envelopes.png',dpi=150)
```

3.A.20 Cross correlation with linear sweep plot (Figure 2.3.2.1)

```
import pylab
from numpy import array,logspace, linspace, \
     log10,log,pi, sin,abs,zeros,ones,cos,exp,\
     real,imag,concatenate,conjugate,correlate,arange
from scipy import fft,ifft
from scipy.signal.waveforms import chirp
from matplotlib.colors import LogNorm

from numpy.random import normal

j=1j

def if_else(a,b,c):
    if a: return b
    return c

def ticks2textuple(ticks,precision=2):
    return tuple([r'$%s$'%str(round(tk,precision)) for tk in ticks])

def cross_correlate(g,h):
    """g and h should both be even. len(h)>len(g)"""
    M=len(h)
    _h=concatenate((h[::-1],h,h[::-1]))
    m=len(_h)
    sym_pad_l=(len(_h)-len(g))/2
    g=concatenate((zeros(sym_pad_l),g,zeros(sym_pad_l)))
    cc=ifft(fft(conjugate(g))*fft(_h))
##    cc=ifft(fft(g*-1.)*fft(h))

    return concatenate((cc[sym_pad_l:],cc[:sym_pad_l]))

def gauss(N,sd=0.5):
    n=array(range(N))
    return exp(-.5*((n-(N-1.)/2.)/(sd*(N-1.)/2.))**2)

rect = lambda N : ones(N)
hamming = lambda N : 0.54 - 0.46*cos((2*pi*array(range(N)))/(N-1))
blackman_harris4 = lambda N: 0.3232153788877343                    \
              -0.4714921439576260*cos(2*pi/(N-1)*array(range(N)))   \
              +0.1755341299601972*cos(4*pi/(N-1)*array(range(N)))   \
              -2.849699010614994e-2*cos(6*pi/(N-1)*array(range(N))) \
              +1.261357088292677e-3*cos(8*pi/(N-1)*array(range(N)))

N=2**14
t_end=10.

def g(n,w, sr=N/t_end, window=gauss):
    m=w*(float(n)/sr)
    return window(n)*exp(j*2.*pi*(array(range(n))-n/2.)*m/n)

t = linspace(0,t_end, N)
x=chirp(t, f0 = 1., t1 = t_end, f1 = 10.)
frq=3.

n,w=2048,frq

pylab.figure()
```

```
pylab.subplot(1,1,1)
pylab.plot(t,x,'m', alpha=.3, label=r'$x(t)$')

g_nw=g(n,w)
cc=cross_correlate(g_nw,x)

pylab.plot(t[:n]+7.00,real(g_nw)-8.0,'b', alpha=.5)
pylab.plot(t[:n]+7.00,imag(g_nw)-8.0,'r', alpha=.5)
pylab.text(5.75,-5.00,r'$Gaussian\; Window$')
pylab.text(6.00,-6.10,r'$duration=1.25 s$')
pylab.text(6.00,-7.10,r'$\omega=3 Hz$')

pylab.plot(t,real(cc[len(t):2*len(t)])*(float(n)**-.5),
           'b', alpha=.5, label=r'$Re\{(x\star g_{\tau,\omega})(t)\}$')
pylab.plot(t,imag(cc[len(t):2*len(t)])*(float(n)**-.5),
           'r', alpha=.5, label=r'$Im\{(x\star g_{\tau,\omega})(t)\}$')
pylab.plot(t,abs(cc[len(t):2*len(t)])*(float(n)**-.5),
           'k', alpha=.5, label=r'$|(x\star g_{\tau,\omega})(t)|$')

pylab.legend()

yticks=pylab.yticks()[0]
pylab.yticks(yticks,ticks2textuple(yticks))

xticks=pylab.xticks()[0]
pylab.xticks(xticks,ticks2textuple(xticks))

pylab.title(r'$x(t)= linear \; sweep \; from \; 1Hz \; to \; 10Hz\;^\dag$')
pylab.text(5.5,-14.,r'$\dag \; sweep \; is \; symmetrically \; padded$')
pylab.xlabel(r'$Time \; (seconds)$')
pylab.savefig('cross_correlation_simple_sym_padding.png',dpi=150)
```

3.A.21 Varied Width Window Envelopes (Figure 2.3.1.3)

```
import pylab

from numpy import array,pi,exp,linspace,real,imag,cos

j=1j

N=4096
t_0=-.5
t_end=.5

hamming = lambda N : 0.54 - 0.46*cos((2*pi*array(range(N)))/(N-1))

def g(n,w, sr=N/(t_end-t_0), window=hamming):
    m=w*(float(n)/sr)
    return window(n)*exp(j*2.*pi*(array(range(n))-n/2.)*m/n)

nws=[(N,5.),(N/2,10.),(N/4,20.)]
numplots=len(nws)
tau=.15

pylab.figure(figsize=(6,8))

for i,(n,w) in enumerate(nws):
    pylab.subplot(numplots,1,i+1, frameon=False)
    pylab.plot(linspace(-.5*(n/float(N))+tau,
                        .5*(n/float(N))+tau,n),
               real(g(n,w)),
               'b',alpha=.6,
               label=r'$Re\{g_{\tau,\omega}(t)\}$')
    pylab.plot(linspace(-.5*(n/float(N))+tau,
                        .5*(n/float(N))+tau,n),
               imag(g(n,w)),
               'r',alpha=.6,
               label=r'$Im\{g_{\tau,\omega}(t)\}$')
    if n==N:
        pylab.text(tau+.05,1.2,r'$\omega=%.0f\;Hz$'%w)
    else:
        pylab.text(.5*(n/float(N))+tau+.05,1.2,
                   r'$\omega=%.0f\;Hz$'%w)

    pylab.axvline(0,color='k')
    pylab.axhline(0,color='k')

    pylab.axvline(tau,color='k',
                  linestyle='--')
    pylab.axvline(-.5*(n/float(N))+tau,color='k',
                  linestyle='--')
    pylab.axvline(.5*(n/float(N))+tau,color='k',
                  linestyle='--')
    pylab.ylim([-1.5,1.5])
    pylab.yticks([])
    pylab.xlim([tau-1.5,tau+1.])
    pylab.xticks([-.5*(n/float(N))+tau,
                  tau,
                  .5*(n/float(N))+tau],
                 (r'$\tau-%.1f \;$'%(.5*(n/float(N))),
                  r'$\tau$',
```

```
                r'$\; \tau+%.1f$'%(.5*(n/float(N))))
    pylab.text(tau+.8,0.05,
               r'$t \; \rightarrow$')

    if i==0: pylab.legend(loc=(0.,.6))

pylab.savefig('multi_width_envelopes.png',dpi=150)
```

3.A.22 Continous Wavelet Transform of logsweep (Figure 2.3.3.1)

cwt_logsweep.py

```
# cwt_logsweep.py
import pylab
from numpy import array,logspace, linspace,log10, log, \
     pi, sin,abs,zeros,ones,cos,exp,mean
from scipy import fft
from scipy.signal.waveforms import chirp
from matplotlib.colors import LogNorm

from numpy.random import normal

def if_else(a,b,c):
    if a: return b
    return c

def ticks2textuple(ticks,precision=2):
    return tuple([r'$%s$'%str(round(tk,precision)) for tk in ticks])

rect = lambda N : ones(N)
hamming = lambda N : 0.54 - 0.46*cos((2*pi*array(range(N)))/(N-1))
blackman_harris4 = lambda N: 0.3232153788877343 \
              -0.4714921439576260*cos(2*pi/(N-1)*array(range(N))) \
              +0.1755341299601972*cos(4*pi/(N-1)*array(range(N))) \
              -2.849699010614994e-2*cos(6*pi/(N-1)*array(range(N))) \
              +1.261357088292677e-3*cos(8*pi/(N-1)*array(range(N)))

def downsample(vector, factor):
    """
    downsample(vector, factor):
        Downsample (by averaging) a vector by an integer factor.
    """
    if (len(vector) % factor):
        print "Length of 'vector' %i is not divisible by 'factor'=%d!" \
              % (len(vector),factor)
        return 0
    vector.shape = (len(vector)/factor, factor)
    return mean(vector, axis=1)

N=2**14
ns=[64,128,256,512,1024,2048,4096,8192]
numplots=len(ns)+1

t_end=100.
t = linspace(0,t_end, N)
x=chirp(t, f0 = 1., t1 = t_end, f1 = 10., method = 'log')
fs=N/100.
freqs=linspace(1./t_end,fs/2.,N)

##import matplotlib.pyplot as plt
import pycwt
import numpy as np

freqs=linspace(10./128.,10.,128)

print len(t),len(freqs)
```

```
fc=6.4
r=pycwt.cwt_f(x,freqs,fs,pycwt.Morlet(fc))
rr=r.real**2+r.imag**2
print rr.shape
d_rr=[]
for rr_ in rr:
    d_rr.append(downsample(rr_,16))
d_rr=array(d_rr)

T,F = pylab.meshgrid(downsample(t[::],16), freqs)
print d_rr.shape,T.shape,F.shape

pylab.figure(figsize=(9,12))
pylab.subplot(2,1,1)
pylab.plot(t,x)
xticks=linspace(0.,100.,11)
pylab.xticks(xticks,['' for i in xrange(12)])
##pylab.xlim(99,100)
pylab.ylim([-1.25,1.25])
pylab.yticks([-1,-.5,0,.5,1.],
             ticks2textuple([-1,-.5,0,.5,1.]))
pylab.title(r'$x(t) = logsweep \; from \; 1Hz \; to \; 10Hz$')

pylab.subplot(2,1,2)
pylab.pcolor(T,F,d_rr,shading='flat')

pylab.xticks(xticks,ticks2textuple(xticks))
yticks=pylab.yticks()[0]
pylab.yticks(yticks,ticks2textuple(yticks))

pylab.title(r'$Continuous Wavelet Transform$')
pylab.ylabel(r'$Frequency \; (Hz)$')
pylab.xlabel(r'$Time \; (seconds)$')

pylab.savefig('cwt_of_sweep_wavelet.png', dpi=150)
```

pycwt.py

```
# from http://pypi.python.org/pypi/swan
#
# author: Alexey Brazhe <brazhe at gmail com>
#
# redistributed under GNU General public license (GPL)
#
# Continuous wavelet transfrom via Fourier transform
# Collection of routines for wavelet transform via FFT algorithm


#-- Some naming and other conventions --
# use f instead of omega wherever rational/possible
# *_ft means Fourier transform

#-- Some references --
# [1] Mallat, S.  A wavelet tour of signal processing
# [2] Addison, Paul S. The illustrated wavelet transform handbook

import numpy
from numpy.fft import fft, ifft, fftfreq

#try:
#    from scipy.special import gamma
#except:



pi = numpy.pi

class DOG:
    """Derivative of Gaussian, general form"""
    # Incomplete, as the general form of the mother wavelet
    # would require symbolic differentiation.
    # Should be enough for the CWT computation, though

    def __init__(self, m = 1.):
        self.order = m
        self.fc = (m+.5)**.5 / (2*pi)

    def psi_ft(self, f):
        c = 1j**self.order / numpy.sqrt(gamma(self.order + .5)) #normalization
        w = 2*pi*f
        return c * w**self.order * numpy.exp(-.5*w**2)

class Mexican_hat:
    def __init__(self, sigma = 1.0):
        self.sigma = sigma
        self.fc = .5 * 2.5**.5 / pi
    def psi_ft(self, f):
        """Fourier transform of the Mexican hat wavelet"""
        c = numpy.sqrt(8./3.) * pi**.25 * self.sigma**2.5
        wsq = (2. * pi * f)**2.
        return -c * wsq * numpy.exp(-.5 * wsq * self.sigma**2.)
    def psi(self, tau):
        """Mexian hat wavelet as described in [1]"""
        xsq = (tau / self.sigma)**2.
        # normalization constant from [1]
        c = 2 * pi**-.25 / numpy.sqrt(3 * self.sigma)
        return c * (1 - xsq) * numpy.exp(-.5 * xsq)
```

```
    def set_f0(self, f0):
        pass

class Morlet:
    def __init__(self, f0 = 1.5):
        self.set_f0(f0)
    def psi_ft(self, f):
        """
        Fourier transform of the approximate Morlet wavelet
        f0 should be more than 0.8 for this function to be correct.
        """
        return (pi**-.25) * (2.**.5) * \
               numpy.exp(-.5 * (2. * pi * (f - self.fc))**2.)
    def set_f0(self, f0):
        self.f0 = f0
        self.fc = f0


def cwt_a(signal, scales, sampling_scale = 1.0, wavelet=Mexican_hat()):
    """ Continuous wavelet transform via fft. Scales version.  """
    # FFT of the signal
    signal_ft = fft(signal)

    # create the matrix beforehand
    W = numpy.zeros((len(scales), len(signal)),'complex')

    # FFT frequencies
    ftfreqs = fftfreq(len(signal), sampling_scale)

    # Now fill in the matrix
    for n,s in enumerate(scales):
        psi_ft_bar = numpy.conjugate(wavelet.psi_ft(s * ftfreqs))
        W[n,:] = (s**.5) * ifft(signal_ft * psi_ft_bar)
    return W


def cwt_f(signal, freqs, Fs=1.0, wavelet = Morlet()):
    """Continuous wavelet transform -- frequencies version"""
    scales = wavelet.fc/freqs
    dt = 1./Fs
    return cwt_a(signal, scales, dt, wavelet)
```

3.A.23 Haar Wavelet (Figure 2.4.2)

```
import numpy as np
import pylab
import math

pylab.figure(figsize=(6,8))
pylab.subplots_adjust(bottom=.05, top=.95, hspace=.1)

# non-scaled and non-translated
pylab.subplot(2,1,1)
pylab.plot([-0.5, 0.0], [ 0.0,  0.0], 'b')
pylab.plot([ 0.0, 0.5], [ 1.0,  1.0], 'b')
pylab.plot([ 0.5, 1.0], [-1.0, -1.0], 'b')
pylab.plot([ 1.0, 1.5], [ 0.0,  0.0], 'b')
pylab.plot([ 0.0, 0.0], [ 0.0,  1.0], 'b--')
pylab.plot([ 0.5, 0.5], [-1.0,  1.0], 'b--')
pylab.plot([ 1.0, 1.0], [-1.0,  0.0], 'b--')
pylab.scatter([0.0, 0.5, 1.0], [1.0, -1.0,  0.0], s=30.,
              marker='o', edgecolor='b', facecolor='b')
pylab.scatter([0.0, 0.5, 1.0], [0.0,  1.0, -1.0], s=30.,
              marker='o', edgecolor='b', facecolor='w', zorder=3)

pylab.ylim([-2.25, 2.25])
pylab.xticks(np.linspace(0,1,3),
             [r'' for x in np.linspace(0,1,3)])
pylab.yticks(np.linspace(-2,2,5),
             [r'$%.1f$'%x for x in np.linspace(-2,2,5)])
pylab.title(r'$s=1.0, \; \tau=1.0$')

# scaled and translated
pylab.subplot(2,1,2)
pylab.plot([-0.5, 0.5], [ 0.0,  0.0], 'b')
pylab.plot([ 0.5, 0.75], [ math.sqrt(2),  math.sqrt(2)], 'b')
pylab.plot([ 0.75, 1.0], [-math.sqrt(2), -math.sqrt(2)], 'b')
pylab.plot([ 1.0, 1.5], [ 0.0,  0.0], 'b')
pylab.plot([ 0.5, 0.5], [ 0.0,  math.sqrt(2)], 'b--')
pylab.plot([ 0.75, 0.75], [-math.sqrt(2),  math.sqrt(2)], 'b--')
pylab.plot([ 1.0, 1.0], [-math.sqrt(2),  0.0], 'b--')
pylab.scatter([0.5, 0.75, 1.0], [math.sqrt(2), -math.sqrt(2),  0.0], s=30.,
              marker='o', edgecolor='b', facecolor='b')
pylab.scatter([0.5, 0.75, 1.0], [0.0,  math.sqrt(2), -math.sqrt(2)], s=30.,
              marker='o', edgecolor='b', facecolor='w', zorder=3)

pylab.ylim([-2.25, 2.25])
pylab.xticks(np.linspace(0,1,3),
             [r'$%.1f$'%x for x in np.linspace(0,1,3)])
pylab.yticks(np.linspace(-2,2,5),
             [r'$%.1f$'%x for x in np.linspace(-2,2,5)])
pylab.title(r'$s=2.0, \; \tau=0.5$')

# save and close
pylab.savefig('haar_simple_w_orthogonal.png', dpi=300)
pylab.close()
```

3.A.24 Haar Wavelet (Figure 2.4.3)

```
import sys
if sys.version_info[0] == 2:
    _strobj = basestring
elif sys.version_info[0] == 3:
    _strobj = str

import pylab
from numpy.random import random
from numpy import linspace, sin,pi,cos,array

import pywt

def flatten(x):
    """_flatten(sequence) -> list

    Returns a single, flat list which contains all elements retrieved
    from the sequence and all recursively contained sub-sequences
    (iterables).

    Examples:
    >>> [1, 2, [3,4], (5,6)]
    [1, 2, [3, 4], (5, 6)]
    >>> _flatten([[[1,2,3], (42,None)], [4,5], [6], 7, MyVector(8,9,10)])
    [1, 2, 3, 42, None, 4, 5, 6, 7, 8, 9, 10]"""

    result = []
    for el in x:
        #if isinstance(el, (list, tuple)):
        if hasattr(el, "__iter__") and not isinstance(el, _strobj):
            result.extend(flatten(el))
        else:
            result.append(el)
    return result

n=128
x=linspace(-6*pi,6*pi,n)
phases = random(8)

f = lambda x : .2*sin(.245*x+phases[0]*2*pi) + \
               .2*sin(.434*x+phases[1]*2*pi) + \
               .2*sin(.745*x+phases[2]*2*pi) + \
               .2*sin(1.53*x+phases[3]*2*pi) + \
               .2*sin(2.73*x+phases[4]*2*pi) + \
               .2*sin(3.30*x+phases[5]*2*pi) + \
               .2*sin(5.60*x+phases[6]*2*pi) + \
               .2*sin(9.10*x+phases[7]*2*pi)

y = f(x)

coeffs = pywt.wavedec(y, 'haar')
cA = coeffs[0]
cDs = coeffs[1:]

for c in coeffs:
    print(len(c))

pylab.figure(figsize=(9,12))
pylab.subplots_adjust(bottom=.05, top=.95, hspace=.6)
yticks = [-2,-1,0,1,2]
```

```
pylab.subplot(len(coeffs),1,1)
pylab.plot(linspace(0,1,n*10), f(linspace(-6*pi,6*pi,n*10)))
pylab.xlim([0,1])
pylab.xticks([0,.25,.5,.75,1], [r'$%.2f$'%x for x in [0,.25,.5,.75,1]])
pylab.ylim([-2.25, 2.25])
pylab.yticks(yticks, [r'$%i$'%x for x in yticks])
pylab.title(r'$f(x)$')

c_ = []
for i,cD in enumerate(coeffs):
    c_.append(cD)
    pylab.subplot(len(coeffs),1,i+1)
    if i > 0:
        r = pywt.waverec(c_,'haar')
        pylab.plot(flatten([[v,v] for v in range(len(r)+1)])[1:-1],
                   flatten([[v,v] for v in r]), 'b')
        pylab.xlim([0,len(r)])
        if len(r) == 2:
            xticks = [0,1,2]
        else:
            xticks = linspace(0,len(r),5)
        pylab.xticks(xticks, [r'$%i$'%x for x in xticks])
        pylab.ylim([-2.25, 2.25])
        pylab.yticks(yticks, [r'$%i$'%x for x in yticks])
        pylab.title(r'$\mathrm{Level} \, %i \, \mathrm{approximation}$'%(i))

pylab.savefig('haar_dwt.png')
pylab.close()
```

Chapter 4: Machine Learning

Physiological signals contain information relevant to cognitive workload (see Chapter 2). However, these signals contain large amounts of noise and are often non-linearly correlated with workload estimates. Much is unknown about how exactly physiological measures represent cognitive states. One possibility for identifying how they relate is to use *machine learning* techniques. Machine learning is a sub-domain of artificial intelligence related to developing and understanding algorithms that *learn* the characteristics in a dataset without having to explicitly program the characteristics in advance (Samuel, 1959). The aim here is to provide an introduction to machine learning to a naïve audience.

Machine learning algorithms can be generally classified into three categories based on the information provided to the algorithm during training. Supervised learners are given training data with the desired output. Reinforcement learners are given the training data, but instead of being given the correct output they are just given quantitative feedback regarding the efficacy of an output leading to a desired state. Lastly, unsupervised learners are just given data and are left to their own devices to figure out relationships in the data (clustering) or how actions map to changes in their environments (Russell & Norvig, 2003).

The type of learner used often depends on the particular problem and the available information. If an artificial agent is being constructed to replace a subject matter expertise and the correct output for a set of inputs is known then a supervised learner can be used. In other circumstances one might not know the correct output, but can determine whether a given output leads to making the problem better or making the problem worse (Russell & Norvig, 2003). For example, if a car driving agent crashes we can tell the agent they did something wrong without knowing exactly what they should have done differently. In contexts where one wishes to identify hidden structures in datasets or develop theoretical understanding of processes unsupervised learners may be appropriate.

One of the interesting lessons of machine learning has been that no known technique has shown to be superior in every circumstance. Much of the science of artificial intelligence is developing a theoretical understanding of constraints that yield *optimal* performance. Here it is important to define optimal as referring to finding the best solution and not how long it takes to find a solution (Russell & Norvig, 2003). As the theoretical underpinnings of various machine learning techniques improves the attributes that make a problem difficult will hopefully be able to formalize. While there is a trend to develop more general problem solvers that do not require tailoring to the specific circumstances, it is sometimes only possible to say technique *X* with parameters *A, B,* and *C* performed better than technique *Y* with parameters *D* and *E* on problem *Z* (Eiben & Smith, 2003).

In addition to performance, the time and space complexity of various techniques is also important to consider. The time complexity of an algorithm describes the number iterations required to solve a problem as a function of the number inputs. The space complexity of an algorithm describes the amount of information that needs to be stored while the algorithm is running. With some problems finding an okay solution in a reasonable amount of time is better than waiting a really long time for the best solution. Problems often exhibit tradeoffs between space and time complexity. Space complexity is often more of a limiting factor than time complexity. Giving an algorithm a magnitude order more time is often feasible. Waiting days for solutions rather than minutes may be acceptable whereas adding a magnitude order more memory to a system is cost prohibitive (Russell & Norvig, 2003).

When encountering new problems, it is likely not possible to know what technique will have the best performance. In such cases, the best route may be to implement and compare a variety of approaches. Some approaches are simple while others can be quite complicated. Some may have low time and space complexity, while others have high time, high space, or high time and

space complexity. Sections of this chapter will be devoted to various techniques. Before we get to the details it might be helpful to briefly introduce these techniques.

4.1 Linear discriminant analysis

One of the most basic, and more common, approaches is linear discriminate analysis (LDA). LDA weights and sums variables to form predictions, the magnitudes of the weights are indicative their importance in the model. As the name implies the classification is linear. More sophisticated algorithms may offer better performance but tradeoff in complexity and human interpretability. For example Algorithms like random-forests and symbolic regression build decision-trees that due to their size and complexity can be difficult for humans to interpret. Large branches of the decision trees may offer no real value to the outcome or evaluate to a constant. Secondly, any advantage offered by *non-linear classifiers* can only be identified through comparison to *linear classifiers*. To illustrate the problem consider the following function:

$$f(x, y) = 5.32 + .032x + 10\sin(8.6 * x) - 0.86y + 20\cos(3.2 * y)$$

Multiple regression could capture the intercept and linear contribution of *x* and *y* but will fail to capture the cyclical contributions of the trigonometric terms. Over a restricted domain (for arguments sake $x = (0,1), y = (0,1)$) this is especially problematic as the cyclical components contribute to the majority of the functions variability. In contrast a *symbolic regressor* with trigonometric non-terminals would be capable of capturing the full variability of the model.

A related approach is to try and make difficult problems linearly separable by projecting the predictor variables into higher dimensional spaces. Support Vector Machines (SVM) function as such. The support vectors are vectors that optimally segregate training case from one another. Decomposing a time signal into time-frequency components, also serves to project a predictor variables onto higher dimensional spaces. With SVM the basis is dependent on the criterion variable, with wavelets the basis is independent from the criterion.

4.2 Decision Trees

A decision tree is a simple yet highly effective algorithm for making a decision given categorical or continuous attributes or for classifying attributes. Decision trees can also be used to learn continuous functions. Their moniker is indicative of how various properties of attributes are represented as branches. A decision is formed by evaluating the root attribute and following the succeeding branches until a determination is made. One of the niceties of decision trees is that their representations can be quite natural for humans to follow (at least when they are not too big). Several methods exist to train decision trees. The method most commonly associated with decision trees is based on information theory. The root attribute is selected as the attribute that best segregates training cases and subsequent branches are selected in a similar manner. The information theoretic approach is considered *greedy* in that it does not look more than one step ahead and is consequently non-optimal. Even though there is no guarantee of optimal the heuristic has shown to be empirically useful. One of the weaknesses of decision trees is that they can be prone to overfitting with noisy datasets and datasets with several irrelevant attributes. Branches based on irrelevant attributes will be added to the tree to classify all the cases in the training data regardless of how well they generalize. Significance testing is often used to try and sort out meaningful attributes from irrelevant attributes (χ^2 pruning).

4.3 Adaboost

A decision tree generated using an information theoretic approach can be conceptualized as one of many possible hypotheses. The accuracy of an outcome rests solely on that single hypothesis. Ensemble learners attempt to develop multiple hypotheses that operate in conjunction with one another. The combined decision of a group of ensemble learners is often better than any single estimate. The notion is similar to “the wisdom of the crowds” phenomena first observed by Galton in 1907. Galton asked a group of nearly 800 experts and non-experts to guess the weight of a

slaughtered and dressed oxen. Although individual estimates were often far off the mark the median of the estimates was remarkably close to the actual weight (1198 lb. actual; 1207 lb. median) (Galton, 1907). Adaboost or *Adaptive Boosting* refers to a meta-algorithm for training an ensemble learner. The creators of Adaboost, (Freund & Schapire, 1999; Freund, 1995), present Adaboost using a rather intriguing horse racing analogy. Imagine a Gambler wishes to construct a computer program that can predict the fastest horse. The Gambler proceeds by asking his fellow Gamblers what sort of rules-of-thumb they use and comes up with a list of possible criteria like the horse that has recently won the most number of races, or the horse with the most favored odds. Each criteria by itself is *weak,* it may be better than chance but would hardly make the Gambler rich and some of the rules-of-thumb are more accurate than others. However if the individual rules-of-thumb are appropriately weighted and combined a fairly accurate classifier may be obtained. Adaboost is the component that figures out the appropriate weights.

4.4 Random Forests

Random forests are another type of ensemble learner developed by Brieman and Cutler (Brieman, 2001; Liaw, 2013) and based on work by Ho (1995; 1998)(1995; 1998) and Amit and Geman (1997). It incorporates bootstrap aggregation or *bagging* with ensemble learners each based on a random subset of features. Bagging resamples cases with replacement to reduce variance in training cases and reduce overfitting. The bagged samples are used to train a set of decision trees (hence the term random forest). Each tree is has a random subset of features. The number of features in this subset is treated as a tuning parameter and is typically small compared to the total number of features. As with decision trees the feature that best segregates the training set are choosen. The trees in the forest vote to form the overall classification.

Because biological signals often represent both linear and non-linear underlying processes symbolic regression may provide better performance compared to the traditional approach of linear discriminant analysis (LDA).

4.5 Genetic Programming

Playing off of the idiom "everything but the kitchen sink" genetic programming is perhaps most easily described as a "kitchen sink approach." It is capable of testing literally hundreds of thousands of models against one another in a matter of hours. This provides the benefit of not having to postulate relationships between variables based on theory. It is highly possible that it will create models which do not accurately model the process of human resource utilization, but still reliable predict workload at a structural level. Despite this caveat the models obtained from GP may also provide theoretical insight into cognitive workload. For example with early variations of ICA all pupil diameter wavelet components contributed equally to the final assessment of workload. The work by Nakayama and Shimizu (2002; 2004) and Nakayama & Katsukura (2007) suggests some of these components contain noise and not information related to mental workload. Genetic programming is capable of sorting out which wavelet components contain information relevant to workload and which components reflect noise. Wavelet coefficients that are used more often are likely to be more informative to cognitive workload.

This chapter is intended to provide a conceptual overview of machine learning to a general audience. Specifics relevant to the exact implementations can be found in the experiment write-ups and Appendix A.

Appendix 4.A Genetic Programming Implementation Details

4.A.1 Terminals and Non-Terminals

Symbolic regression represents programs as tree structures. The branches or non-terminals belong to a set of mathematical operators. Here the non-terminal set consisted of addition (of 2 *arity*), subtraction (2), multiplication (2), division (2), power (2), absolute value (2), square root (1), exponential (1), sin (1), cos (1), tan (1), and if-less-than (4). All operators were "protected" by checking to see if the operator overflowed, errored (divide by zero), or returned something other than a number. An overflow refers to when a number exceeds the range in which it can be represented by the IEEE 32 bit floating point standard. When any of the proceeding occurred the individual is given a fitness value of 1, which is the worst possible fitness value an individual can take with Keizer's scaled squared error. This pressures the model to select individuals which at least evaluate, but has the disadvantage of eliminating more solutions that might be reasonable within restricted parameter spaces.

The leaves of the tree structure are refered to as terminals. Here the terminal set consisted of scalars of pi, -100, -10, -1, 0, 1, 10, and 100; as well as the vectors holding raw skin conductance, raw pupil diameter, the 9 skin conductance wavelet coefficients, and the 9 pupil diameter coefficients.

4.A.2 ALPS-SS Algorithm

The model can be grossly described as a steady-state age-layered population (ALPS) genetic program (Hornby, 2009). An age layered population sacrifices speed for "robustness." The layers only allow individuals to compete with other individuals their layer and the layer younger than themselves. When individuals reach their maximum age they must have a better fitness than one of the individuals in the layer above them to move up. If they don't then they are deleted making room for individuals to move up. Ideally, the "diversity" of the population stays high which allows for ALPS to search out multiple optima without quickly converging. Individuals in the bottom (youngest) layer are replaced by freshly generated random individuals whose age starts at 1. New individuals generated through crossover and/or mutation take the age of their youngest parent. The top layer has no maximum age and individuals stay until they are replaced by a better individual. Hornby suggests using a Fibonacci progression for defining maximum ages as well as the formula for calculating age. Hornby typically uses ten layers with 70 to 100 individuals.

Implementing a steady-state ALPS surprising isn't that much more difficult than implementing a vanilla GP algorithm. The pseudo-code below illustrates the basic flow of the algorithm. As the code suggests most of the work is actually accomplished elsewhere in the tree objects which support symbolic regression.

function TRYMOVEUP(*layers, j, k*)
description: Tries to move the individual in *layers*[*j*][*k*] with an individual with poorer fitness in *layers*[*j+1*] if such an individual exists. Either way it deletes *layers*[*j*][*k*] before returning

function ALPS-SS()
local variables: *layers,* a list of lists containing the individuals in each age layer
num_layers, an integer specifying the number of layers in *layers*
imax, the number of iterations to evolve solution

layers ← initialize each layer with random individuals

i ← 0
while *i* < *imax* **do**

for *j* = 0**;** *j* < *num_layers*; j++ **do**
parent$_1$ ← using tournament select from *layers*[j] (and *layers*[j-1] **if** *j* != 0)
parent$_2$ ← using tournament select from *layers*[j] (and *layers*[j-1] **if** *j* != 0)

```
        new1, new2 ← CROSSOVER(parent1, parent2)
        new1.MUTATE()
        new2.MUTATE()
        new1. SETFITNESS()
        new2.SETFITNESS()
        layers[j].REPLACEWORSTIND(new1)
        layers[j].REPLACEWORSTIND(new2)

        i ← i + 1
        for k = 0; k < LENGTH(layers[j]); k++ do
            layers[j][k].UPDATEAGE()

            if layers[j][k].TOOOLD() do
                TRYMOVEUP(j, k)

return the best individual in layers
```

4.A.3 Initial Population

The GP utilized a technique known as Age Layered Populations to increase the robustness of search and as a preventative measure against premature convergence (Hornby, 2009). Each initial population consisted of ten layers of 100 random "full" trees of depths of randomly selected depths of 3, 4, and 5. The term full describes trees in which all the terminals are at the maximum depth of the tree. They are built using the following class method:

```
def __full__(self,node, N):
    if node.depth==N:
        # randomly pick terminal
        node.add_child(terminals[randint(len(terminals))])
    else:
        # randomly pick non-terminal
        operator,arity = nonterminals[randint(len(nonterminals))]
        node.add_child(operator)
        for i in range(arity):
            self.__full__(node.children[-1],N)
```

4.A.4 Scaled Mean Squared Calculation

The target values are normalized such that $\sum t_i^2 = N$. Since E_s is upper bounded by $\frac{1}{N}\sum t_i^2$ (Keizer, 2004)after normalizing the new upper bound of the fitness measure becomes 1. The Because the upper bound is well defined it makes adding parsimony pressure much simpler. A parsimony penalty of .0005 per node was used for all simulations described in Experiment 3.

```
def set_fitness(self, parsimony_penalty):
        Y=self.evaluate()     # The predicted values
        T=self.Targets        # The unscaled target values
        N=self.N              # The length of Y, T, and t

        t=deepcopy(T)         # The scaled target values
        t-=numpy.mean(t)      # set mean(t) = 0
        t=sqrt(N)*(t)/norm(t) # set sum(t**2) = N

        try:
            if numpy.isnan(Y).any() or numpy.isinf(Y).any():
                raise Exception

            y = Y-numpy.mean(Y)
            v = 1./N*numpy.sum(y**2)
            w = 1./math.sqrt(v)
            r = 1./N*numpy.sum(y*w*t)

            self.fitness = 1.-r**2+len(self)*parsimony_penalty

        except:
            self.fitness=1.+j*len(self)*parsimony_penalty
```

4.A.5 Crossover

My crossover function applies "standard" node swapping using the 90/10 rule. The 90/10 rule dictates that 90% of the time a non-terminal will be chosen for crossover and only 10% of the time a terminal will be chosen. This is accomplished by building a list of all the terminal nodes and a list of all the non-terminal nodes. From there selecting a random non-terminal with a probability of 90% or a random terminal with a probability of 10% is straightforward. The "try" and "except" statements below are for the case when the tree contains no non-terminals which can occasionally occur with the grow mode.

```
def crossover(self,other):
    # Make copies of self, and other
    new1,new2=deepcopy(self),deepcopy(other)

    # Select nodes for crossover using 90 / 10 rule
    # each individual has a list of terminals and a
    # list of non-terminals
    try:
        if random() < .9 :
            node1 = new1.nonterms[randint(len(new1.nonterms))]
        else:
            node1 = new1.terms[randint(len(new1.terms))]
    except:
        node1 = new1.root

    try:
        if random() < .9:
            node2 = new2.nonterms[randint(len(new2.nonterms))]
        else:
            node2 = new2.terms[randint(len(new2.terms))]
    except:
        node2 = new2.root

    # Swap nodes
    tmp1,tmp2 = deepcopy(node1),deepcopy(node2)

    node1.data,node2.data = tmp2.data,tmp1.data
    node1.children,node2.children = tmp2.children,tmp1.children

    # rebuld terms and nonterms node lists
    new1.build_typelists()
    new2.build_typelists()

    # return new individuals
    return new1,new2
```

4.A.6 Mutation

A simple point mutation is performed by first selecting a random node. If the randomly chosen node is a constant terminal (a scalar) mutation is performed with probability of .1 (10%). If the randomly chosen node is a non-terminal or a non-constant terminal (one of the data columns) then mutation is performed with a probability of .01 (1%).

If a constant is selected for mutation the constant is mutated by a maximum of +/- 5% of its absolute value. If a non-constant terminal is selected it is swapped with different randomly selected non-constant terminal. If a non-terminal is chosen it is swapped with a different randomly selected non-terminal with the same arity.

Since iflte (if-less-than-or-equal-to) is the only non-terminal with an arity of four iflte nodes are not selected for mutation. The mutation algorithm will reselect until a node not containing iflte is chosen. Future analyses will experiment with using other non-terminals of 4-arity (if greater than, if equal to, etc.).

```
    def mutate(self, c_rate=.1, o_rate=.01):
        # c_rate defines mutation rate for constants
        # o_rate defines mutation rate for all other terminals
        # and non-terminals

        # Pick random node for crossover
        nodelist = self.terms+self.nonterms
        node = nodelist[randint(len(nodelist))]

        while node.data=='iflte':
            node = nodelist[randint(len(nodelist))]

        if isfloat(node.data) and random() < c_rate:
            # node is a constant scalar terminal
            node.data=str(float(node.data)+random()-.5)

        elif random() < o_rate:
            arity1 = [op for (op,arity) in nonterminals if arity==1]
            arity2 = [op for (op,arity) in nonterminals if arity==2]

            if node.is_terminal():
                # node is non-constant terminal
                node.data = ['Xs[%i]'%i for i,X in enumerate(Xs)
                             if 'Xs[%i]'%i != node.data][randint(len(Xs)-1)]

            elif node.data in arity2:
                # node is a non-terminal of arity 2
                node.data = [op for op in arity2
```

```
                if op != node.data][randint(len(arity2)-1)]

        elif node.data in arity1:
            # node is a non-terminal of arity 1
            node.data = [op for op in arity1
                         if op != node.data][randint(len(arity1)-1)]
```

Chapter 5: Empirical Evaluation

Having now reviewed the finer details of spectral analysis and genetic programming I now turn to the application of these techniques to assessing mental workload from physiological measures. Here I present the results of series of seven experiments that tested the application of novel analytical tools for using physiological measures to assess mental workload in a variety of tasks. Experiment 1 simulated a process control task in which faults were introduced to induce changes in task difficulty. Experiment 2 examined a two-dimensional visual-manual compensatory tracking task which manipulated task difficulty by instantaneously and repeatedly reversing the directional mappings of the control. Neither of these experiments produced conclusive results, however I include them here illustrate how the experimental manipulations for subsequent experiments evolved and to provide the most accurate representation of the work. As skeptical evaluators it is appropriate to ask "Are the results obtained here a true reflection of the efficacy of workload measures, or are they an illusion based on the right combination of experimental parameters, participants, the evaluation methodology, and chance?" By providing results from the earliest, and in hindsight, flawed empirical investigations I hope to help the reader answer this question for themselves and to learn from my mistakes. Readers are welcome, and perhaps even encouraged, to start with Experiment 3: Pursuit Tracking (Normal vs. Rotated). Experiments 1 and 2 are less central to this dissertation.

The tasks manipulations used in Experiments 3-7 benefitted from the lessons learned in Experiments 1-2. Experiment 3 used a two-dimensional tracking task similar to Experiment 2, but manipulated task difficulty by instantaneously rotating the control mappings 90 degrees, rather than reversing them. The results clearly demonstrated that analysis of pupil diameter and skin conductance with genetic programming techniques can produce sensitive workload indicators for large, instantaneous changes in task difficulty. To determine whether these results generalize to less extreme and slower changes in task difficulty, Experiments 5-7 assessed these indicators in the

context of a critical instability tracking task (McDonnell & Jex, 1967; McRuer & Graham, 1965) whose difficulty could be varied continuously over time. Experiments 5 and 6 aimed to validate the effect of my manipulation of task difficulty on mental workload, measured subjectively (Exp. 5) and using a secondary task (Exp. 6). These experiments found a clear relationship between task difficulty and mental workload, assessed either subjectively or with a secondary task. Based on this validation, Experiment 7 used the same critical instability task to examine whether physiological indicators are sensitive to subtle changes in task difficulty, and whether they might also serve as reliable leading indicators of poor task performance.

5.1 Experiment 1: Process Control Simulator (DURESSJ)

My first attempt at assessing workload from pupil diameter and skin conductance used a process-control simulation requiring the participants to monitor and manipulate a system of pumps, valves, and heaters to met changing flow and temperature demands (Lew, Dyre, Werner, Wotring, & Tran, 2008). The process control task was chosen for its ecological validity to control monitoring tasks. At the time I was aware of Marshall's work with wavelet decomposition, and analyzed the time frequency domain using short-time Fourier transformations with rectangular windowing.

5.1.1 *Method*

5.1.1.1 Participants. Four university students participated in this experiment. All had normal or corrected to normal Snellen visual acuity (20/30 or better). Participant 1, one of the experimenters, had detailed knowledge of the experiments hypothesis and methods but did not know when the faults occured, while the others had only limited knowledge of the experimental measures and manipulations. All participants were ethically treated in accordance with experimental protocols approved by the University of Idaho's Human Assurance Committee (see Appendices 5.1.A – 5.1.C).

5.1.1.2 Stimuli and Apparatus. Participants operated the DURESS process-control simulation (Cosentino & Ross, 1999; Vincente & Pawlak, 1994). The simulator modeled a system of two pumps, eight valves, two aquifers, and two heaters. Water flowed from a common source through the pumps, valves, aquifers, and heaters to two outputs (see **Figure 5.1.1**The flow and temperature of these outputs changed throughout the duration of the trial. The participants were tasked with matching changing flow and temperature demands. To manipulate workload, each participant was given an identical set of plant failures and events at predefined instances (see Table 5.1.1).

The simulation was displayed on a 60 inch rear-projection monitor at a spatial resolution of 1280 x 1024 and a temporal resolution of 60 Hz with a viewing angle of 45° X 33.75°. The display was presented in a darkened room with the participant sitting 1.5m from the display. The participants used a standard optical mouse with their right hand to control the elements of the DURESS simulation.

A model ASL5000 head-mounted eye/head tracker was used to measure gaze direction, pupil diameter, and blink rate at 60 Hz (see Figure 5.1.2). Skin conductance was measured at a temporal resolution of 256 Hz with a Thought Technologies ProComp5 Infiniti encoder using two finger-mounted sensors placed around the index and ring fingers of the left hand. Participants were instructed to leave their left hand in a stationary position during the course of the trial.

5.1.1.3 Procedure. Each participant received a short tutorial explaining goals, controls, and caveats of the DURESS program. Participants were then given a 10-minute DURESS practice session with coaching from the experimenter without the eye/head tracking or SC monitor attached.

After training, the eye/head tracker and GSR devices were mounted and calibrated to the individual. The participants then operated and experienced the DURESS fault trial which ended after 15 minutes (900 s). All data analysis and visualization was performed using Python, with the

Figure 5.1.1 *User Interface for the DURESS Simulator*
The participant is tasked with matching flow and temperature demands for two outputs as well as reservoir levels. PA and PB are pumps, VA, VB, VA1, VA2, VB1, and VB2 control the flow of water to the aquifers. Participants are instructed they should not let the aquifers run dry or overflow. VO1 and VO2 are used to set the flow to demand. H1 and H2 set the heater levels for the aquifers. The green regions show the operator the requested flow and temperature demand.

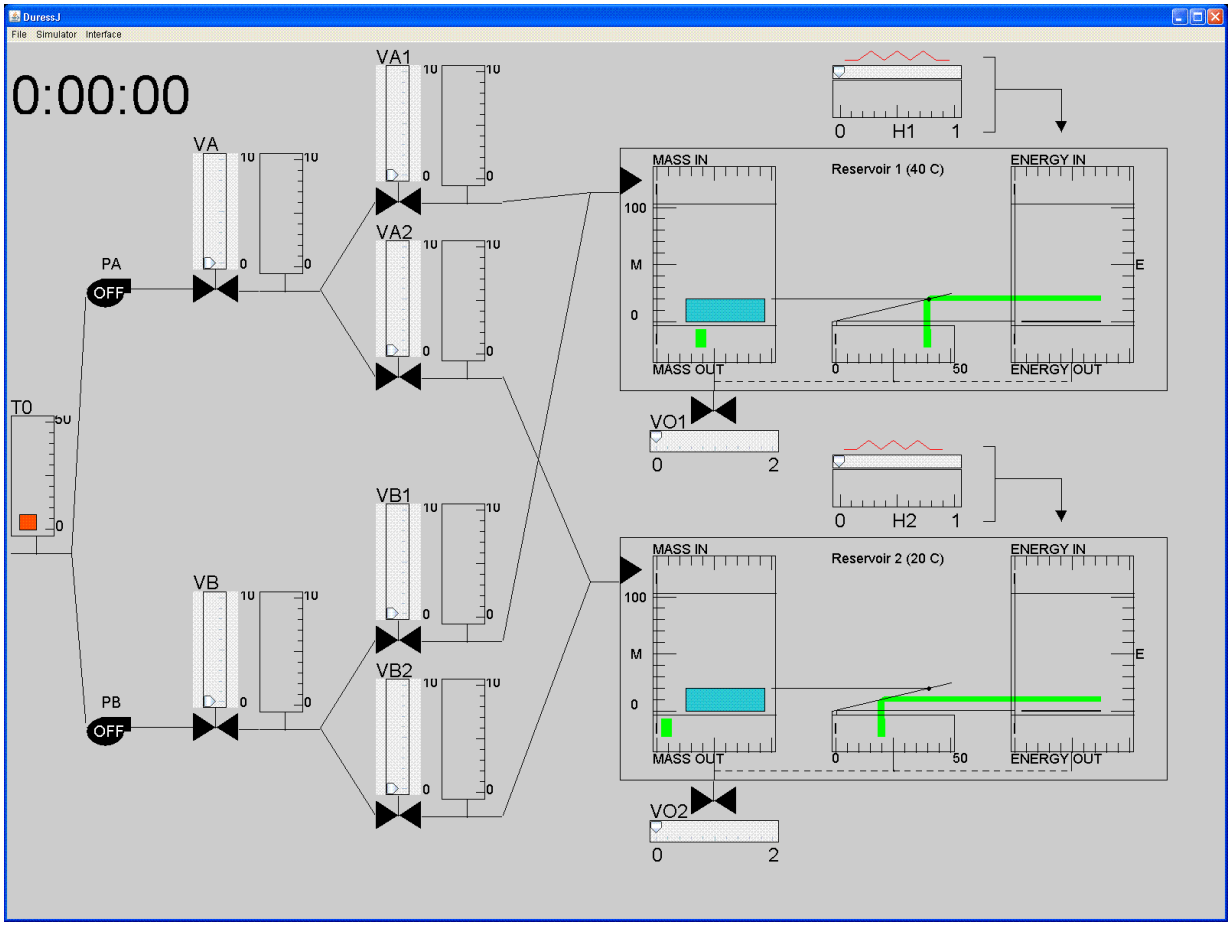


Figure 5.1.2 *Participants wore the head-mounted eye tracker shown below. Due to the weight and design of the unit it becomes moderately uncomfortable after a period of about 20 minutes. For this reason I tried to design all the experiments so that participants would not have to wear the tracker for longer than 20 minutes.*

NumPy, matplotlib, and SciPy packages. Skin conductance, pupil diameter, and duress measures where linearly interpolated to a sampling rate of 60 Hz (the slowest of original sampling rates). Additionally, blinks were removed from the pupil diameter data by holding the last non-blink value during the duration of the blink, and data below three standard deviations of the mean where deemed as unreliable due to measurement error and removed (less than 0.1% of data, typically indicate a partially blink). More complex treatments of blinks exist but they typically do not significantly alter the power content below .5 Hz according to Nakayama and Shimizu (2004).

Examining the data in the frequency domain provides a mechanism by which changes in physiological responses to events in a complex task can be estimated on all time scales at once, in effect a "brute force" approach to the problem of determining the appropriate time scale. To do this I first transform short samples of the time-series measures into the frequency domain. The time-series "windows" defining these samples may overlap to provide a more continuous measure of the change in the spectral characteristics of the measures.

SC and pupil diameter were decomposed using short time Fourier transform (STFT) with time windows of approximately 34 s (2048 samples) with each window overlapping its neighbor by 50%. For each time slice the power spectrum was estimated using a fast Fourier transform (FFT) algorithm. Finally, spectrograms (plots depicting power amplitudes with color, time on the x-axis and frequency on the y-axis) were created. The spectrographs make low frequency variations in the time domain easily distinguishable from noise. This treatment is especially useful for long trial durations (15 min.) where distinguishing nuances in time domain plots is not possible due to how the time axis gets compressed. For this experiment no inferential statistics were conducted. The plots were more or less visually correlated with the DURESS faults and performance error.

5.1.2 *Results.*

Figures 5.1.3.- 5.1.7. depict the physiological measures and performance of the four participants over time. For each the top of panel shows raw SC, and the third panel shows

Figure 5.1.3 *Experiment 1, Participant 1. Physiological* measures *and performance of the four participants over time. For each the top of panel shows raw SC, and the third panel shows interpolated pupil diameter, and the vertical magenta lines of the bottom panel show when participants made changes to the DURESS simulation.*

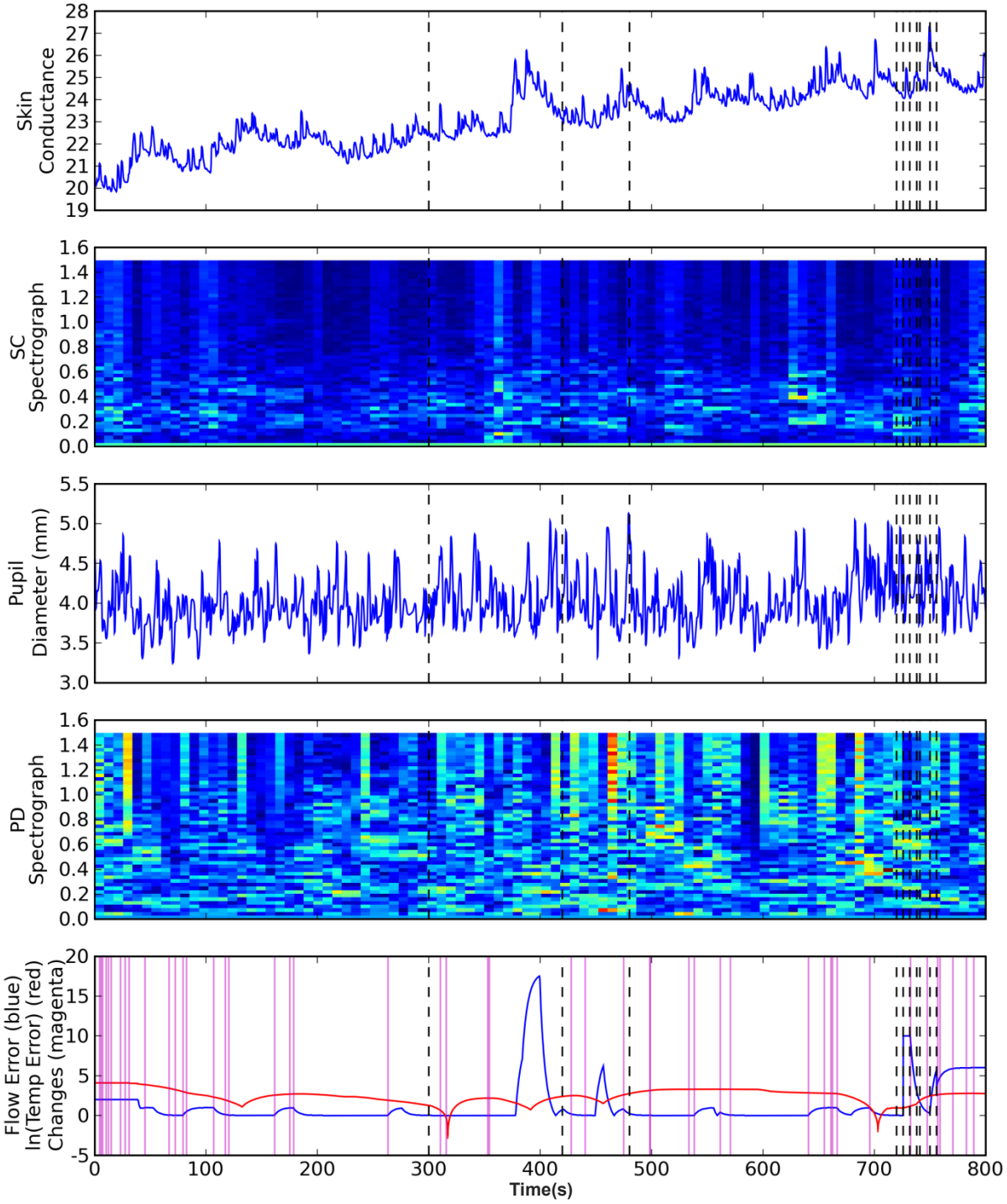


Figure 5.1.4 *Experiment 1, Participant 2.*

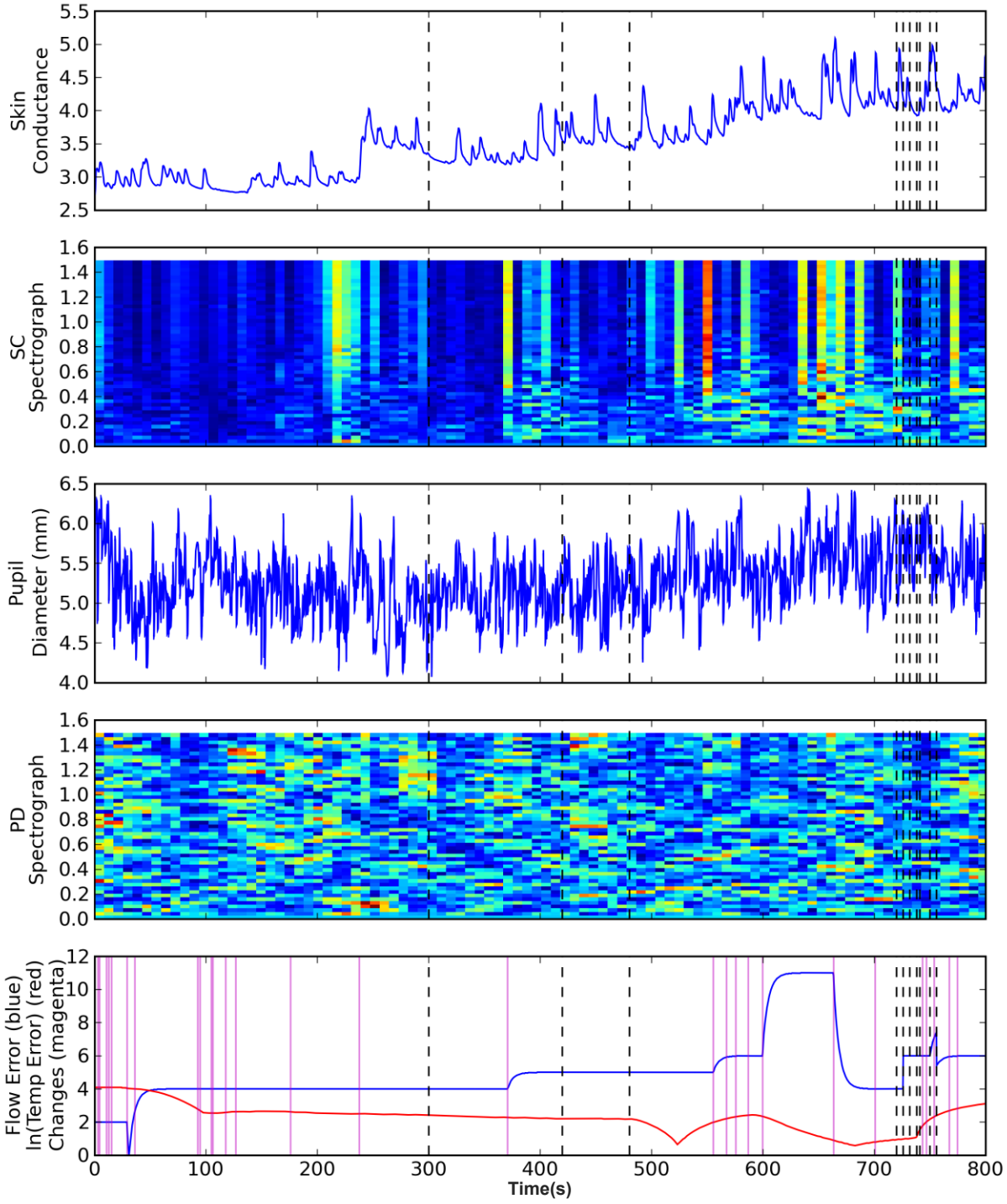


Figure 5.1.5 *Experiment 1, Participant 3.*

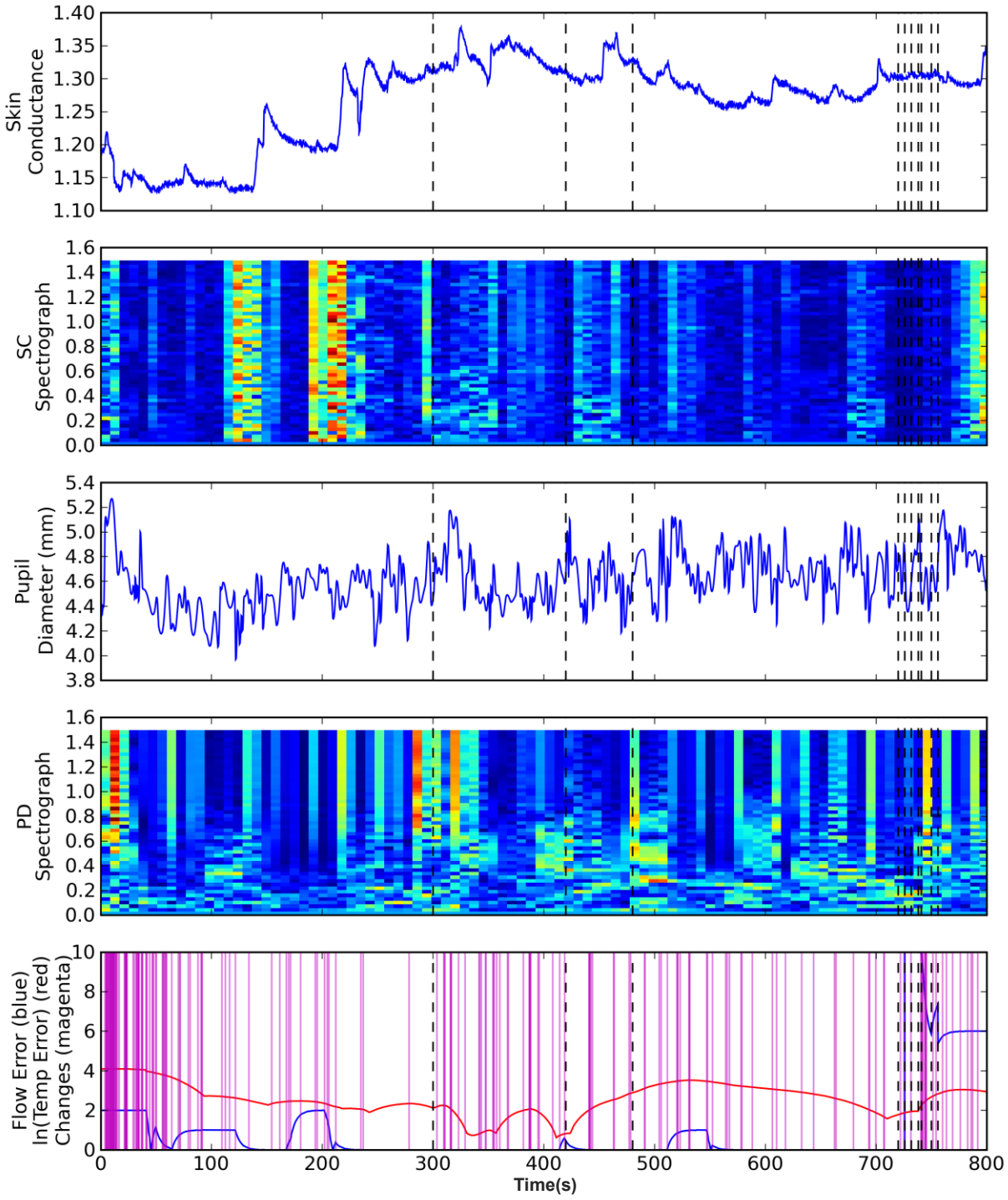
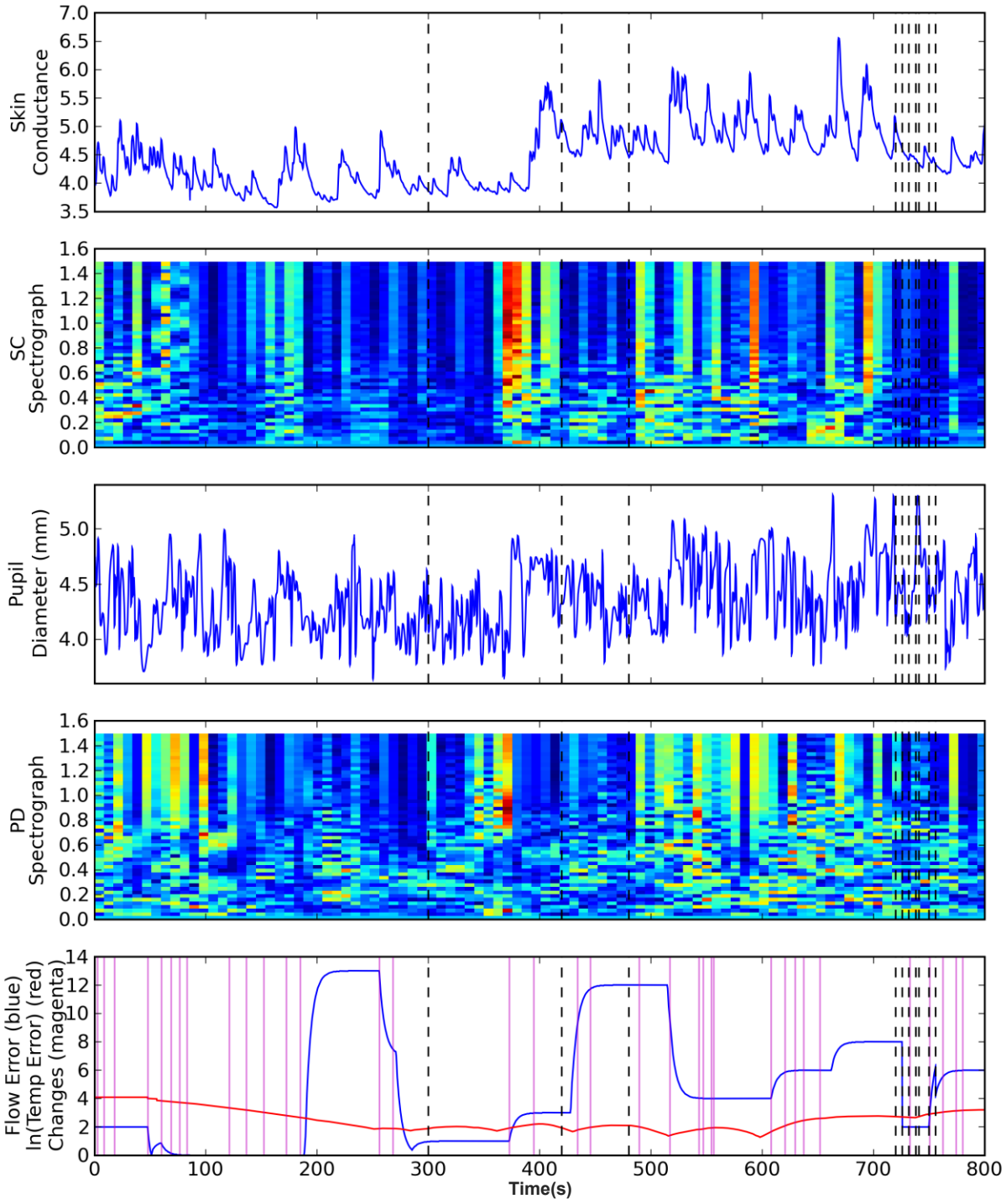


Figure 5.1.6 *Experiment 1, Participant 4.*



interpolated pupil diameter, and the vertical magenta lines of the bottom panel show when participants made changes to the DURESS simulation. These interactions include setting heat levels or valve flows to new levels, and turning a pump on or off. The blue trend in the bottom panel depicts the average in actual out flow and demand. The red trend reflects the natural log of the actual temperature of the outflow from the demanded temperature. In all panels the hashed vertical black lines show when the pre-specified DURESS faults occurred.

5.1.3 *Conclusions and Discussion.*

Inspection of the spectrographs reveals that visually-obvious increases in power across a fairly broad range of frequency occur commonly at the beginning of the trial while the participant is actively adjusting the simulator to achieve steady state and just after the DURESS fault events. In contrast, during time-sequences with relatively little activity, the spectrograms appear dark blue, indicating very little change in the physiological measures. Across several participants "hot spots" appeared shortly after DURESS fault events which I interpret as promising indications that these measures could prove to be a reliable and sensitive real-time indicators of mental workload and stress when analyzed using STFT.

When interpreting the spectrographs it is important to note that fault events were often not noticed by participants. Faults did not trigger visual or auditory alarms; only the unresponsiveness of the flow and temperature states to system changes indicated system failures. Events potentially went unnoticed for several seconds or even minutes. I know when the events occur but not when the participants become cognitively aware of them. The spectrographs (2nd and 4th panels from the top) represent time (in seconds) on the x-axis and frequency (in Hz) on the y-axis. Color is used to represent the power (in decibels) for each particular time and frequency sample. Areas of high amplitude changes in the measures appear red, while areas of low amplitude appear dark blue. Intermediate amplitudes are represented along dichromatic continuum from red to blue.

Despite some compelling evidence that physiological measures do convey some degree of cognitive workload no truly reliable trends were found in this dataset. Sometimes it appears to have spectral power in the right places, but there are equally many (if not more) occasions where increased power should be visible (if the hypothesis is correct) and it is not visibly higher. As previously mentioned one of the major difficulties in interpreting this dataset is that participants did not always notice the fault events.. Faults did not trigger visual or auditory alarms; only the unresponsiveness of the flow and temperature states to system changes indicated system failures. Experiment 2 attempted to remedy this problem by using a task where changes in workload were more salient.

Table 5.1.1

DURESS Fault Events

Time (s)	*Event Description*
300	Flow rate in valve VA2 changes
420	Demand change for upper reservoir
480	Flow rate in valve VB1 changes
720	Demand changes for upper reservoir
726	Demand changes for lower reservoir
732	Flow rate in VB2 changes
738	Inflow of water to upper reservoir
741	Outflow of water to lower reservoir
750	Output valve outflow increases and inflow temperature changes
756	Demand changes for lower reservoir and flow rate in valve VA2 changes

Appendix 5.1.A Consent Form

CONSENT FORM

Idaho Visual Performance Laboratory
Department of Psychology and Communication Studies
College of Liberal Arts and Social Sciences
University of Idaho
Control of speed during altitude changes

During this experiment you will be presented a display in a virtual environment. Various parameters of this display will be manipulated to examine stress and mental workload. In this experiment you will be asked to control movement in the virtual world using an input device such as a joystick.

The data you provide will be kept anonymous. There will be absolutely no link between your identity and your particular set of data.

Your participation will help increase knowledge of stress and mental workload. Subsequent to your participation the purpose and methods of the study will be described to you and questions about the study will be answered. It is our sincere hope that you will learn something interesting about your visual system from this debriefing.

The risks in this study are minimal, however displays simulating movement may on rare occasion cause motion sickness or eye fatigue in sensitive individuals. If at any time during the experiment you feel eye fatigue, dizziness, headache or nausea, please let the experimenter know immediately so that you can take a break before these symptoms become too intense. We endeavor to design our displays to minimize eye fatigue and motion sickness, and schedule periodic breaks to further reduce their occurrence. As a result, these phenomena have not been a common problem in previous similar studies.

Your participation will require **1** session of approximately **60** minutes. You may withdraw from this study at anytime without penalty. You will receive partial credit for your time spent. However, please be aware that your data is useful to us only if you complete the experiment in its entirety.
This research project has been approved by the University of Idaho Human Assurance Committee. As such, new information developed during the course of the research which may relate to your willingness to continue participation will be provided to you.

Thank you for your participation

Signature_____ Date _____

If you have further questions or encounter problems please contact:
Dr. Brian P. Dyre
(208) 885-6927
bdyre@uidaho.edu

Appendix 5.1.B Debreifing Form

Debriefing Form

Department of Psychology and Communication Studies

College of Letters, Arts, and Social Sciences

INL Physiological Predictors of Workload

Experiment 1

Participant:____

Date:_____

1. Did you move your left hand during the course of the trial while the GSR was still hooked up?

2. How often do you play video games?

    a. What is your video game skill? (Bad, okay or good)

3. Did you notice that the controls changed part way through the trial?

    a. How many times?

4. How difficult was the task when you first started? (1-10)

5. How difficult were the normal vs. reversed controls? (1-10)

6. How long did it take you to feel confident you were performing well at this task?

    a. Normal mappings

    b. Reversed mappings

7. How uncomfortable was the eye-tracker when you first started? (1-10)

8. How uncomfortable was the eye-tracker when you finished? (1-10)

9. Did you find the eye-tracker distracting from the task at hand?

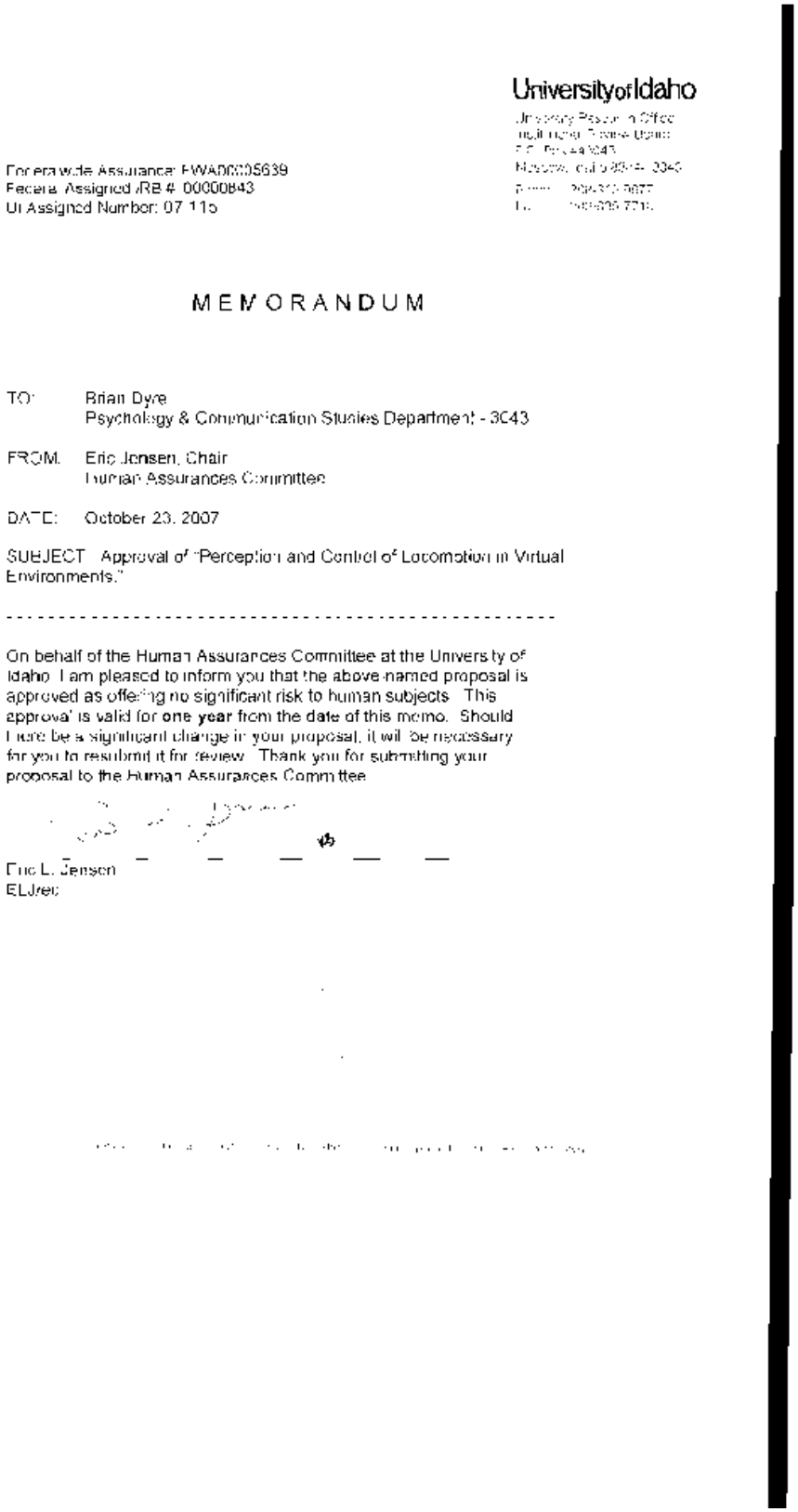
10. Do you think that fatigue played a role in your performance?

    a. How about fatigue from the eye-tracker?

11. Did you have any eye-strain?

Any additional comments

Appendix 5.1.C Human Assurances Approval



University of Idaho

Research Office
Moscow, ID 83844-3040
Phone: 208-885-6162
Fax: 208-885-5751

Federalwide Assurance: FWA00005639
Federal Assigned IRB #: 00000843
UI Assigned Number: 07-115

MEMORANDUM

TO: Brian Dyre
Psychology & Communication Studies Department - 3C43

FROM: Eric Jensen, Chair
Human Assurances Committee

DATE: October 23, 2007

SUBJECT: Approval of "Perception and Control of Locomotion in Virtual Environments."

On behalf of the Human Assurances Committee at the University of Idaho, I am pleased to inform you that the above-named proposal is approved as offering no significant risk to human subjects. This approval is valid for **one year** from the date of this memo. Should there be a significant change in your proposal, it will be necessary for you to resubmit it for review. Thank you for submitting your proposal to the Human Assurances Committee.

Eric L. Jensen
ELJ/ec

5.2 Experiment 2: Pursuit Tracking (Normal vs. Reversed)

Our second attempt (Lew, Dyre, Soule, Ragsdale, & Werner, 2010) tried to provide more obvious workload changes by using a dual-axis tracking task instead of a process-control simulation. In this task participants used a joystick to control a cursor and follow a dot moving in a pseudo-random fashion. To manipulate task difficulty the control mappings were abruptly reversed during half way through the experimental trial. The data analysis still used STFT but used Principle Component Analysis (PCA) and discriminant analysis to provide an objective means of discriminating workload. For the most part this experiment found inconclusive results due to an unforeseen problem with my manipulation of task difficulty, which though having an immediate and obvious change to system control, did not reliably affect task difficulty: not all participants found the reverse mappings more difficult, and in fact some found the "reverse" mappings to be easier than the "normal" mappings I include the experiment despite these inconclusive results because it provides context for my rationale in future experiments.

5.2.1 *Method*

5.2.1.1 Participants. Six university students participated in this experiment and were compensated with course credit. All had normal or corrected to normal Snellen visual acuity (20/30 or better). All participants were naïve to the hypotheses of the experiment. All participants were ethically treated in accordance with experimental protocols approved by the University of Idaho's Human Assurance Committee (see Appendices 5.2.A – 5.2.C).

5.2.1.2 Stimuli and Apparatus. Participants tracked a balanced dot on a grey background moving in a pseudo random fashion with a black cursor. A balanced dot used as precaution against having pupil dilations due to luminance changes. The balanced dot maintains equal-luminance with the background by having a small dot with high luminance surrounded by a larger dot with low luminance (See Figure 5.2.1). The dot's movement was determined by two sum-of-sine disturbances defining the x and y coordinate locations relative to the operator's field of view. The

frequencies and amplitudes were set such that the full balanced dot would always stay within the screen, and tracking performance would be fairly good for most participants. The horizontal disturbance had prime frequencies of 0.027, 0.067, 0.125, 0.154, 0.176 Hz and respective displacement amplitudes of 11.2, 4.513, 2.419, 1.964, 1.718 degrees of visual angle. The vertical disturbance had prime frequencies of 0.039, 0.079, 0.131, 0.167, 0.197 Hz and respective displacement amplitudes of 7.754, 3.828, 2.308, 1.811, 1.535 degrees. Both the horizontal and vertical disturbances had random phases.

Participants wore the head mounted eye tracker and skin conductance apparatus described in the previous in Experiment 1. Participants also viewed the same 60-inch rear projection monitor described in Experiment 1. There were two differences in apparatus from Experiment 1: participants used a gaming joystick instead of a mouse, and the displays were generated with ViEWER (Dyre, Grimes, & Lew, 2009) instead of DURESS.

5.2.1.3 Procedure. Participants controlled the cursor using a right-hand joystick with first order control dynamics and a gain of 25° per second at maximum deflection. For the first ten minutes of the experiment the control mappings were "normal." Normal meant moving the joystick forward moved the cursor up, moving the joystick backward moved the cursor down, moving the joystick right moved the cursor right, and moving the joystick left moved the cursor left. At the ten minute mark the control mappings "reversed" such that moving the joystick forward now moved the cursor down, moving the joystick left now moved the cursor right, and so forth (see Figure 5.2.1). At the fifteen minute mark the control mappings reverted back to "normal" for the remaining five minutes of the trial. In total participants tracked the balanced dot for 20 minutes. The abrupt changes were hypothesized to elicit transient physiological responses, and the "reversed" mappings were hypothesized to cause lower performance (higher tracking error defined by the Euclidean distance between the center of the balanced dot and the cursor) and changes to physiological indicators reflecting increased workload.

Figure 5.2.1 *Screenshot of tracking task. Participants used a joystick to control the black cursor (crosshairs) shown below. The "balanced dot" shown to the left of the cursor moved about in a pseudo-random fashion.*

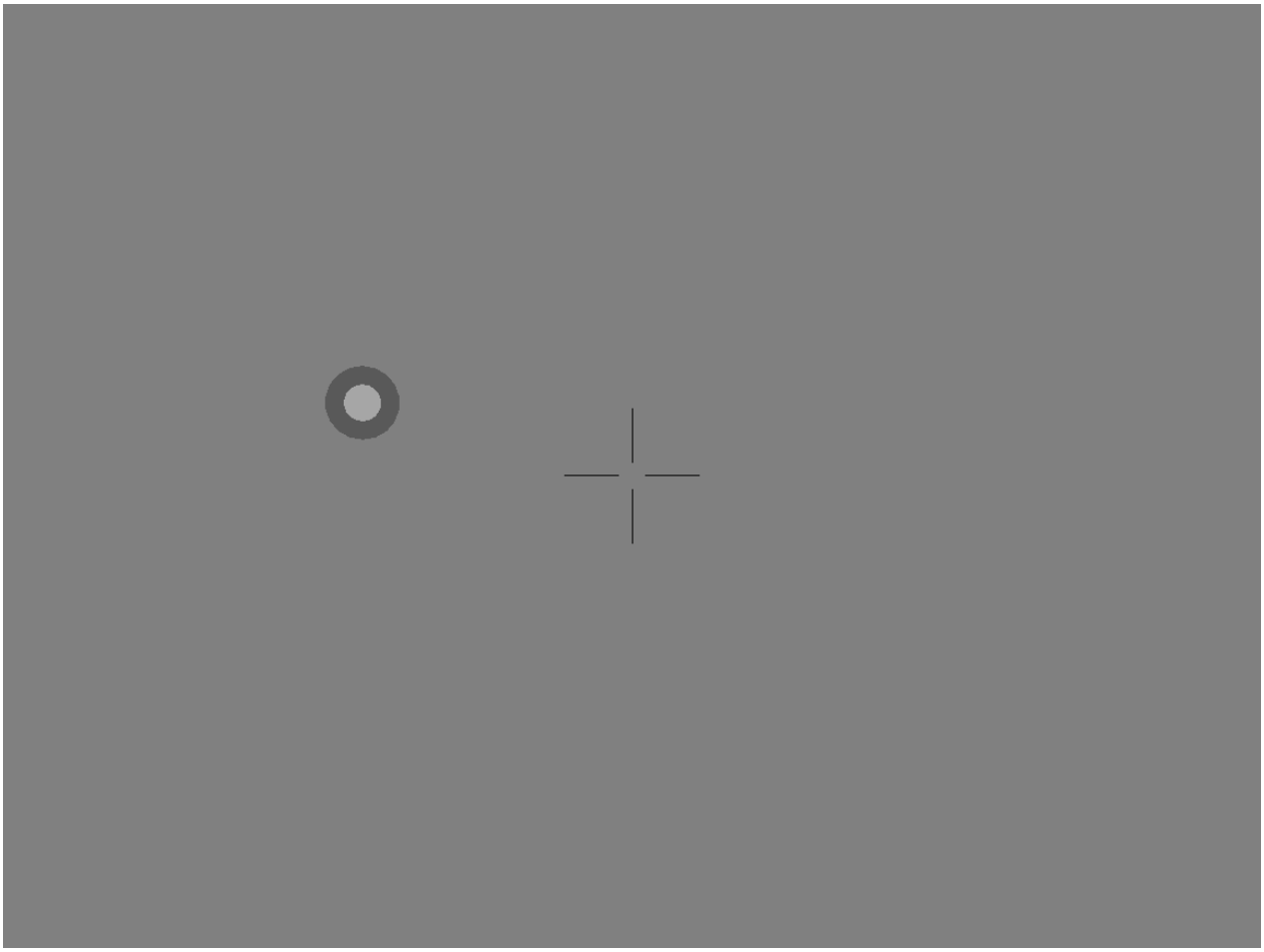
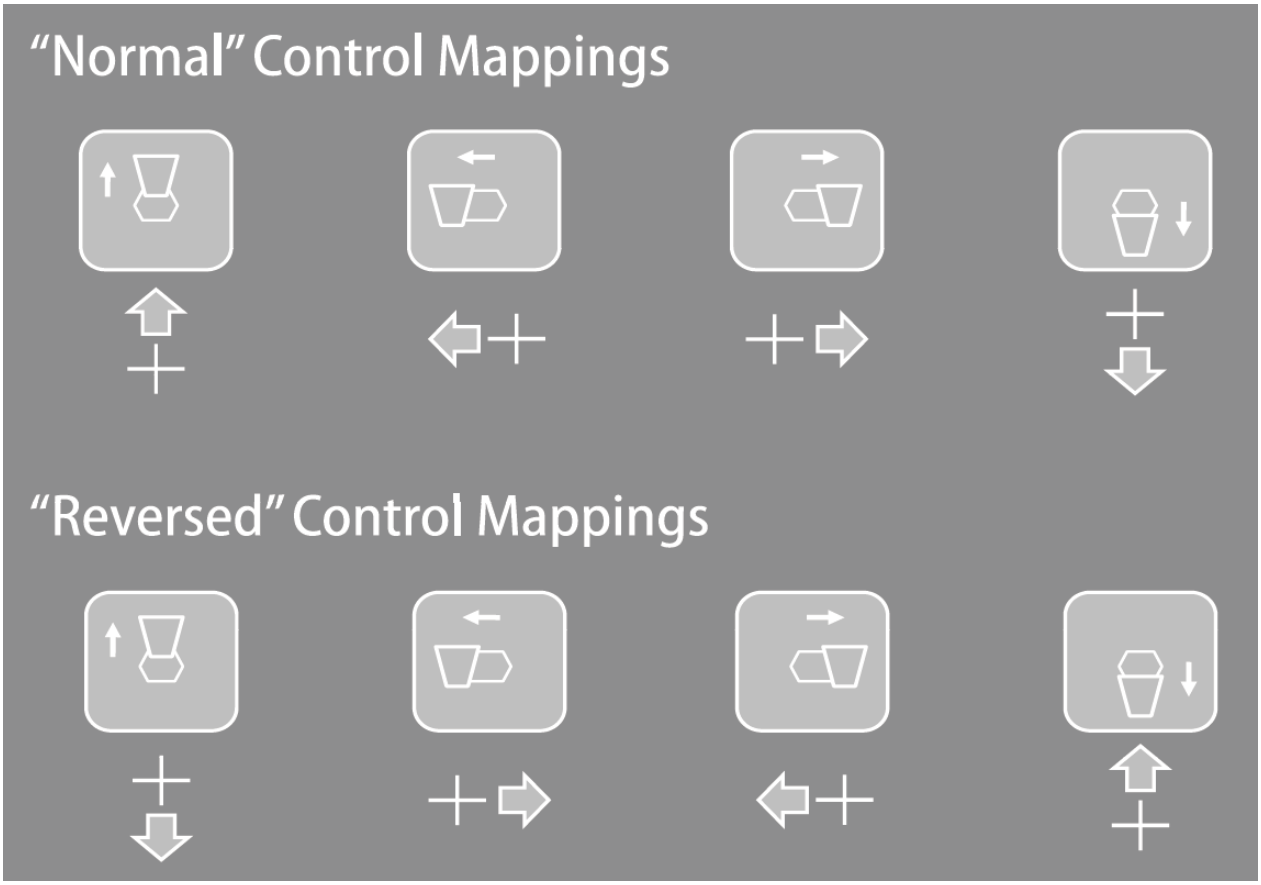


Figure 5.2.2 *Normal and reversed mappings. Ten minutes into the trial the control mappings switched from "normal" to "rotated." At the fifteen minute mark the control mappings abruptly switched back to "normal" for the remaining five minutes of the experiment.*



5.2.2 *Results*

The initial data processing proceeded as described in Experiment 1 with one notable exception. Eye blinks were interpolated with a 3rd order cubic spline before the trial was broken into 34 second segments with 50% overlap and FFT was applied. Figure 5.2.3 through Figure 5.2.8 depict the physiological measures in both the time and frequency domain as well as tracking performance over time. A visual inspection of the plots reveals that the control mapping reversal at the 10 minute mark elicited large transient responses in 4 of the 6 participants. From the plots we can also see that skin conductance transients also occur spontaneously (participant 7 for instance).

After the STFT, Principle Component Analysis (PCA) was independently applied to the GSR and pupil diameter Fourier components. PCA is an analysis technique which attempts to transform a number of correlated variables into a smaller number of uncorrelated variables called the principle components. The first principle component accounts for as much of the variability as possible, and each succeeding variable accounts for as much remaining variability as possible. The Modular Toolkit for Data Processing (MDP) for Python was used to obtain principle components for this analysis. For each measure the first 10 components determined by PCA were used to predict tracking error. The scatterplots in Figure 5.2.9 through Figure 5.2.14 depict the correlations between the principle components and the physiological measures. The top panels depict correlations for skin conductance components and the bottom panels depict correlations for pupil diameter. To assess the combined predictive power of multiple PCA components I ran 2 linear discrimant analyses (one for SC, and one for pupil diameter) per participant using the components as predictors and tracking error as the response variable. No family-wise error correction was applied. Of the 12 analyses only skin conductance with participant 2 was significantly related to tracking error $F(10,58) = 2.42$, $p = 0.017$, $MSE = 12.305$. Table 5.2.1 provides of full summary of these results. When viewed in context these results suggest that the transient skin conductance spike at the 10 minute mark corresponds to the tracking error transient after the 10 minute mark.

5.2.3 *Conclusions and Discussion.*

The observation that 4 of the 6 participants had transients in skin conductance when the mappings first changed is encouraging. Part of the problem with the discrimination analysis is that most of the participants performed the tracking task well, even with the "reverse" control dynamics (see Figure 5.2.2). As a result, the tracking error distributions show extreme positive skew, reflecting a limited range of variation in task difficulty . A task that has a wider and perhaps more normally distributed range of task difficulty and error may be needed for physiological measures and principle component analysis to work reliably. Another problem is that stress induced by the mapping change (as observed by the raw skin conductance trends) seems to subside as participants become accustomed to the reversed mappings. My next experiment addressed these concerns by using normal and rotated mappings and increasing the frequency with which the mappings were switched. I also introduce potentially more powerful analytical tools (wavelet decomposition and genetic programming) to predict tracking error and the control mapping state.

Table 5.2.1
LDA results for SC and PD on Tracking Error

Participant	Phyiological Measure	(df_{source} , df_{error})	*F*	*p*	MSE
2	Skin Conductance	(10,58)	2.42	0.017	12.305
	Pupil Diameter	(10,58)	0.64	0.777	0.414
3	Skin Conductance	(10,58)	0.33	0.969	5.750
	Pupil Diameter	(10,58)	0.85	1.587	0.414
5	Skin Conductance	(10,58)	0.91	0.530	0.048
	Pupil Diameter	(10,58)	1.45	0.071	0.071
6	Skin Conductance	(10,58)	0.35	0.962	1.780
	Pupil Diameter	(10,58)	0.84	0.594	3.940
7	Skin Conductance	(10,58)	0.54	0.852	0.665
	Pupil Diameter	(10,58)	1.35	0.225	1.471
8	Skin Conductance	(10,58)	0.69	0.721	0.955
	Pupil Diameter	(10,58)	0.65	0.765	0.894

Table 5.2.2

LDA vs. Symbolic Regression on Tracking Error (r^2)

	LDA		*Symbolic Regression*	
Participant	*Best Training*	*Best Test*	*Best Training*	*Best Test*
1	0.273	0.115	0.680	0.119
2	0.399	0.179	0.722	0.314
3	0.415	0.069	0.617	0.187
4	0.240	0.177	0.568	0.189
5	0.397	0.019	0.567	0.181
6	0.449	0.314	0.579	0.332
7	0.307	0.089	0.705	0.104
8	0.588	0.481	0.636	0.443
9	0.276	0.175	0.527	0.160
10	0.411	0.115	0.760	0.121
Averages:	0.376	0.173	0.636	0.215

Figure 5.2.3 *Experiment 2, Participant 2.*

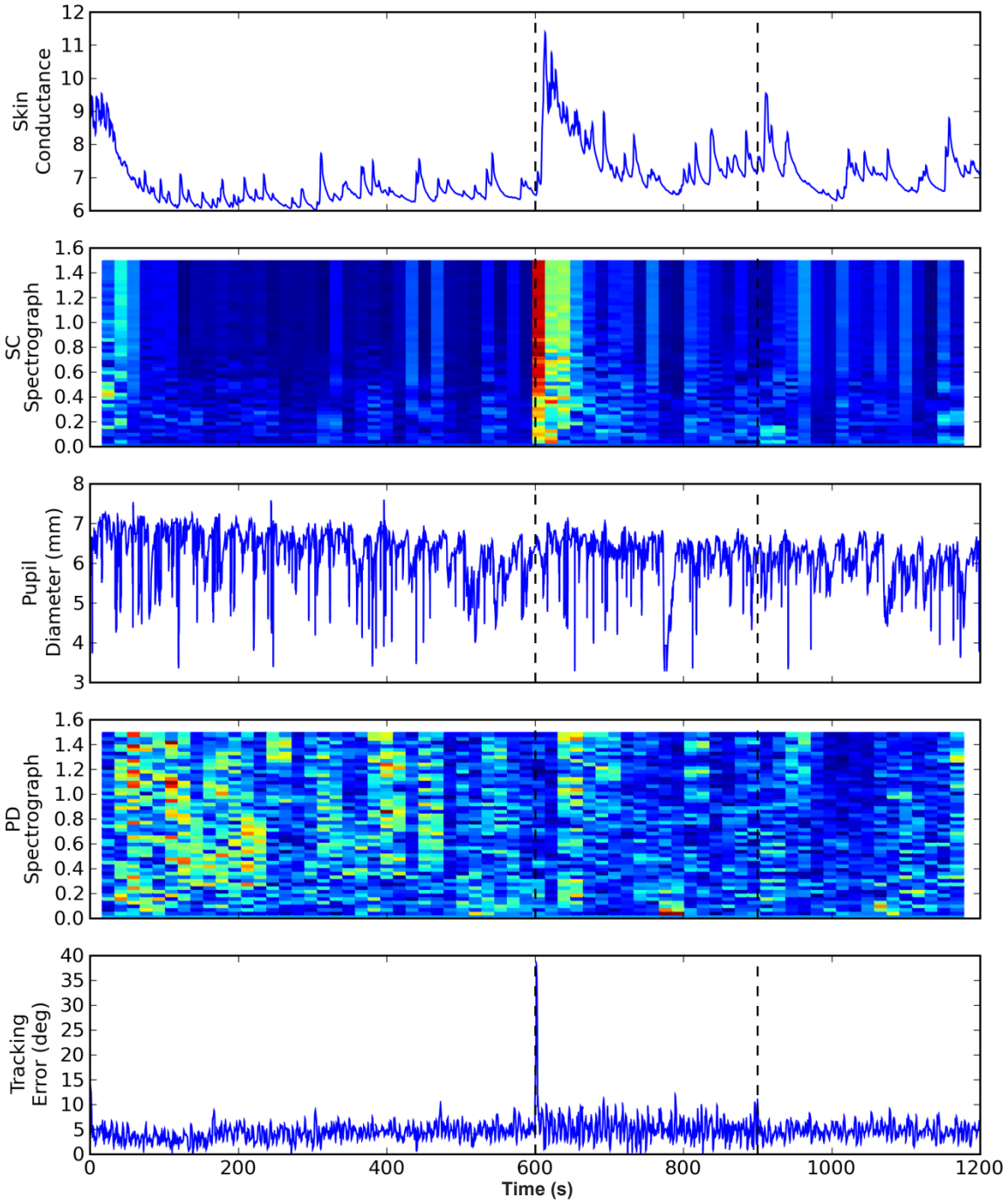


Figure 5.2.4 *Experiment 2, Participant 3.*

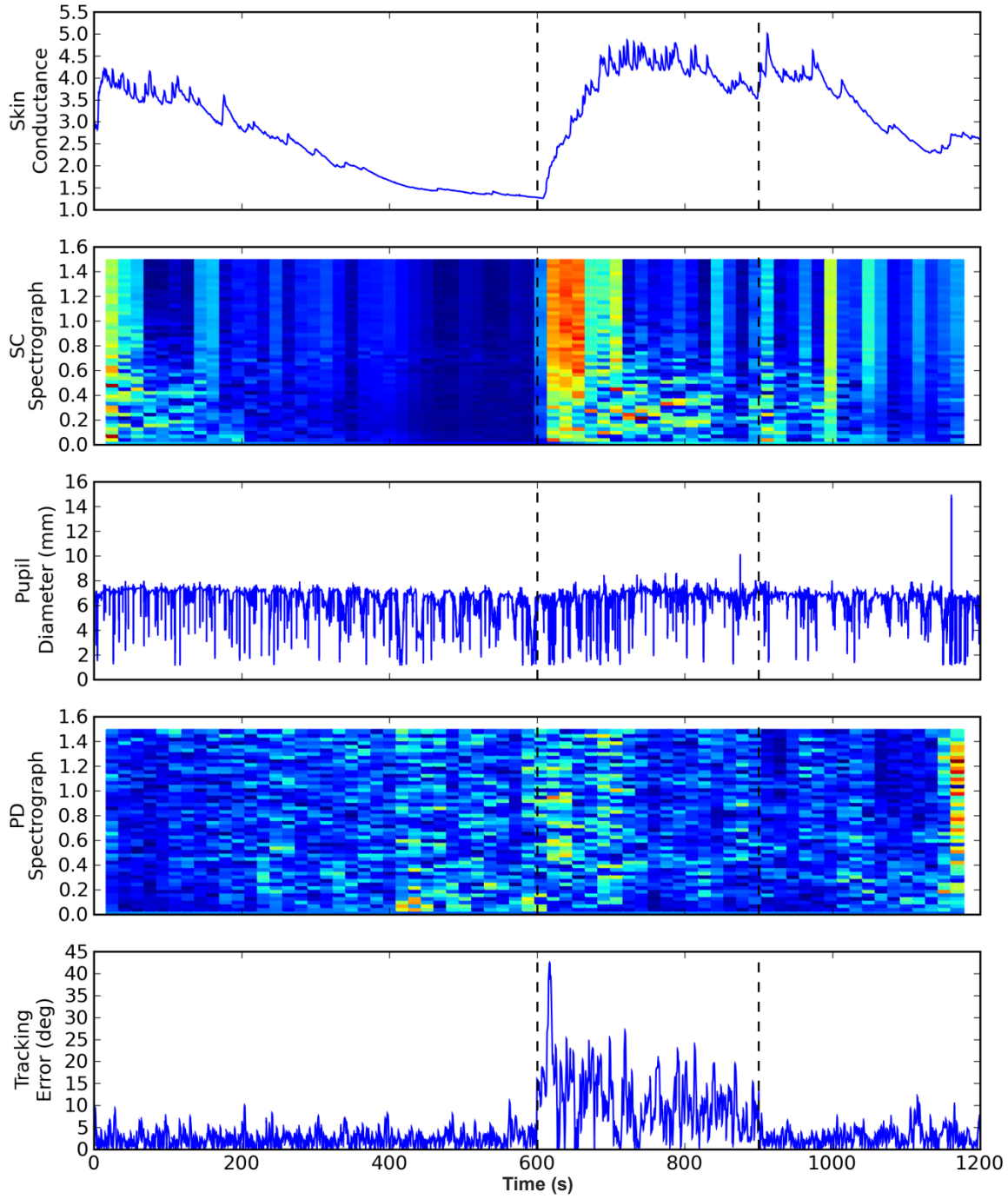


Figure 5.2.5 *Experiment 2, Participant 5.*

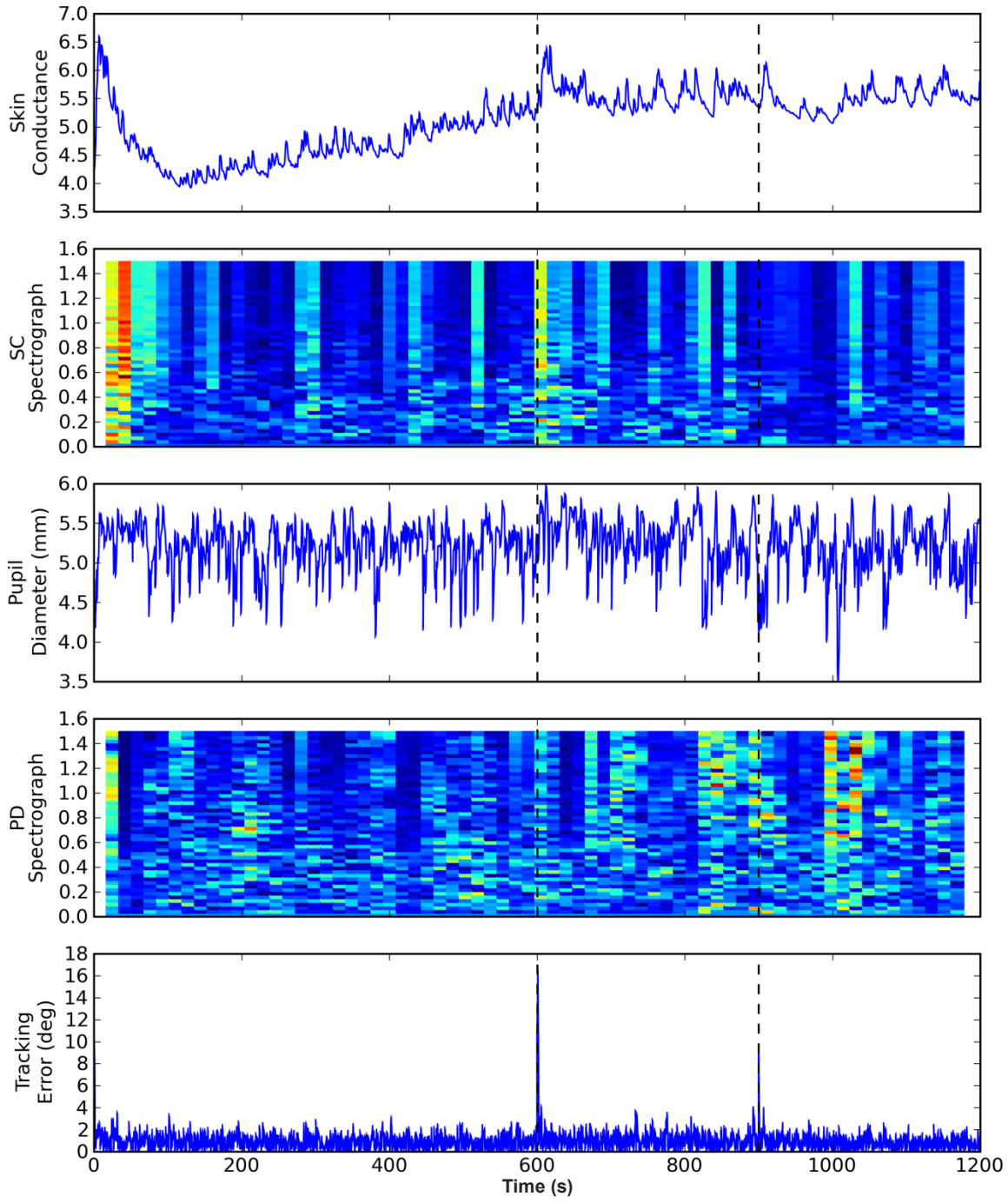


Figure 5.2.6 *Experiment 2, Participant 6.*

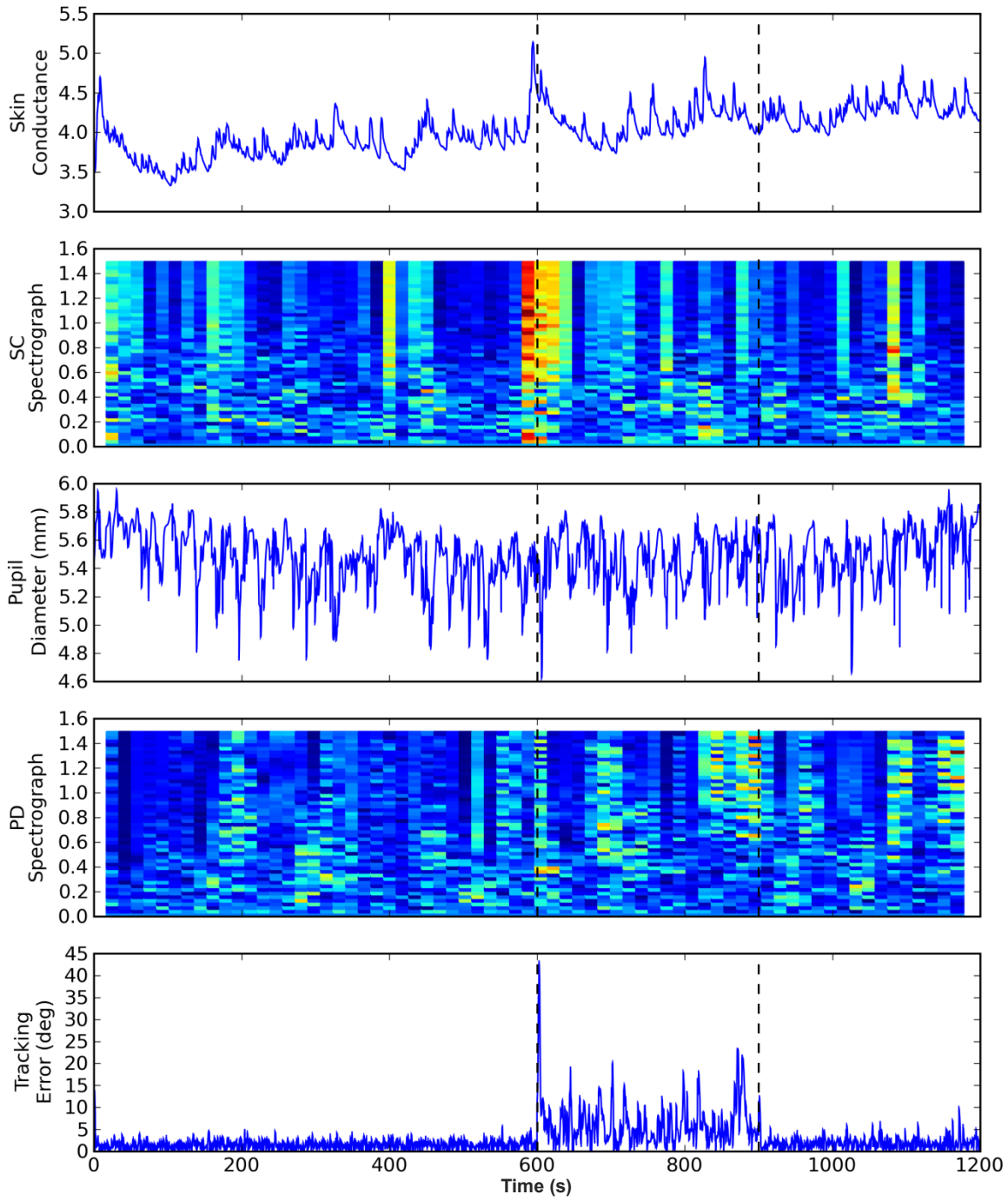


Figure 5.2.7 *Experiment 2, Participant 7.*

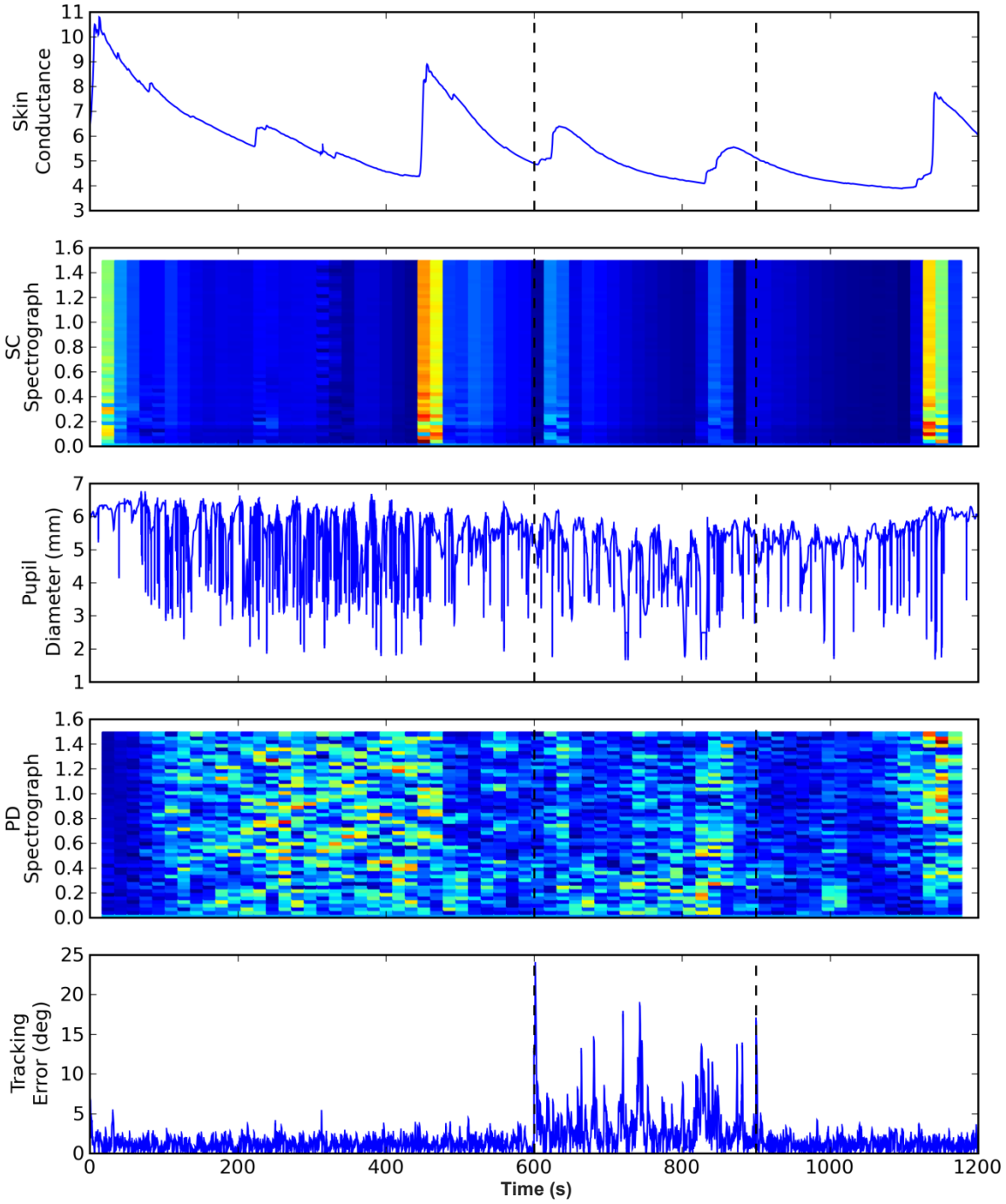


Figure 5.2.8 *Experiment 2, Participant 8.*

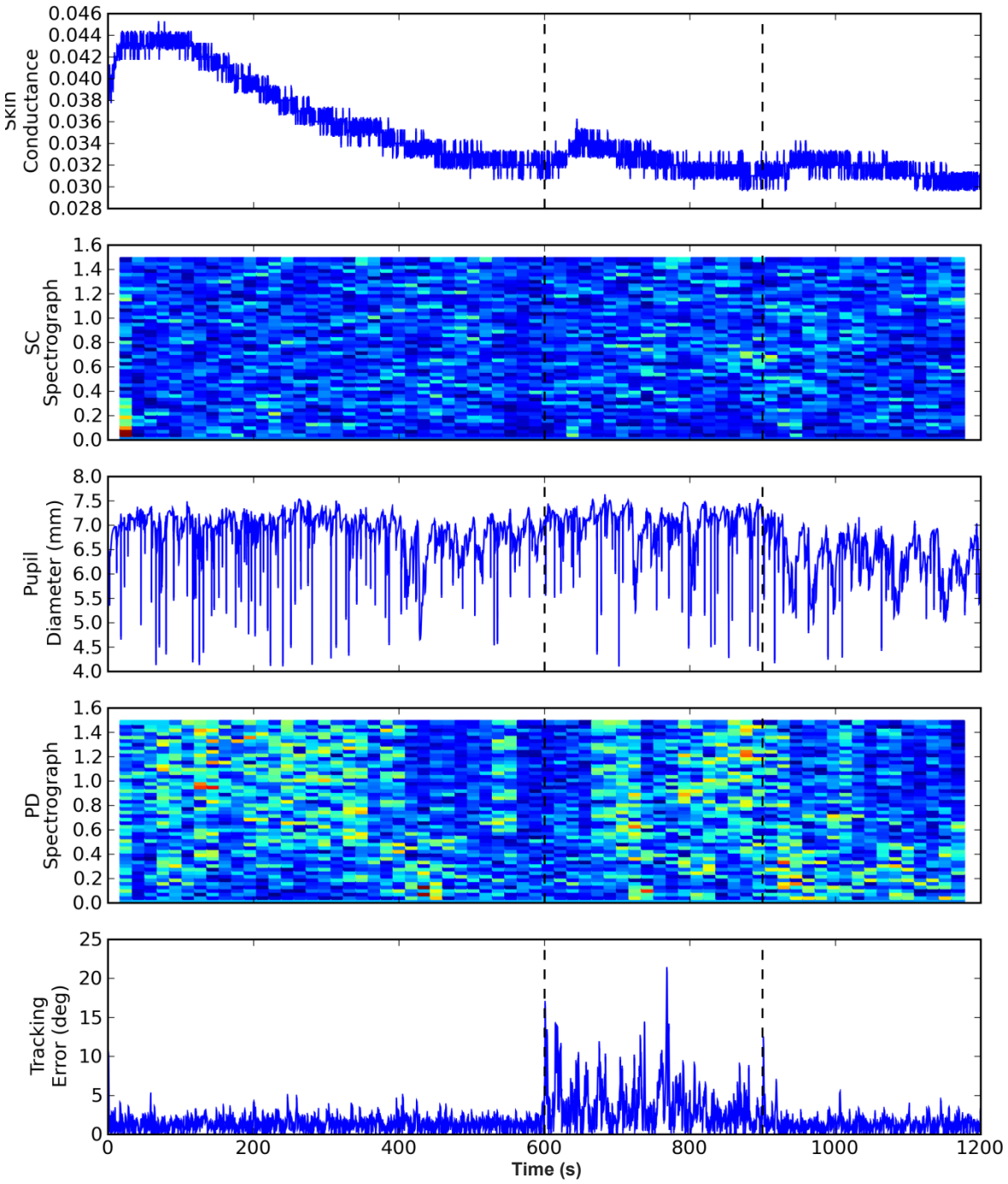


Figure 5.2.9 *Experiment 2, Participant 2 PCA scatterplots. X-axis is dimensionless.*

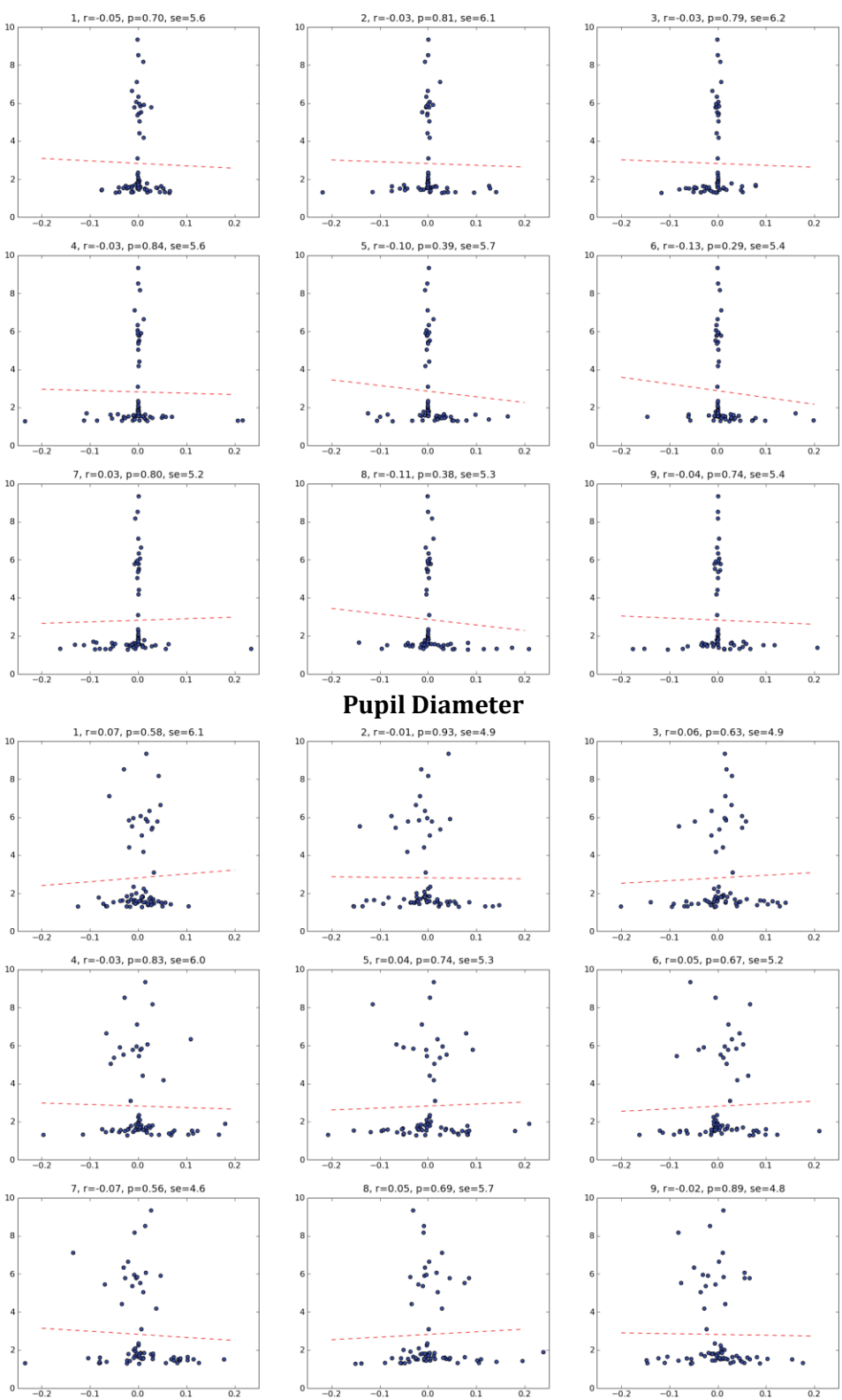


Figure 5.2.10 *Experiment 2, Participant 3 PCA scatterplots. X-axis is dimensionless.*

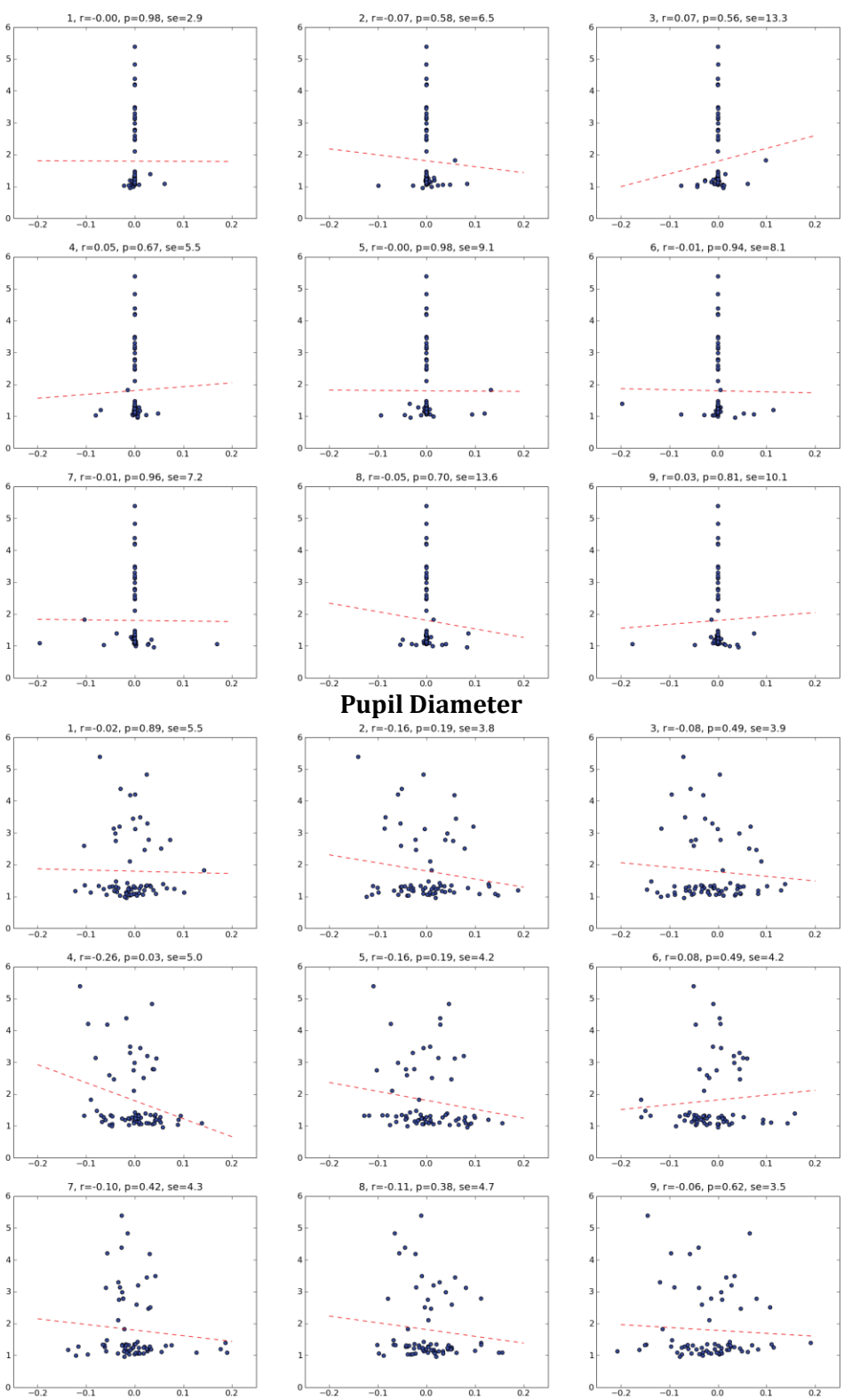


Figure 5.2.11 *Experiment 2, Partipant 5 PCA scatterplots. X-axis is dimensionless.*

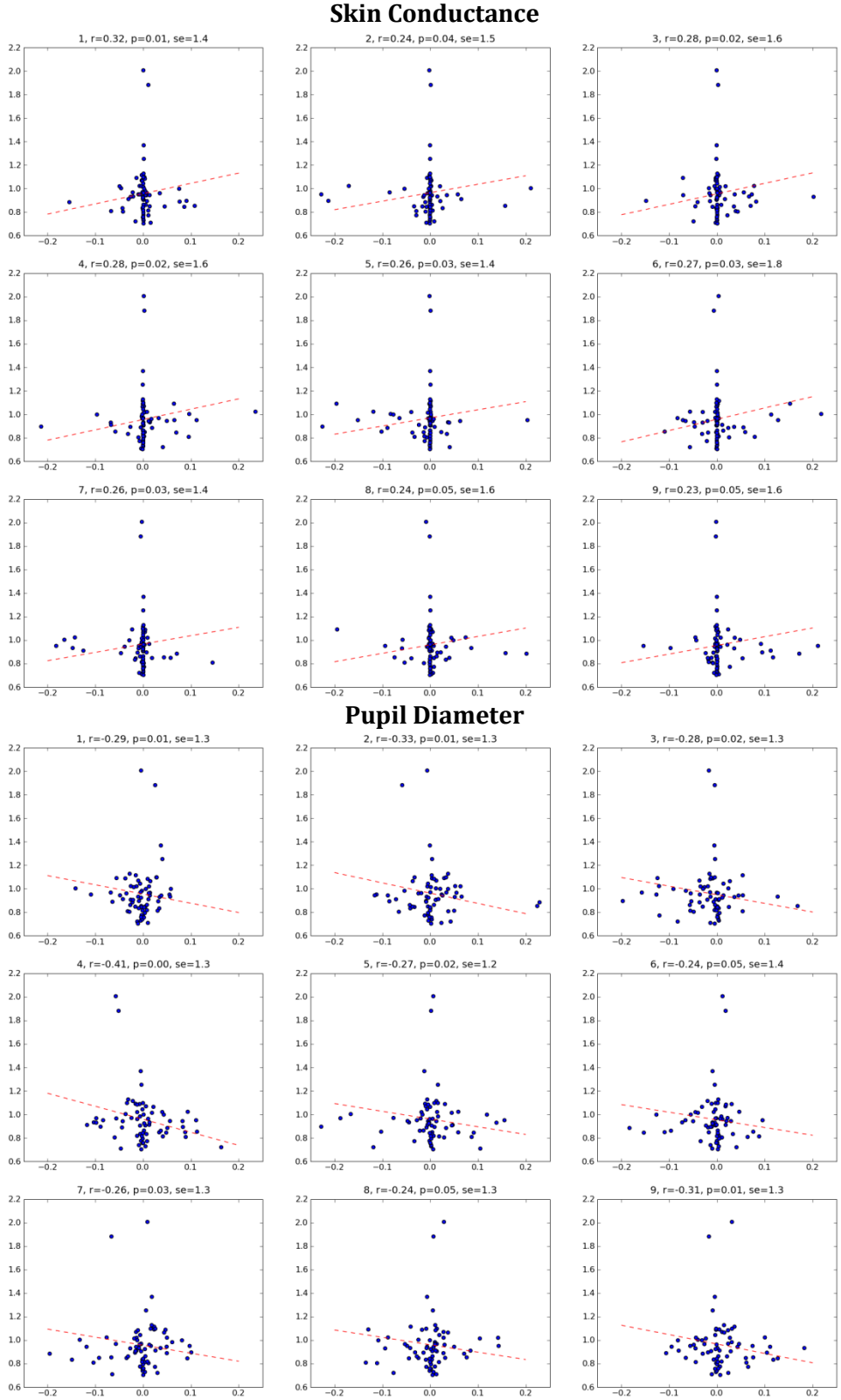


Figure 5.2.12 *Experiment 2, Participant 6 PCA scatterplots. X-axis is dimensionless.*

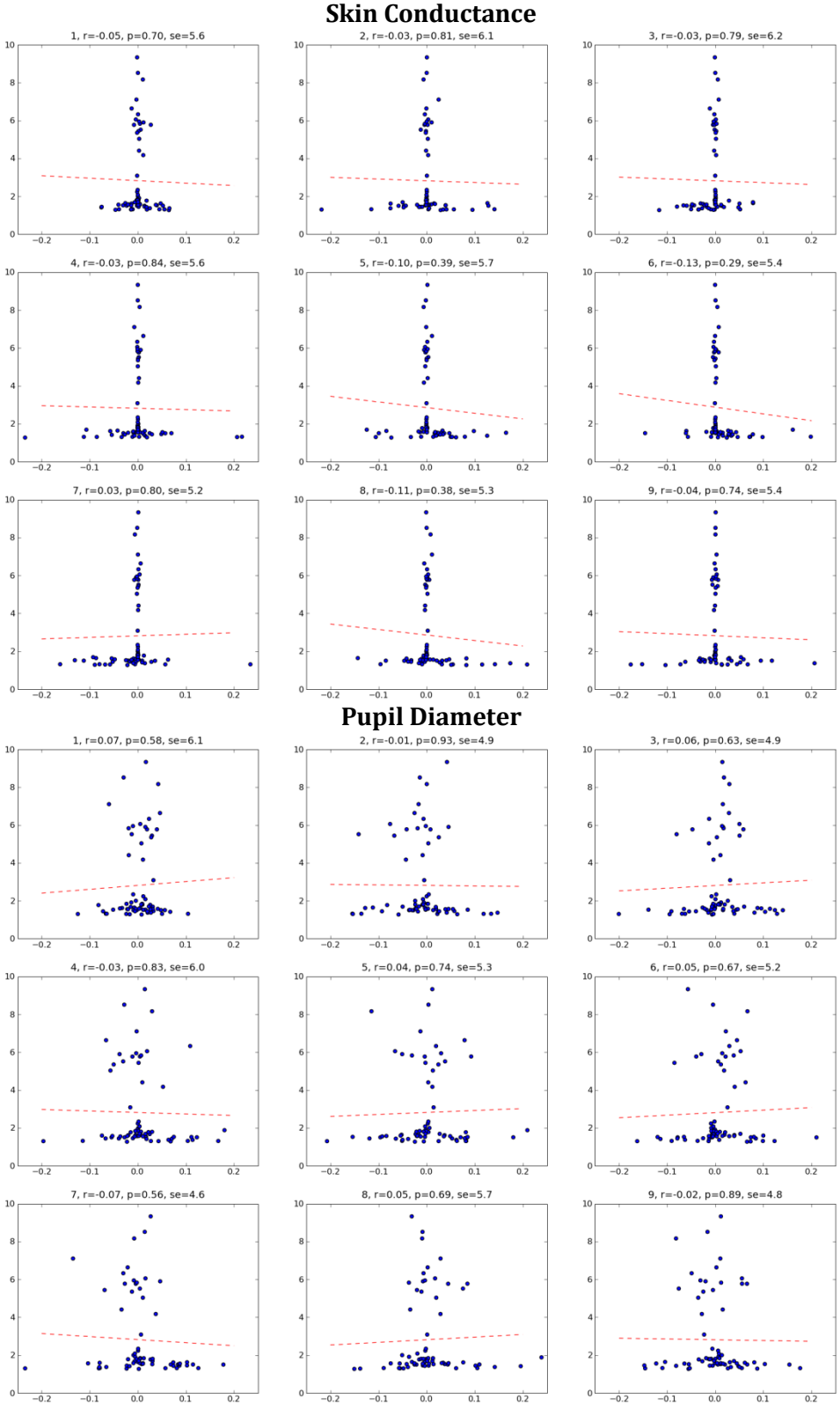


Figure 5.2.13 *Experiment 2, Participant 7 PCA scatterplots. X-axis is dimensionless.*

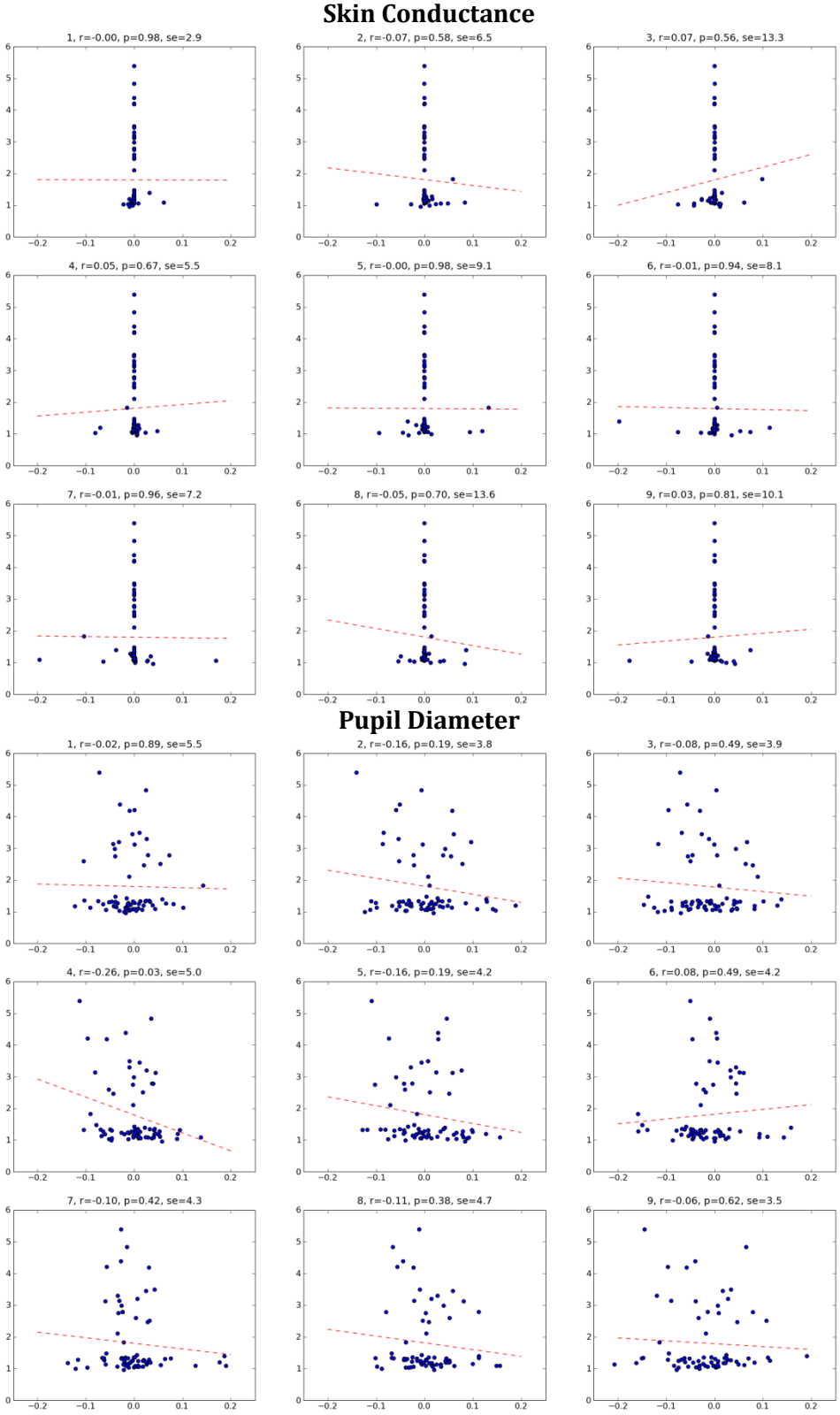
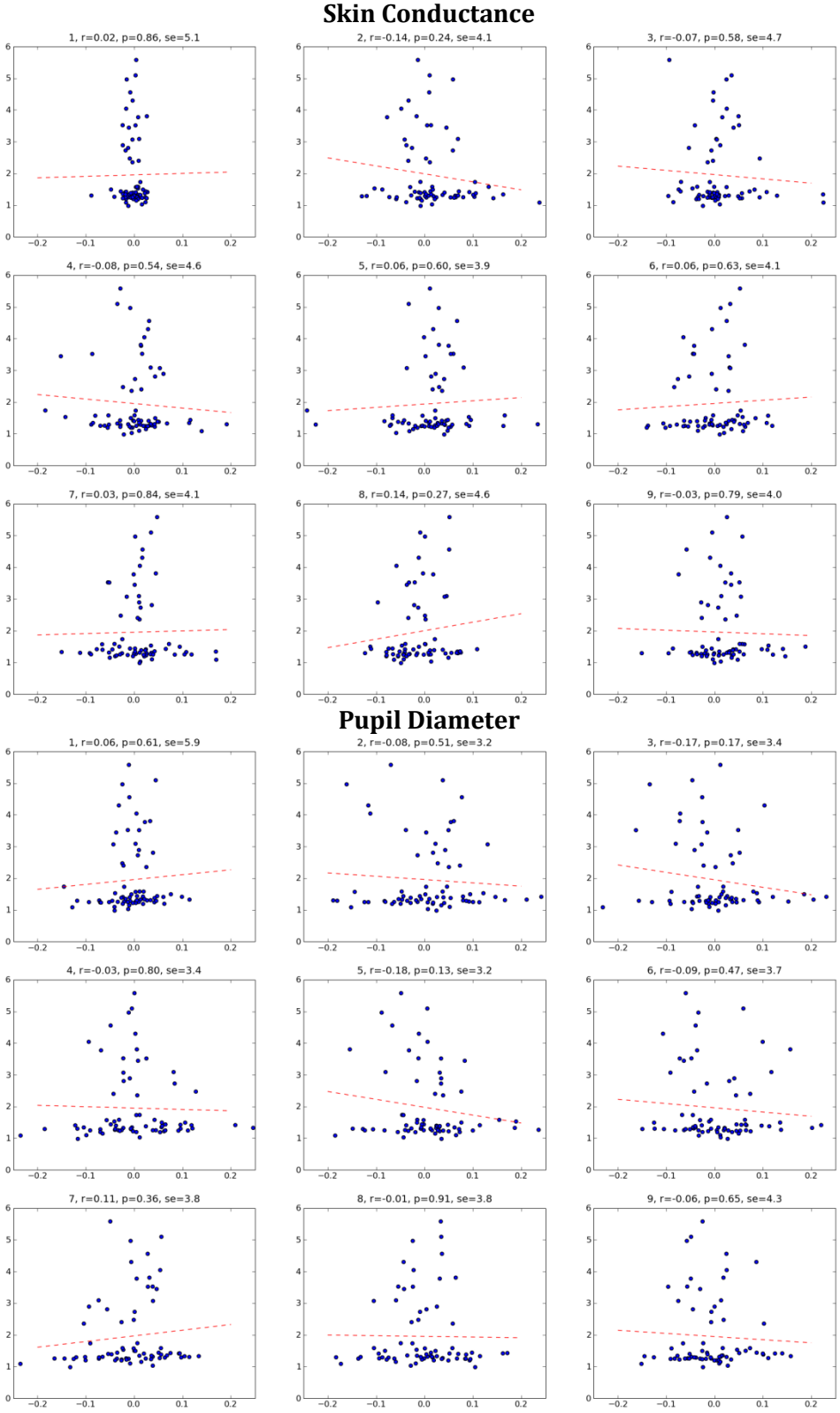


Figure 5.2.14 *Experiment 2, Participant 8 PCA scatterplots. X-axis is dimensionless.*



Appendix 5.2.A Consent Form

CONSENT FORM

Idaho Visual Performance Laboratory
Department of Psychology and Communication Studies
College of Liberal Arts and Social Sciences
University of Idaho
Control of speed during altitude changes

During this experiment you will be presented a display in a virtual environment. Various parameters of this display will be manipulated to examine stress and mental workload. In this experiment you will be asked to control movement in the virtual world using an input device such as a joystick.

The data you provide will be kept anonymous. There will be absolutely no link between your identity and your particular set of data.

Your participation will help increase knowledge of stress and mental workload. Subsequent to your participation the purpose and methods of the study will be described to you and questions about the study will be answered. It is our sincere hope that you will learn something interesting about your visual system from this debriefing.

The risks in this study are minimal, however displays simulating movement may on rare occasion cause motion sickness or eye fatigue in sensitive individuals. If at any time during the experiment you feel eye fatigue, dizziness, headache or nausea, please let the experimenter know immediately so that you can take a break before these symptoms become too intense. We endeavor to design our displays to minimize eye fatigue and motion sickness, and schedule periodic breaks to further reduce their occurrence. As a result, these phenomena have not been a common problem in previous similar studies.

Your participation will require **1** session of approximately **30** minutes. You may withdraw from this study at anytime without penalty. You will receive partial credit for your time spent. However, please be aware that your data is useful to us only if you complete the experiment in its entirety.
This research project has been approved by the University of Idaho Human Assurance Committee. As such, new information developed during the course of the research which may relate to your willingness to continue participation will be provided to you.

Thank you for your participation

Signature_____ Date _____

If you have further questions or encounter problems please contact:
Dr. Brian P. Dyre
(208) 885-6927
bdyre@uidaho.edu

Appendix 5.2.B Debriefing Form

Debriefing Form

Department of Psychology and Communication Studies

College of Letters, Arts, and Social Sciences

INL Physiological Predictors of Workload

Experiment 3

Participant:____

Date:_____

1. Did you move your left hand during the course of the trial while the GSR was still hooked up?

2. How often do you play video games?

    a. What is your video game skill? (Bad, okay or good)

    b. Are you right or left handed?

3. Did you notice that the controls changed throughout the trial?

    a. How many times?

4. How difficult was the task when you first started? (1-10)

5. How difficult were the normal vs. reversed controls? (1-10)

6. Did you feel that you had enough time to feel confident with:

    a. Normal mappings?

   b. Rotated mappings?

7. How uncomfortable was the eye-tracker when you first put it on? (1-10)

8. How uncomfortable was the eye-tracker when you finished? (1-10)

9. Did you find the eye-tracker distracting from the task at hand?

10. Do you think that fatigue played a role in your performance?

   a. How about fatigue from the eye-tracker?

11. Did you have any eye-strain, fatigue, blurred vision, problems focusing on the target, etc. ?

Any additional comments

Appendix 5.2.C Human Assurances Approval

University of Idaho

University Research Office
Moscow, ID 83844-3040

Federalwide Assurance: FWA00005639
Federal Assigned IRB #: 00000843
UI Assigned Number: 07-115

MEMORANDUM

TO: Brian Dyre
Psychology & Communication Studies Department - 3C43

FROM: Eric Jensen, Chair
Human Assurances Committee

DATE: October 23, 2007

SUBJECT: Approval of "Perception and Control of Locomotion in Virtual Environments."

On behalf of the Human Assurances Committee at the University of Idaho, I am pleased to inform you that the above-named proposal is approved as offering no significant risk to human subjects. This approval is valid for one year from the date of this memo. Should there be a significant change in your proposal, it will be necessary for you to resubmit it for review. Thank you for submitting your proposal to the Human Assurances Committee.

Eric L. Jensen
ELJ/ec

5.3 Experiment 3: Pursuit Tracking (Normal vs. Rotated)

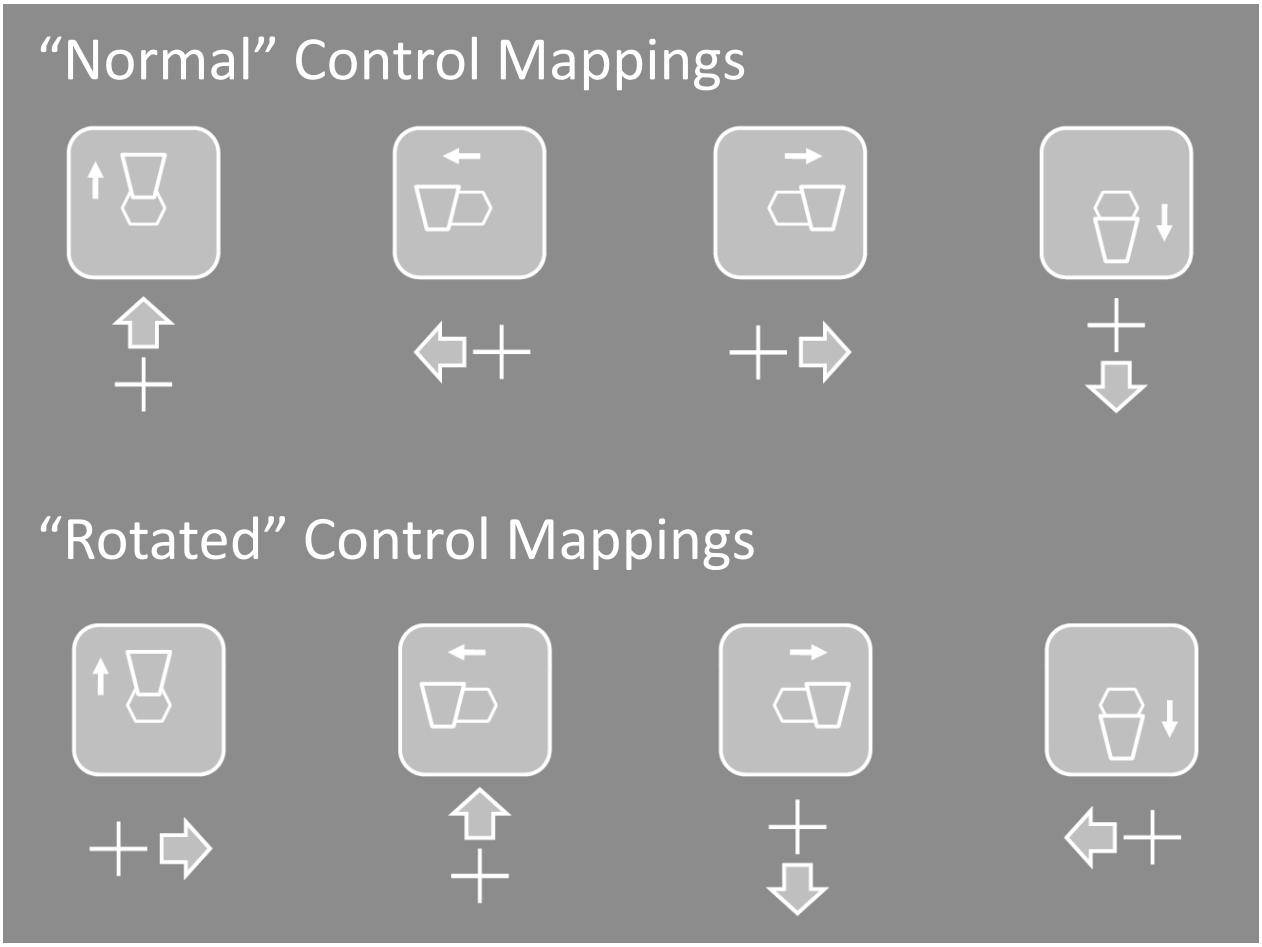
The reversed control mappings used in Experiment 2 did not reliably manipulate task difficulty, which likely contributed to the experiment's inconclusive results. The aim of Experiment 3 was to replicate Experiment 2 with a more reliable manipulation of task difficulty. To that end, I pilot tested alternative difficulty manipulations, and found that rotating the control mappings 90 degrees appeared to reliably reduce tracking performance. In addition, tracking performance continued to be negatively impacted up to 5 minutes after the mappings were rotated. Experiment 3 was similar to Experiment 2 with the exception of the manipulation of task difficulty, which replaced the reversed mappings with rotated mappings, and the control mappings switched every 60 seconds for a total of seven transitions compared to the two transitions used in Experiment 2.

To assess whether control mapping stated produced measurable differences in physiological signals wavelet decomposition was applied to skin conductance (SC) and pupil diameter (PD). Then Genetic programming (GP), a machine learning technique, was used to build models of tracking performance based on SC and PD as well as classifiers of the control mapping state. The performance of Genetic Programming was compared to linear discriminant analysis (LDA).

5.3.1 Method

5.3.1.1 Participants. Ten University of Idaho students participated in this experiment. All had normal or corrected to normal Snellen visual acuity (20/30 or better). Participant 3 had limited knowledge of the hypotheses of the experiment; the remaining participants were naïve to the hypotheses of the experiment. All participants were ethically treated in accordance with experimental protocols approved by the University of Idaho's Human Assurance Committee (see Appendices 5.3.A – 5.3.C).

Figure 5.3.1 *Normal vs. rotated mappings. Every 60 seconds for 480 seconds the control mappings switched from "normal" to "rotated."*



5.3.1.2 Stimuli and Apparatus. As with the previous experiment task difficulty was manipulated using a pursuit tracking task in which participants tracked a balanced dot moving in a pseudo-random fashion with a black cursor superimposed on a gray background. However in this experiment the control mappings were rotated (see **Figure 5.3.1**) instead of reversed; and the mappings were changed several times throughout the trial. The abrupt changes in control mappings were hypothesized to elicit transient physiological responses, and the rotated mappings were hypothesized to cause lower performance (higher tracking error defined by the Euclidean distance between the center of the balanced dot and the cursor) and physiological indicators reflecting increased workload.

The simulation was presented in the same darkened room with the same equipment previously described in the first two experiments. As with the previous experiments PD and SC were recorded while the participants performed the task.

5.3.1.3 Procedure. Participants controlled the cursor using a right-hand joystick with first order control dynamics and a gain of 25° per second at maximum deflection. For the first minute of the experiment the control mappings were normal: moving the joystick forward moved the cursor up, moving the joystick backward moved the cursor down, moving the joystick right moved the cursor right, and moving the joystick left moved the cursor left. After 60 s the joystick control mappings were rotated abruptly 90° clockwise, such that moving the joystick forward-backward moved the cursor rightward-leftward, and moving the joystick leftward-right moved the cursor upward-downward. For the eight minute duration of the experiment the control dynamics were rotated from the normal orientation to 90 degrees, then back to normal, etc. every 60 seconds. Participants performed the tracking task for a total of eight minutes comprised of four minutes of pursuit tracking with the normal mappings and for minutes of pursuit tracking with the rotated mappings.

5.3.2 *Tracking Error Models*

5.3.2.1 Preprocessing. Offline the 480 second trial was divided into eight epochs of 60 seconds each. Epochs 1, 3, 5, and 7 had the normal control mappings, and the remaining even epochs had the rotated mappings. Data analysis proceeded by smoothing the PD data using a third-order cubic spline, centering the data by subtracting the mean diameter, and then log transforming the result. Negative values were treated by log-transforming the absolute value and then multiplying by -1. Next, to prepare the data for wavelet analysis, SC and PD were linearly interpolated to a sampling rate of 34-1/3 Hz so that each 60 second epoch contained 2048 samples. A discrete wavelet transform (DWT) was applied to these 60 second epochs resulting in a 1024 vector of detail coefficients and eight approximation coefficients (of lengths 512, 256, 128, 64, 32, 16, 8, 8). Each of the approximation coefficients and smoothed tracking error were then stretched to a length of 1024 so they would be in a convenient representation for model fitting. Models were fit to each participant's data independently. It is important to note that although these predictions were calculated offline subsequent to data collection, in principle, once a model is developed and trained, its predictions could be calculated in real-time by a modestly fast computer. Figures Figure 5.3.2 - Figure 5.3.10 contain spectrograms (frequency by time) and phaseograms (phase by time) for pupil diameter of all ten participants. These are visual depictions of the data contained in the PD wavelet coefficients.

5.3.2.2 Genetic Programming. An introduction to genetic programming (GP) can be found in Chapter 4. The symbolic regressor GP model utilized a technique known as Age Layered Populations to increase the robustness of search and as a preventative measure against premature convergence (Hornby, 2009). Scaled symbolic error, as described by (Keizer, 2004), was used in place of root-mean-squared error to increase model performance. Scaled symbolic error allows the GP try and fit the fine changes of the fitness landscape, while leaving the gross fitting to a simple linear regression analysis.

Figure 5.3.2 *Participant 1 raw pupil diameter, CWT scalogram, and CWT phaseogram.*

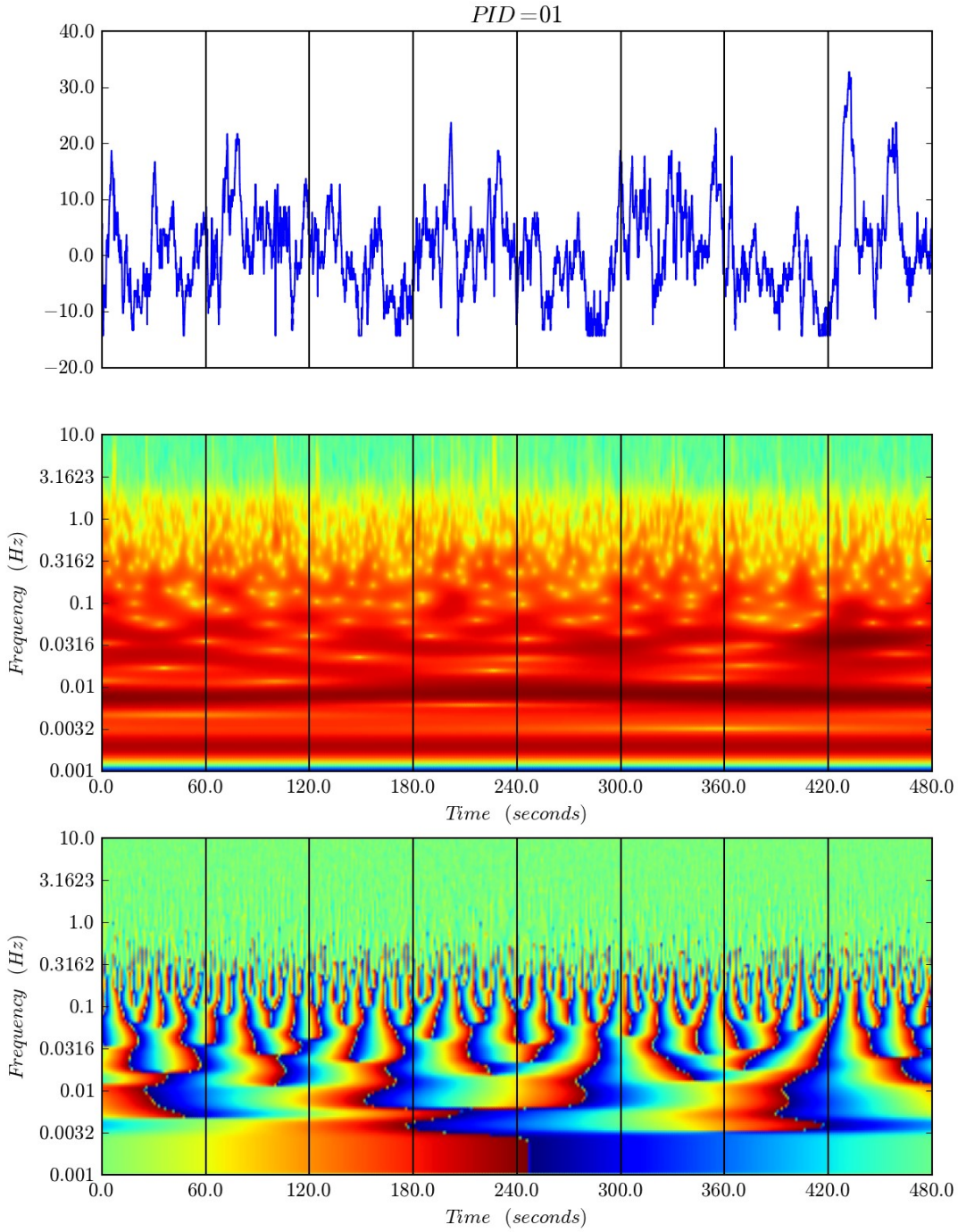


Figure 5.3.3 *Participant 2 raw pupil diameter, CWT scalogram, and CWT phaseogram.*

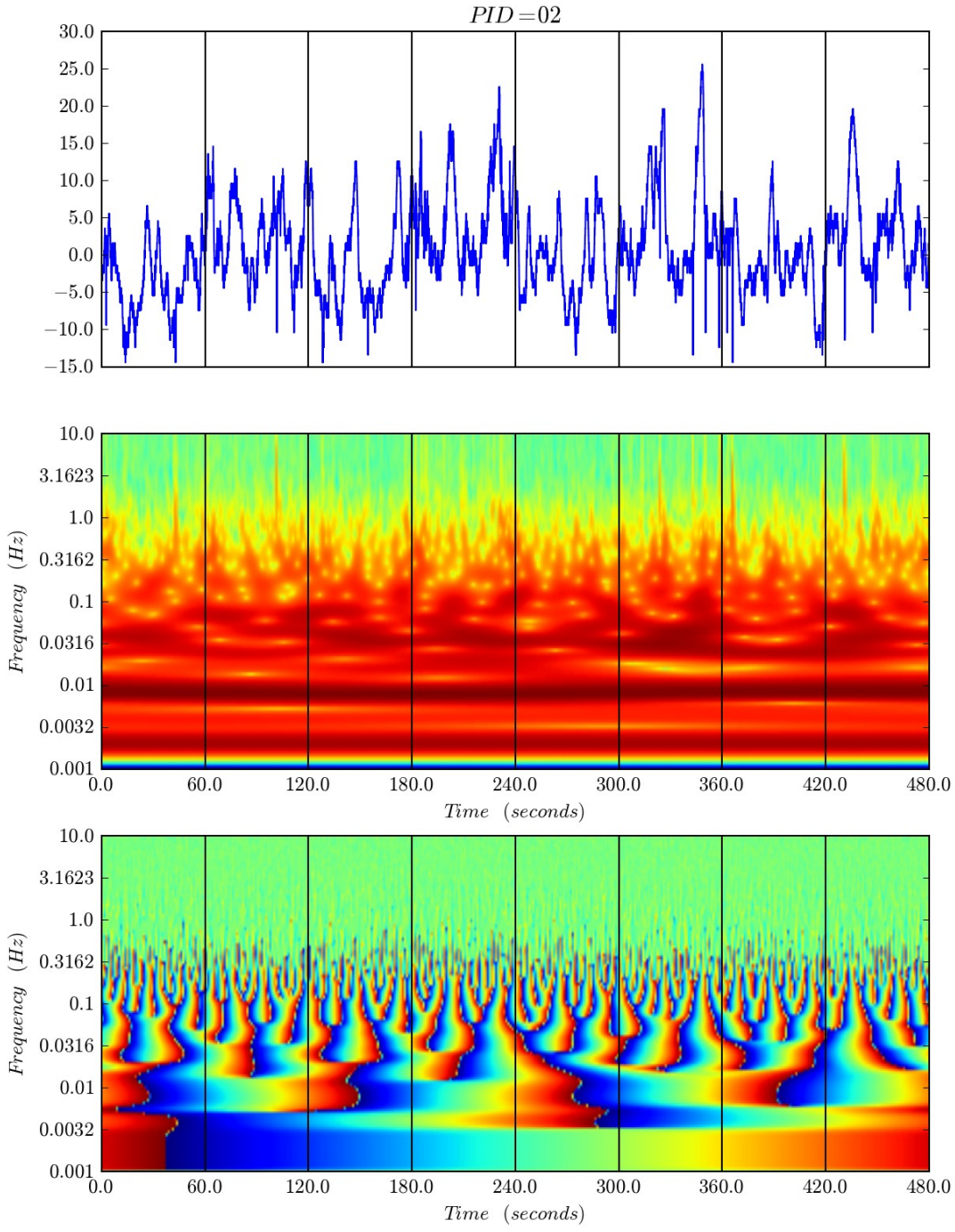
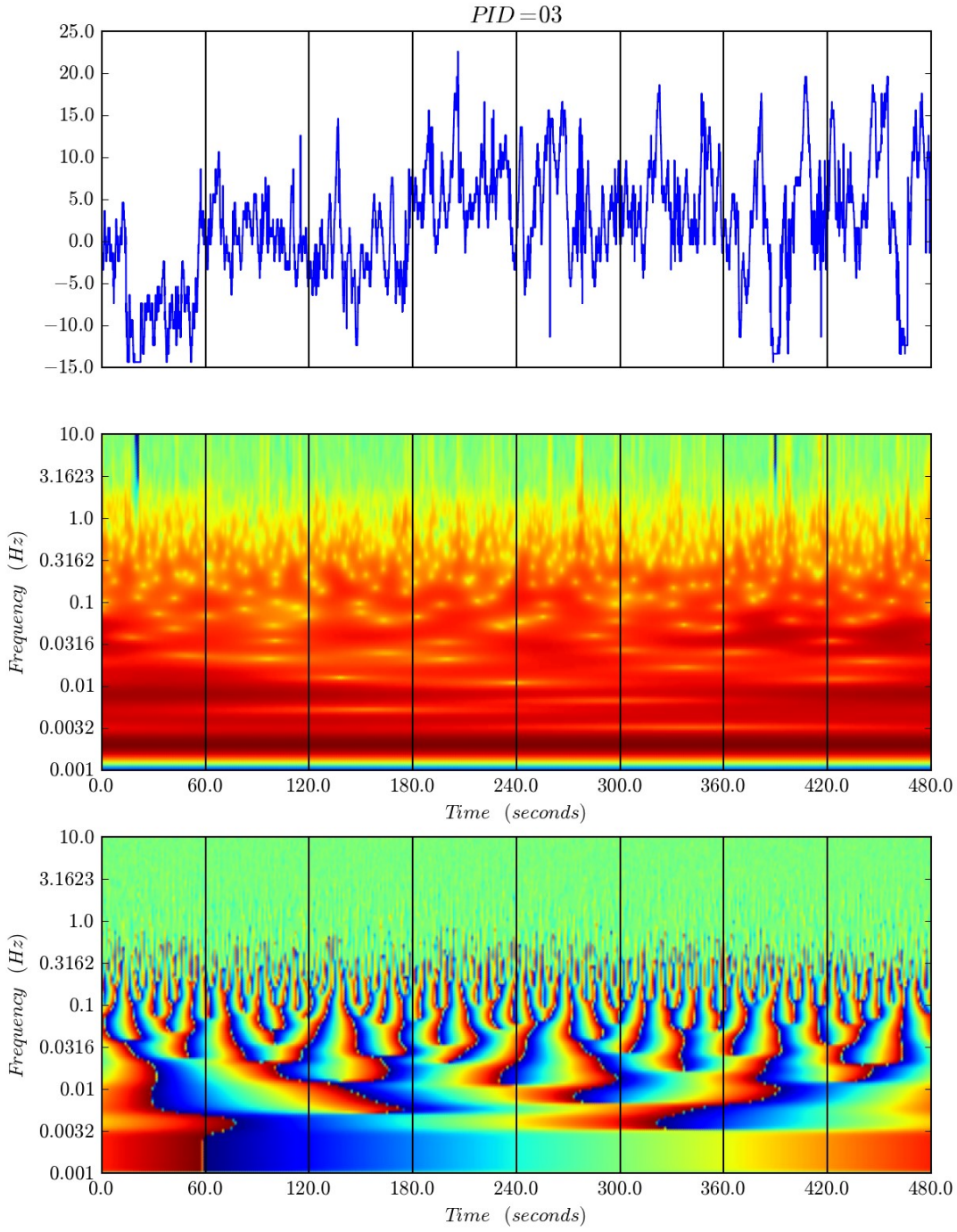


Figure 5.3.4 *Participant 3 raw pupil diameter, CWT scalogram, and CWT phaseogram.*



.

Figure 5.3.5 *Participant 4 raw pupil diameter, CWT scalogram, and CWT phaseogram.*

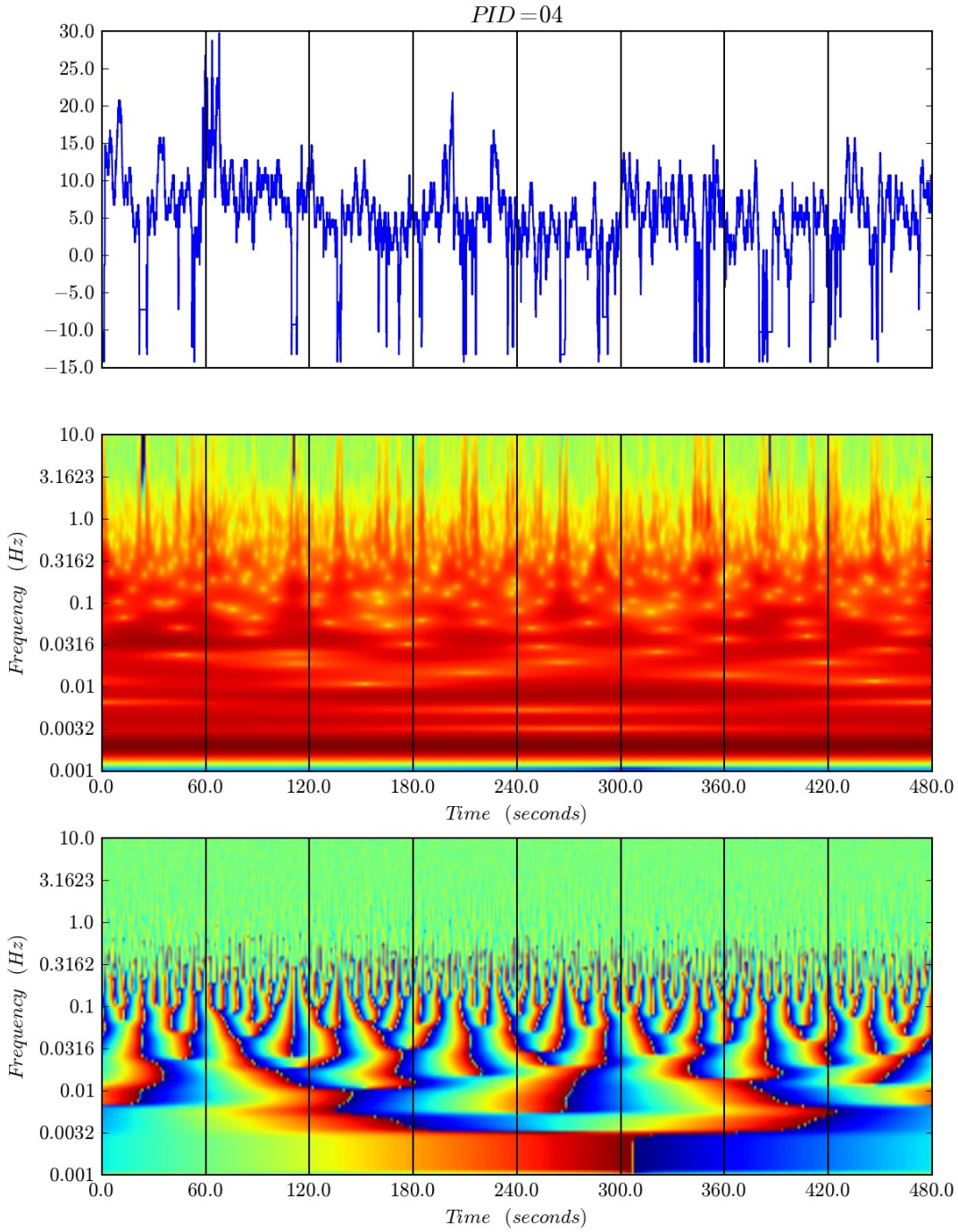


Figure 5.3.6 Participant 5 raw pupil diameter, CWT scalogram, and CWT phaseogram.

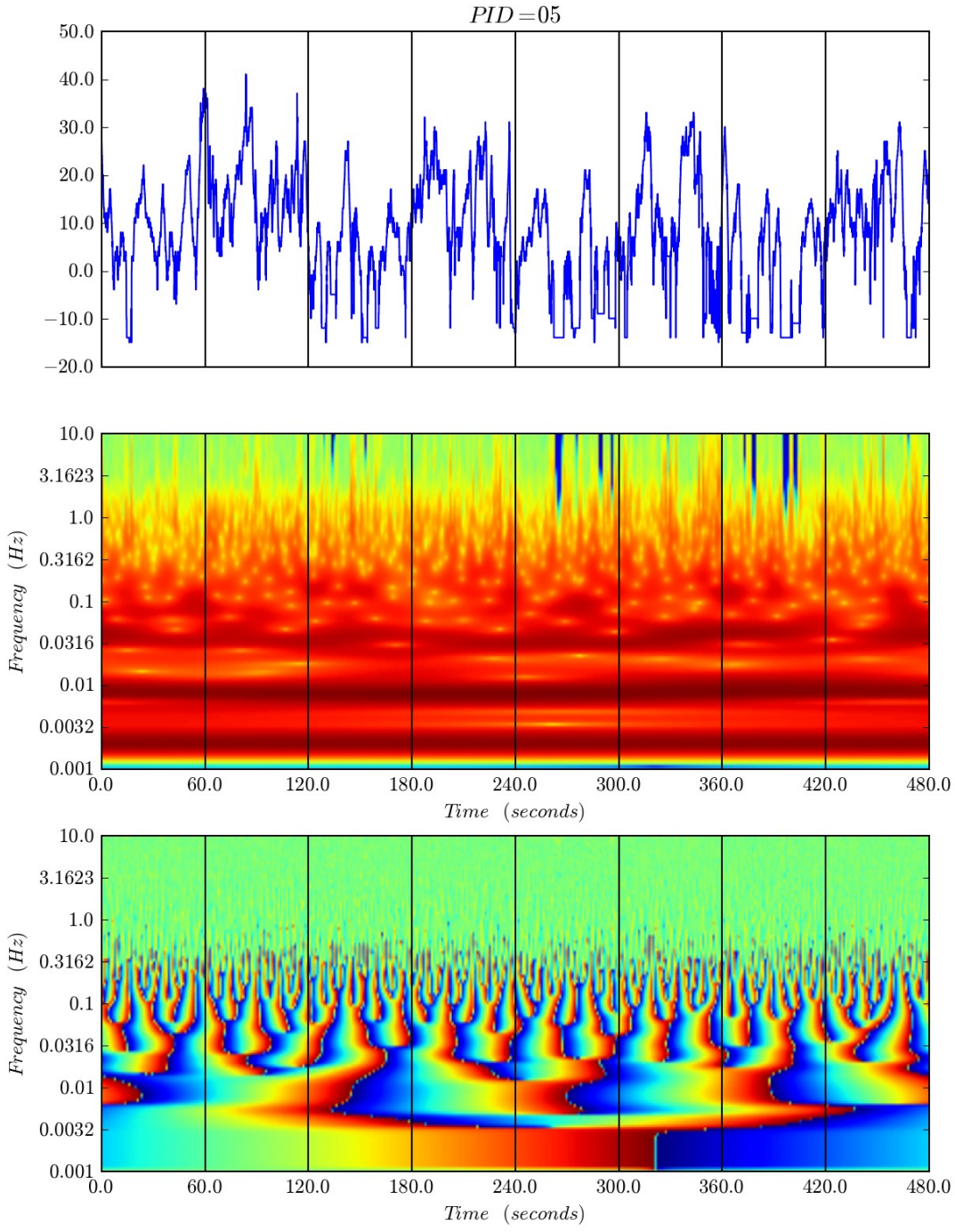


Figure 5.3.7 Participant 5 raw pupil diameter, CWT scalogram, and CWT phaseogram.

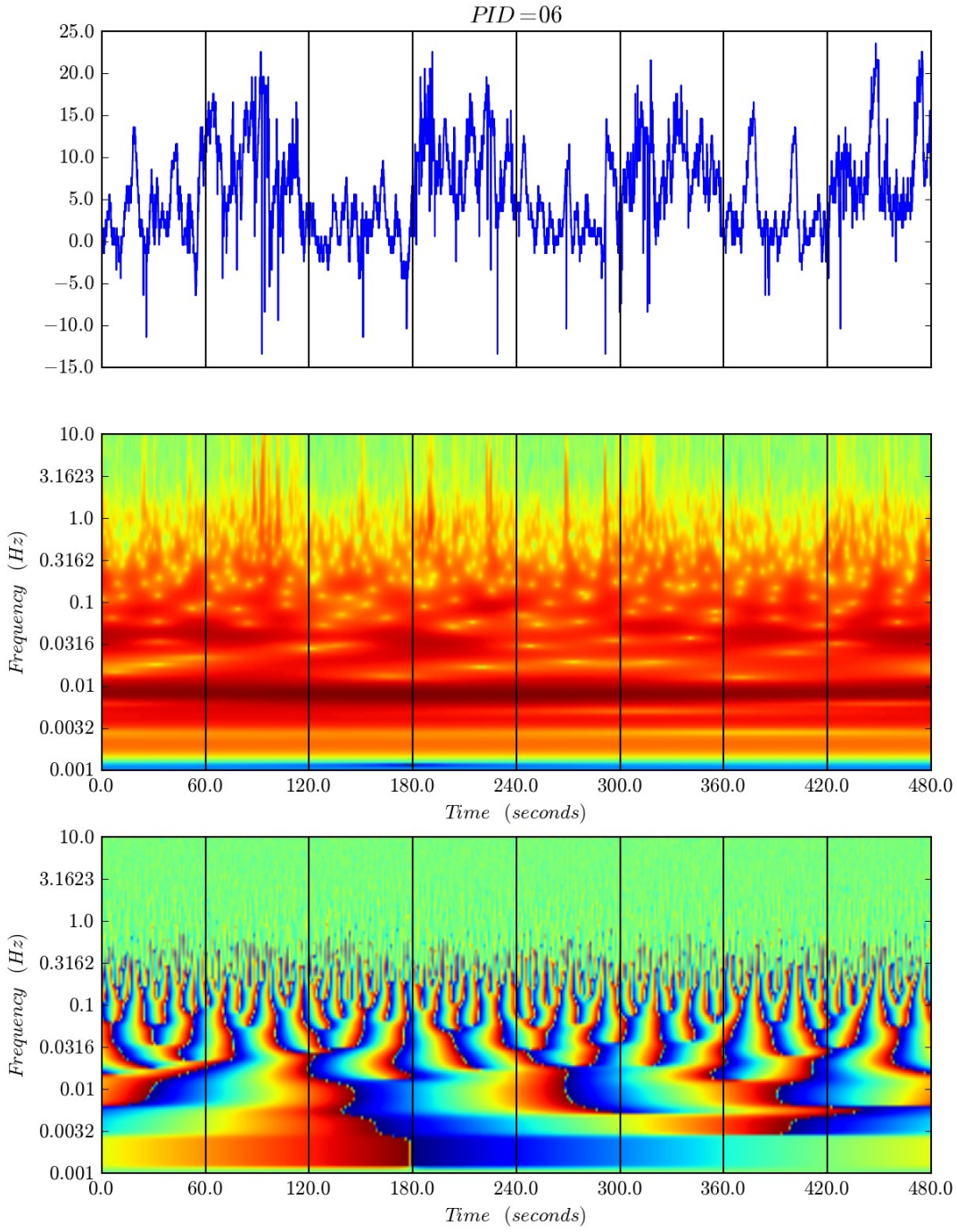


Figure 5.3.8 *Participant 7 raw pupil diameter, CWT scalogram, and CWT phaseogram.*

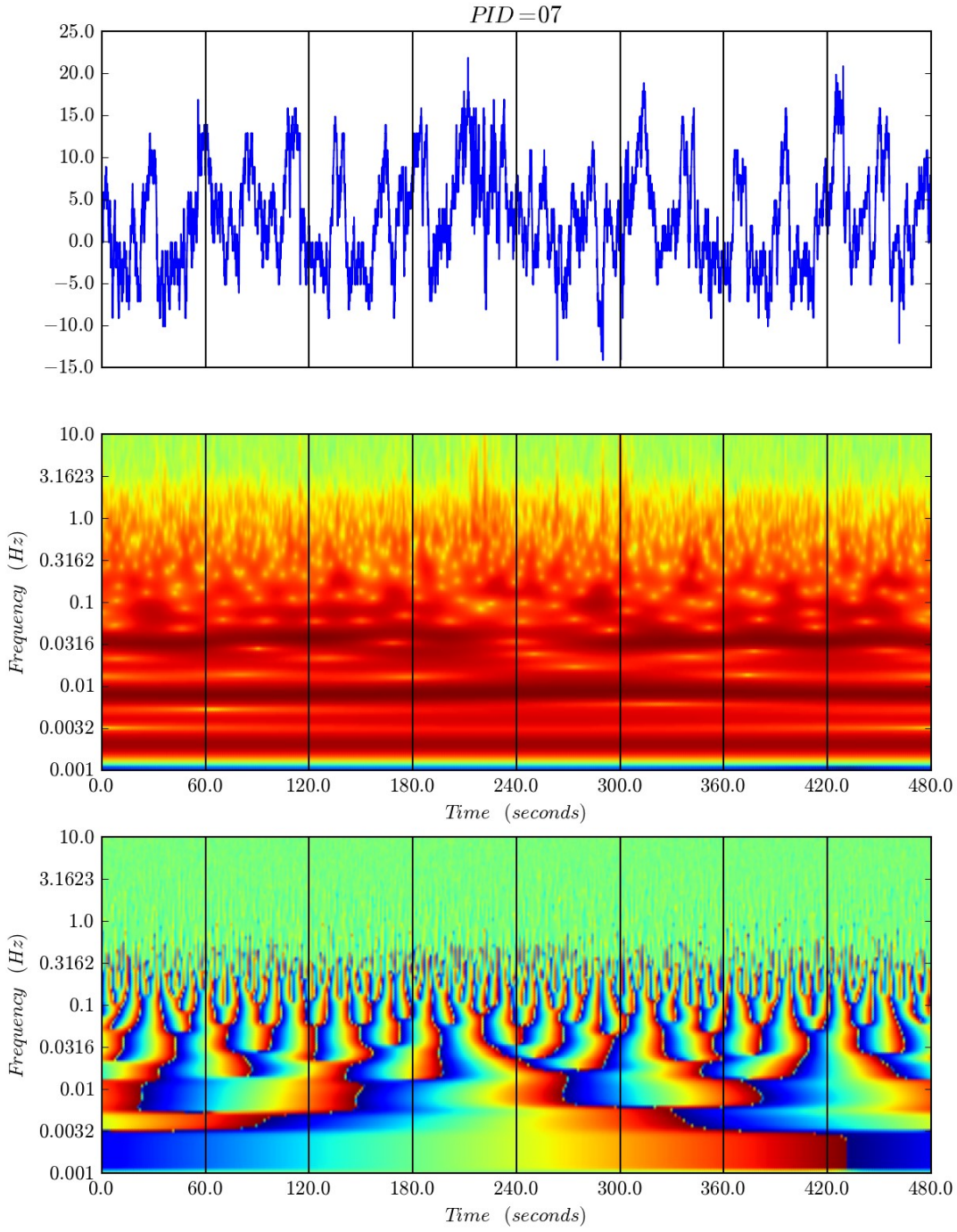


Figure 5.3.9 *Participant 8 raw pupil diameter, CWT scalogram, and CWT phaseogram.*

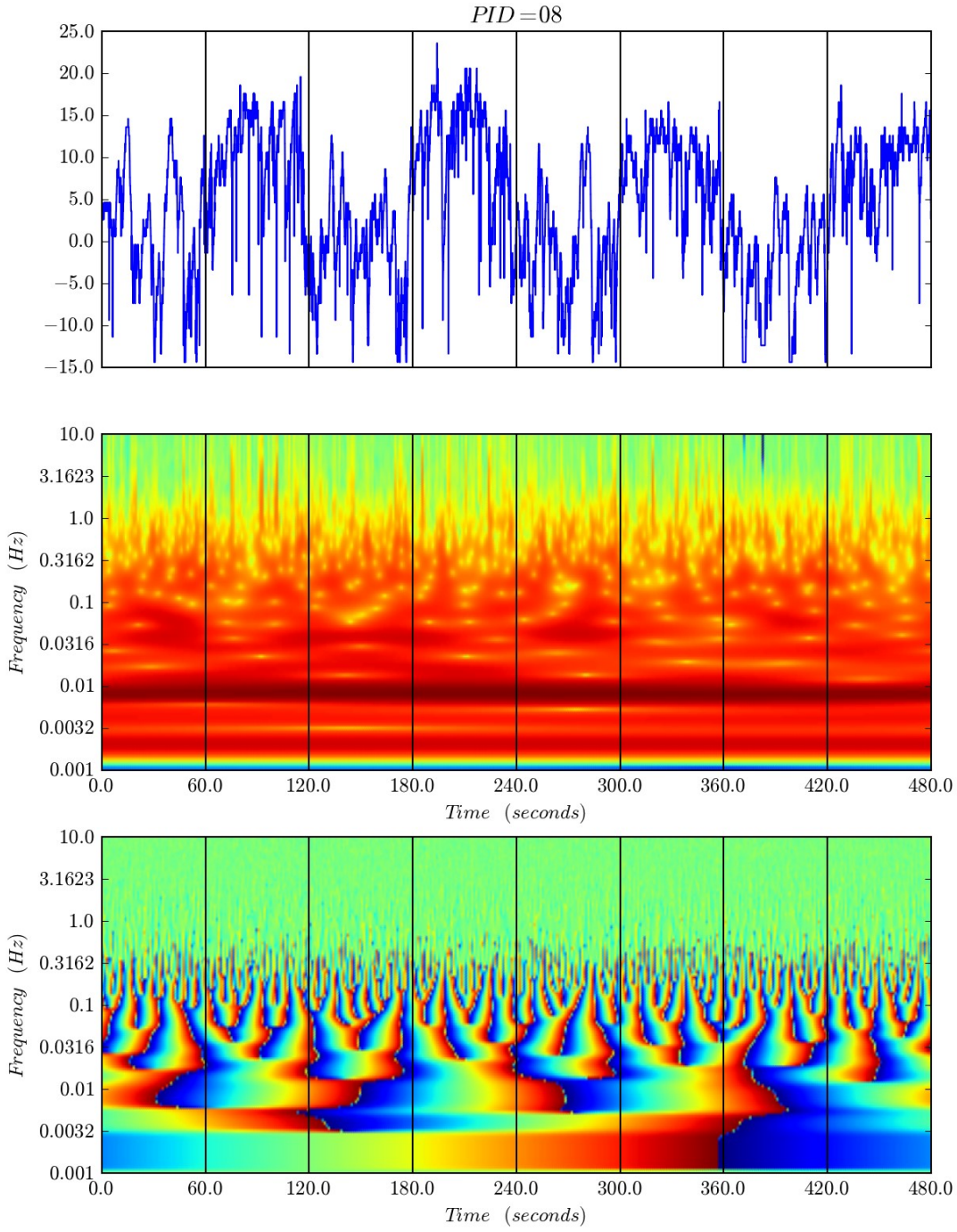


Figure 5.3.10 *Participant 9 raw pupil diameter, CWT scalogram, and CWT phaseogram.*

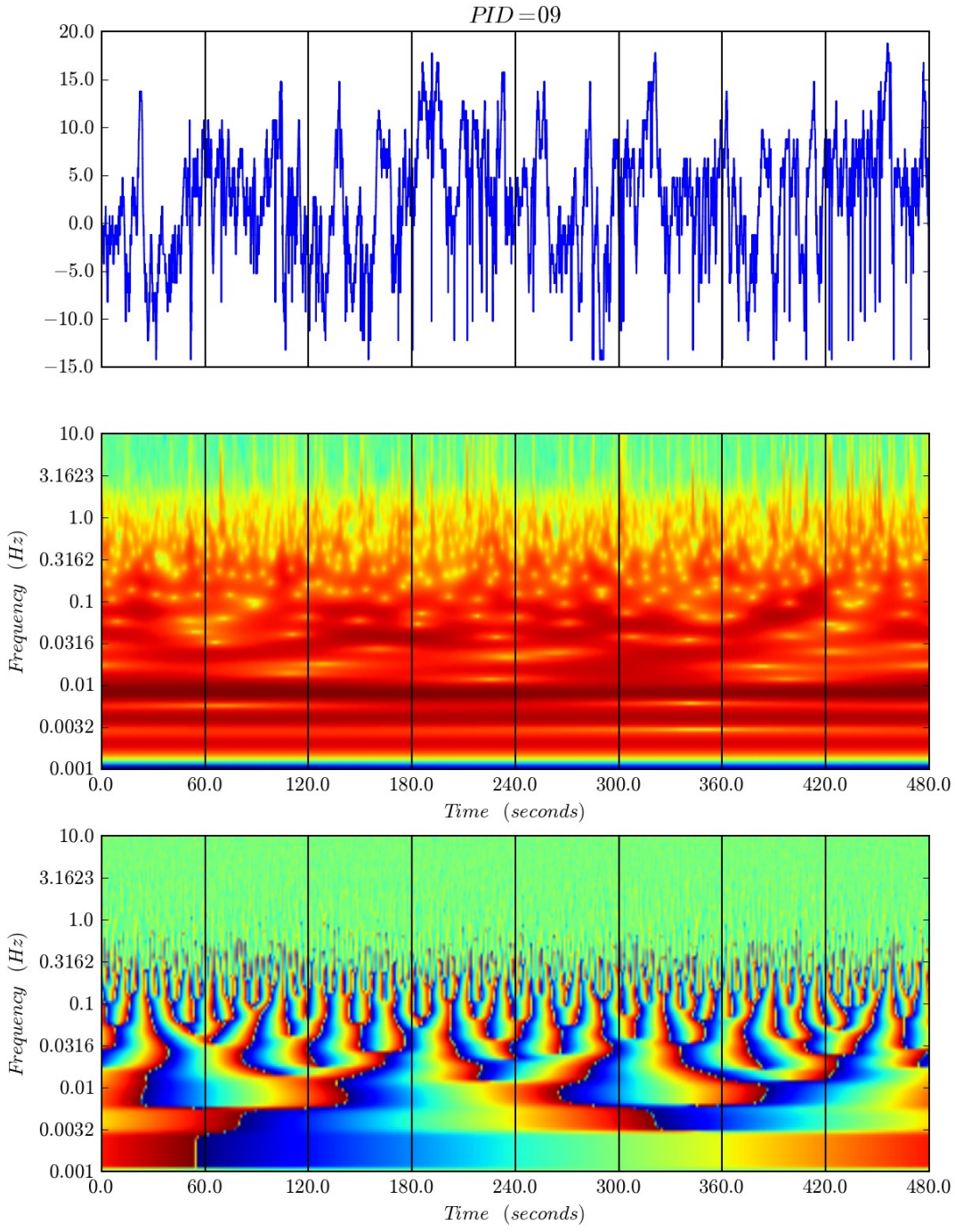
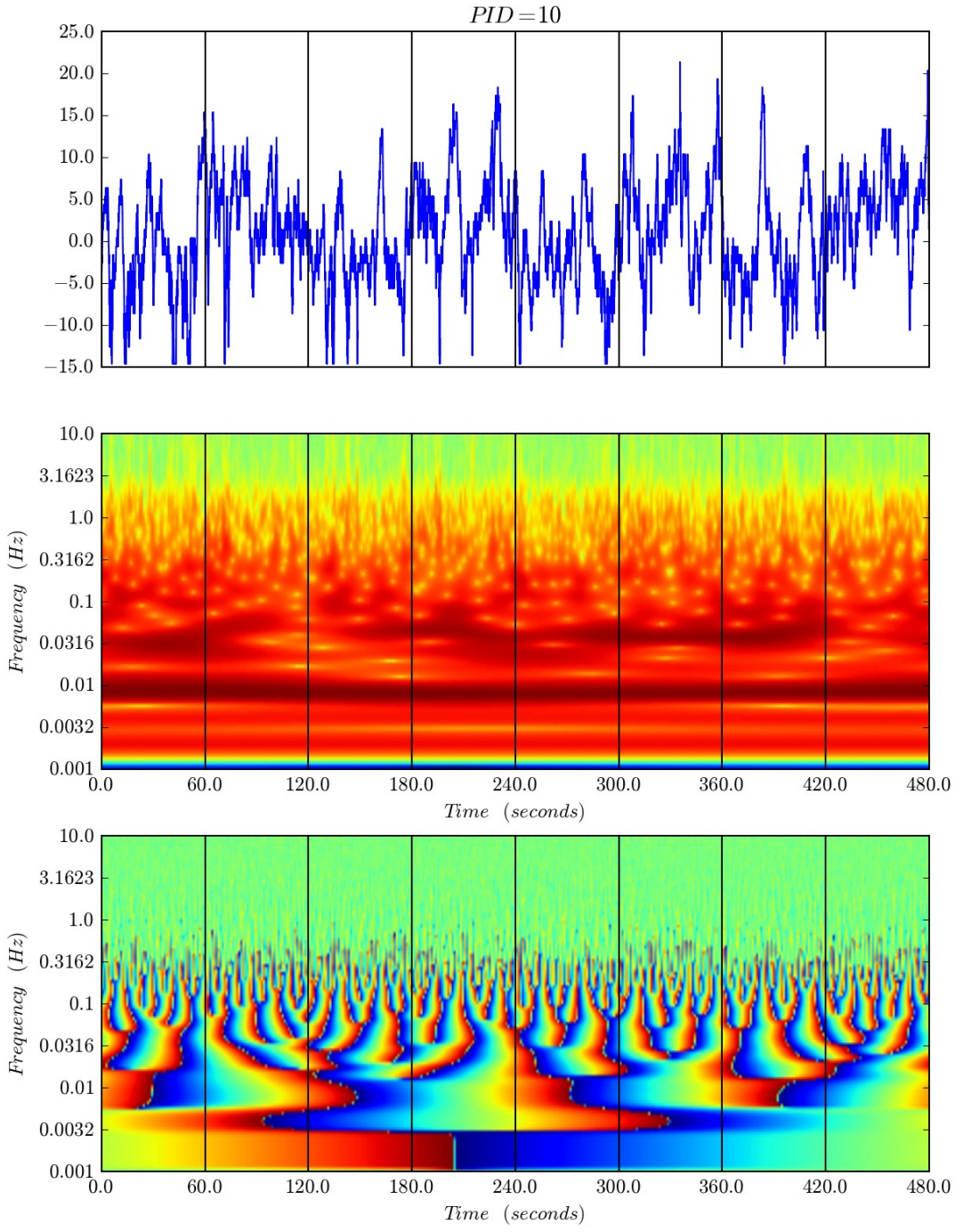


Figure 5.3.11 *Participant 10 raw pupil diameter, CWT scalogram, and CWT phaseogram.*



With genetic programming models tend to grow as the programs evolve. Parsimony pressure was added as a preventative measure against code growth and also to try and force GP to yield smaller more succinct programs. Without some parsimony pressure programs will grow exponentially which can lead to potential memory paging or availability problems.

The goodness of fit of model estimates were manually optimized through four design iterations. These iterations are summarized in Appendix 5.3.D. A summary of final model parameters is provided in Table 5.3.1. For each participant this regressor was ran eight times.

5.3.2.3 Linear Discriminant Analysis. To compare the Iteration IV GP models to LDA, random subsets of 1024 random points were generated and regressed on the 20 variables available to the GP. For each participant LDA was run 100 times. On the test data, a paired samples t-test was used to compare the best r-squares for each participant found by GP to the best r-squares found by LDA. Comparisons were made on the best solutions because GP is conceptually more of a "shotgun" approach compared to LDA. The r-squared distributions obtained by multiple regression are normal with far less variability between runs which implies LDA is fairly robust to the subset of time points but this also results in its top performance being somewhat restricted. Due to the amount of randomness in GP there are no guarantees of converging on a "good" solution and in practice many mediocre and poor solutions often arise. Despite this fault, GP can sometimes come up with solutions that are several times better than average. Here I was interested in top performance rather than average performance. In a real world setting the training would most likely take place offline so time is not a critical factor and several potential solutions could be evaluated before they are put into practice.

5.3.3 *Tracking Error Model Results*

To verify that the rotated mappings did in fact hamper tracking performance (increase tracking error) a 2 x 4 Analysis of Variance (ANOVA) evaluated the effects of mapping (normal vs. rotated) and block (1-4) on mean tracking error. As expected a significant main effect of mapping

was found [F(1, 9)=133.682, p<.001] indicating tracking error was higher under the rotated mappings (see Figure 5.3.12). In addition, a main effect of block [F(3,27)=9.247, p=.003, ε=0.559] as well as a mapping by block interaction [F(3, 27)=7.221, p=.007, ε=0.602] were found. This interaction suggests that virtually no improvement in tracking performance occurs for the normal control mapping across the four blocks, but most participants improve by about 35% on average with the rotated mappings. One exception was Participant 3, who showed about 6 degrees of mean error across all four blocks; the remaining 9 participants showed improvement from block 1 to block 4.

A matched pairs t-test between the r-squared values obtained from the best training results from symbolic regression and LDA revealed that symbolic regression accounted for significantly more variability in the training set [$t(9) = 6.450$, $p < 0.001$] On average GP accounted for 26% more variability (see Figure 5.3.13). Essentially, the more flexible GP approach is able to account for nuances in the training data better than the more constrained linear approach. A matched-pairs t-test on the test epochs revealed that GP was able to significantly outperform multiple linear regression in predicting test performance by 4% [$t(9) = 1.86$, $p < 0.05$].

5.3.4 *Control Mapping State Classification Models*

5.3.4.1 Genetic Programming. When built as a classifier the GP previously described converges quickly and drastically over-fits the training data. Many programs were able to account for every single case of the training data. To deter over-fitting a steady-state symbolic regression without ALPS was used. The population size was set to 1,000, and each simulation was run through only 10,000 iterations. Since fitness reflects the goodness-of-fit for a randomly chosen 1024 subset of time points in the training set and does not necessarily summarize general performance to the training data or the testing data (see Figure 5.3.14), the "best" models for each run were selected by choosing the models which had high accuracy for both training and testing and were of human interpretable sizes. Selecting the "best" models was part science but admittedly part art.

Figure 5.3.12 *Mapping by Block Interaction on mean tracking error. Plot depicts the mapping by block interaction over all ten participants. Error bars represent +/- one standard error of the marginal means. Tracking for the normal mapping does not improve from over the course of the experiment. Tracking for the rotated mapping is significantly higher over all four blocks (as implied by the error bars; no post hoc comparisons were made) but improves over the course of the experiment.*

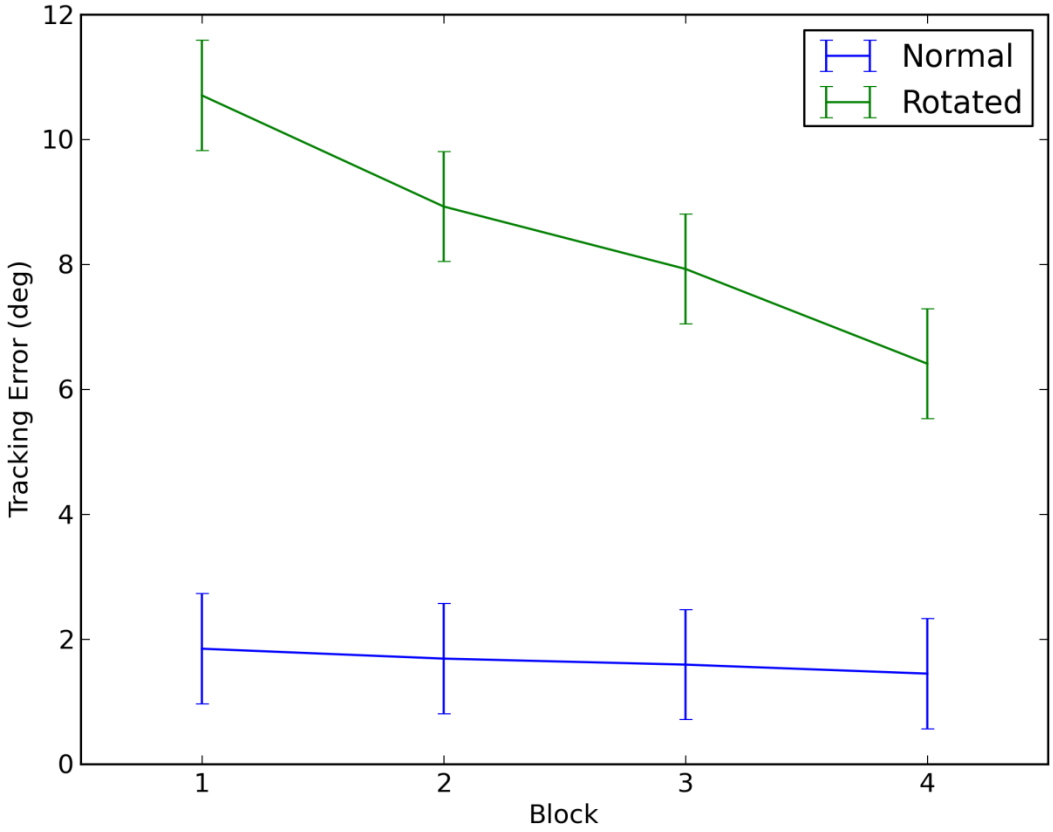
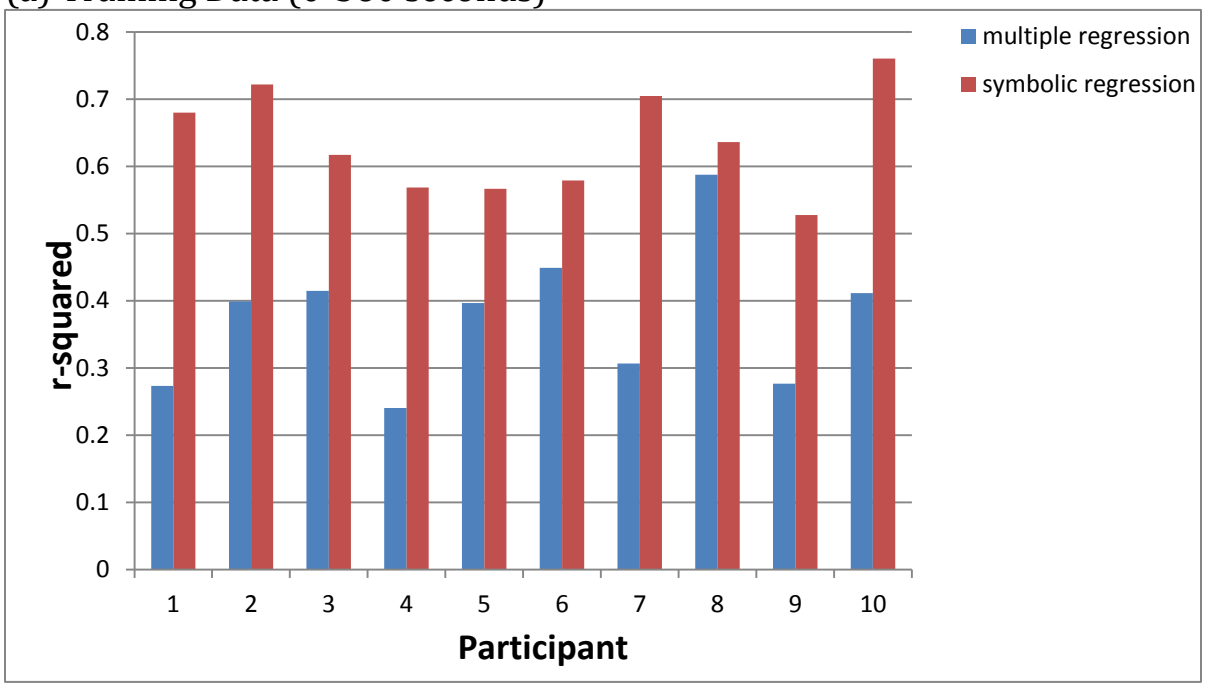


Figure 5.3.13 *r-squares obtained from the training set and the test set.*

(a) Training Data (0-360 seconds)



(a) Test Data (360-480 seconds)

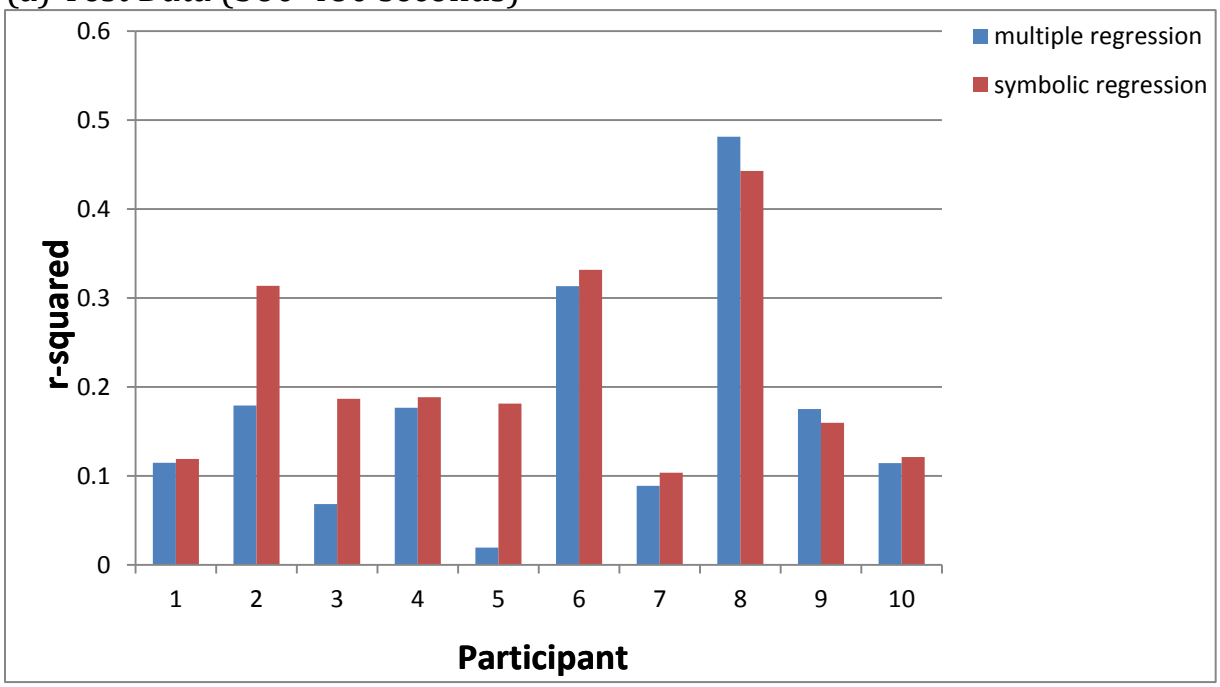
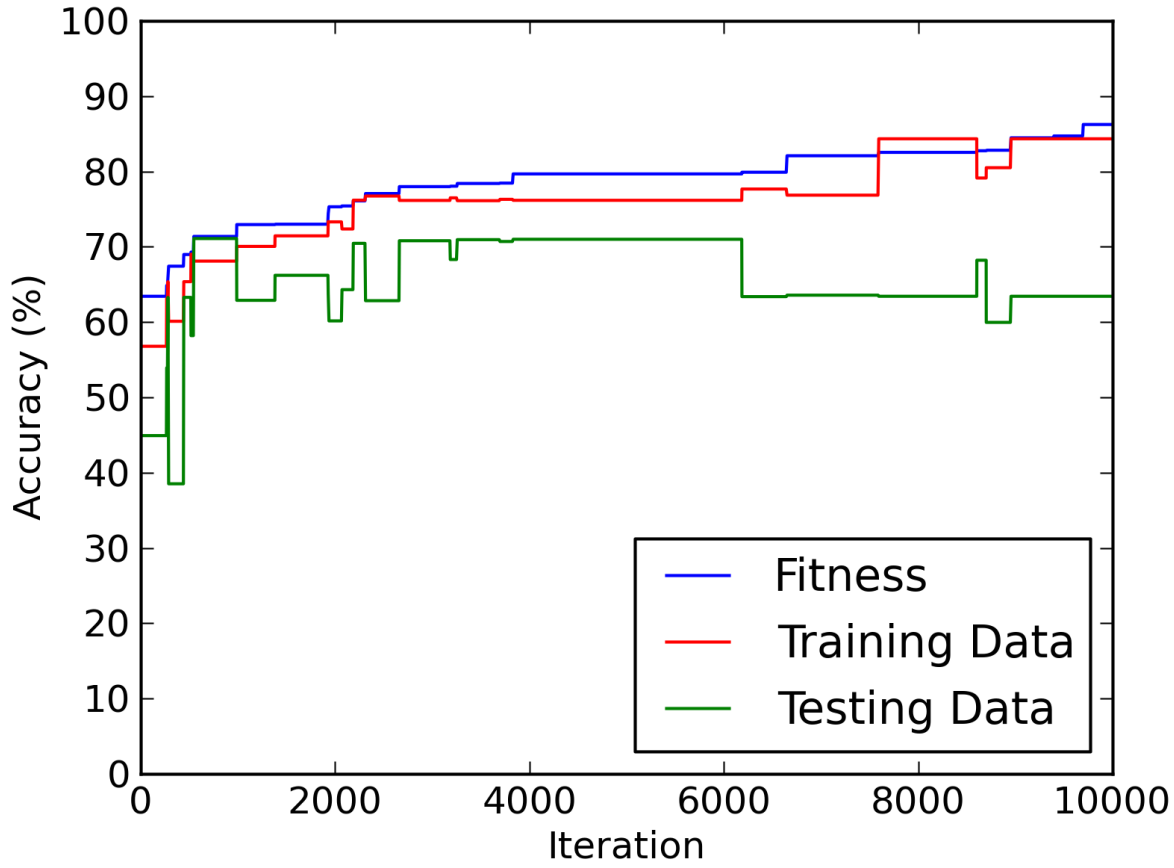


Figure 5.3.14 *Accuracy vs. Goodness-of-fit.The "best" model may not be the model with the highest fitness. Models with high fitness may perform worse on the testing data than models with lower fitness.*



Since the testing data is used to select models (but not to train them) future studies plan on incorporating a second testing set to validate model performance before evaluating test predictions. Dividing the 480 second trial in to training (50%), validation (25%), and testing (25%) sets (as suggested by Marshall, 2007) did not provide enough data to effectively capture the models performance. Subsequent experiments will use longer trials to alleviate this problem.

5.3.4.2 Linear Discriminant Analysis. To compare the performance of GP to more traditional techniques dichotomous Linear Discriminate Analysis (LDA) classifier were generated for each participant. The classifiers were trained using 1024 subsets from the training data in the manner described in the tracking model.

5.3.5 *Control Mapping State Classification Model Results*

Table 5.3.3 lists the classification accuracies for GP and LDA on the training and test data. On the training data LDA averaged 67.1% accuracy across participants compared to 80.6% with GP [$t(9) = 3.06$, $p < 0.01$]. On the testing data LDA averaged 62.5% compared to 77.2% [$t(9) = 3.42$, $p < 0.01$]. It should also be noted that with LDA three of the participants are at chance levels with the training data, and four with the test data. In all but one case GP correctly classified the control mapping at least 2/3 of the time.

5.3.6 *Conclusions and Discussion*

To date, few researchers have employed GP within the domain of augmented cognition. The use of GP in Augmented Cognition is not unprecedented (Simoni, 2008), but is not common. Here evidence has been provided that suggests symbolic regression can be used to predict human workload and can do so significantly better than simple linear techniques. On tracking error symbolic regression performed better with eight out of ten participants. When GP did worse on the test data it did so by 4% with participant 8 and 1.5% with participant 9 (see Table 5.3.3). This suggests that GP has more potential for improving results than for making them worse. Overall

performance is generally better even at the individual level. Furthermore, the predictions produced LDA are noisy and often have transients that are well outside the range of actual tracking error. In contrast, the predictions produces by GP are surprisingly well behaved (see Figure 5.3.15).

Both symbolic regression and LDA commonly used some input vectors more than others. Almost all the models used the lowest two gsr coefficients (GSR_COEFF0, and 1) and the lowest two pupil diameter coefficients (PUP_COEFF0, and 1). The raw SC and pupil diameter values also contributed more to the final value. For some participants the faster pupil coefficients (5, 6, and 7) also contributed to the tracking error prediction. Figure 5.3.16 through Figure 5.3.25 compare the power spectral density of the participant's pupil diameter between normal and rotated mappings.

At this point not much has been done to optimize or examine the size of the solutions generated by symbolic regression. Although, it is worth pointing out that the over the course of the 80 runs from Iteration IV the average solution size was 49 (including Keizer's scaled symbolic coefficients and addition and multiplication terminals) nodes and the best test solutions have an average size of 45. In comparison, the multiple regression models would have an equivalent node size of 81 if represented as a tree structure which suggests that symbolic regression is not better just because it has more parameters. In the case of subject 8 the best symbolic regression model used only 13 nodes to predict 98.5% of the variability predicted by the 81 nodes of multiple regression (see Figure 5.3.26). The 13 node solution is:

```
(+ (7.728) (* (0.460) (+ (+ (+ (+ (PUP_COEFF0) (PUP_RAW)) (PUP_RAW)) (PUP_RAW)) (PUP_RAW))))
```

While the performance of the GPs discussed here is noteworthy, even greater performance may be attainable through further application of GP theory and technique. For instance, applying cooperative co-evolution within multi-agent systems has been found to increase the performance of GP classifiers (Soule & Heckendorn, 2008).

Figure 5.3.15 *GP vs. LDA tracking error predictions. The models in the above panels (top: multiple regression, bottom: symbolic regression) account for about the same amount of variability of tracking error, but as the reader can see the multiple regression predictions are noisy and not as bounded as the symbolic regression model.*

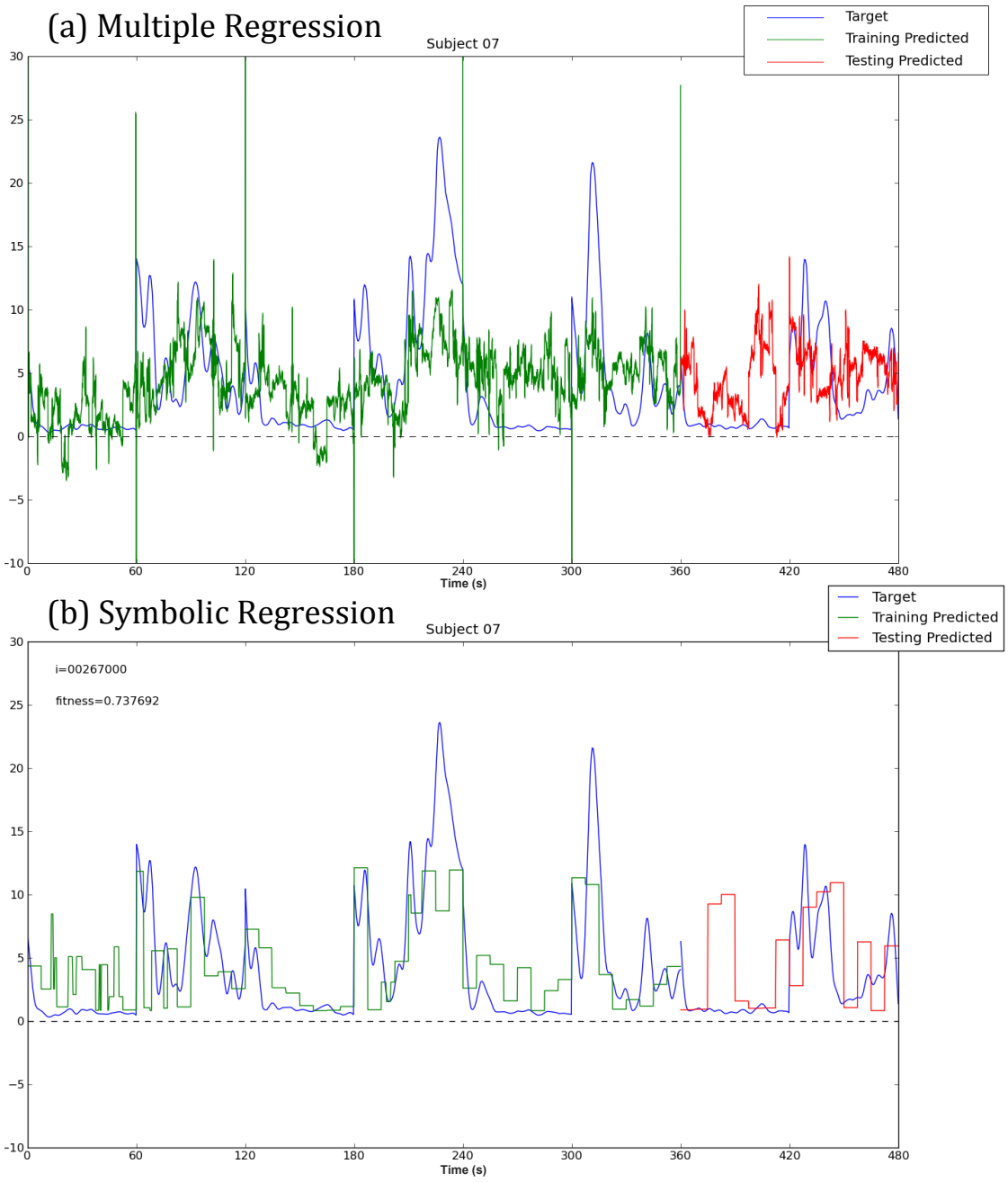


Figure 5.3.16 *Participant 1 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

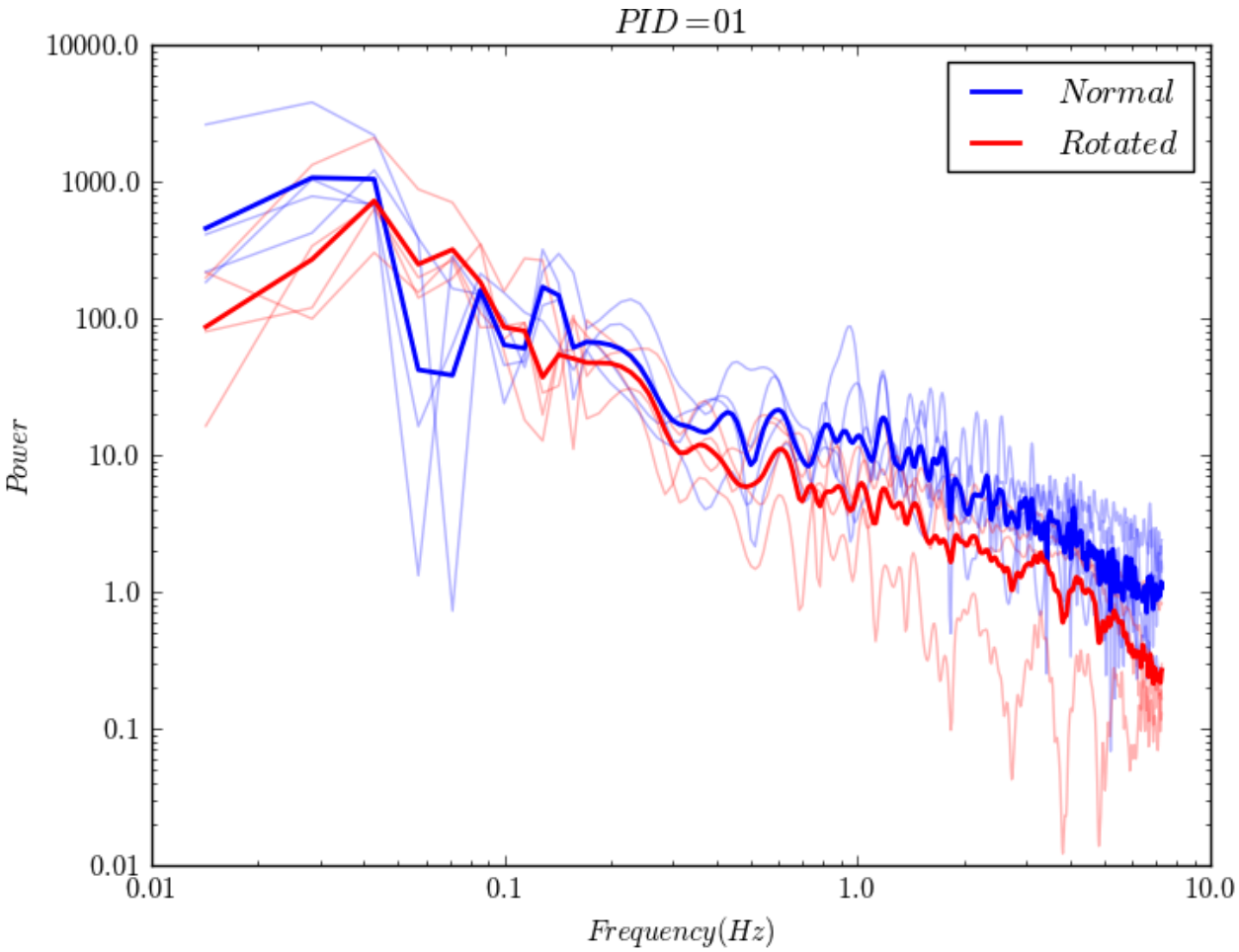


Figure 5.3.17 *Participant 2 normal versus rotated power spectral density.Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

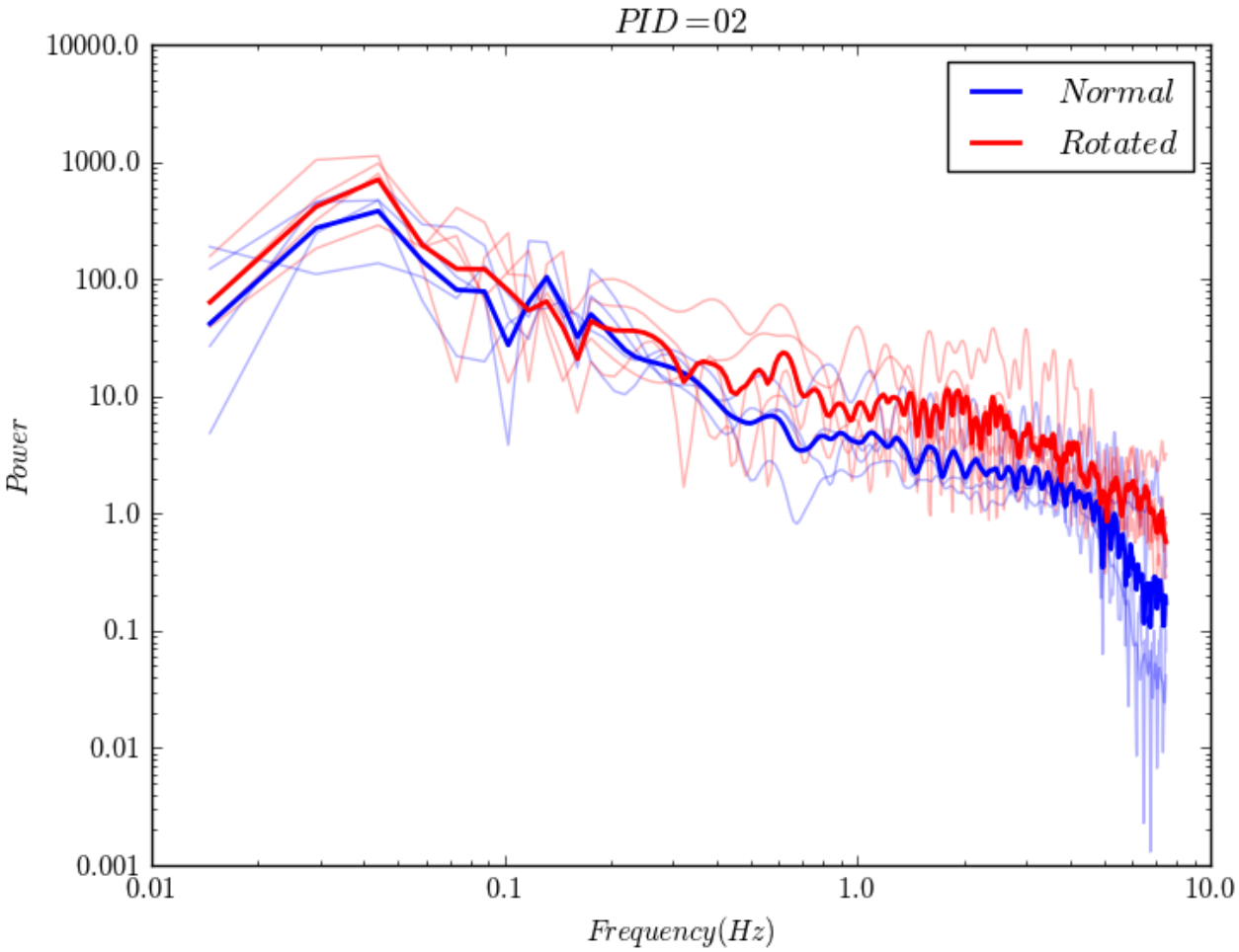


Figure 5.3.18 *Participant 3 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

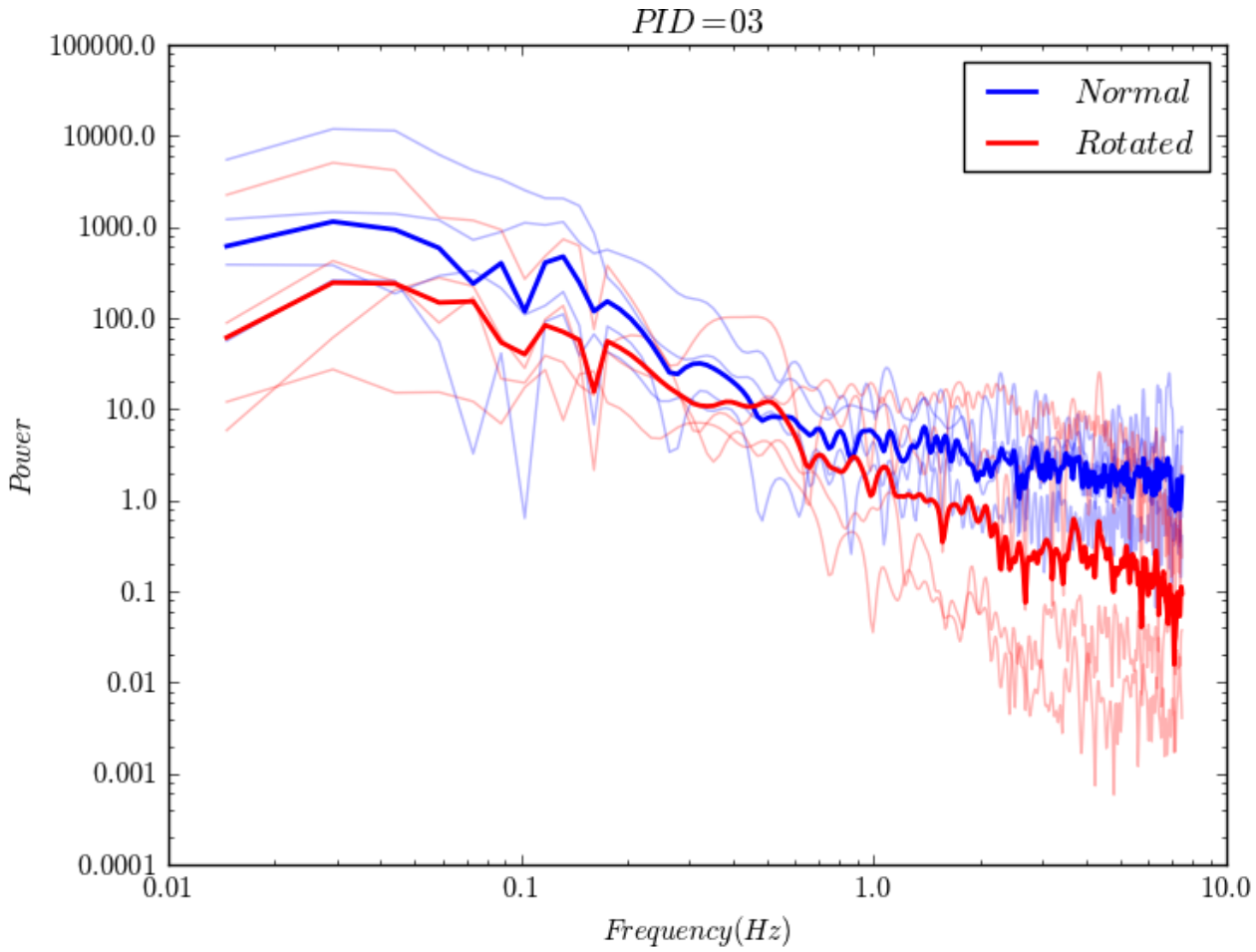


Figure 5.3.19 *Participant 4 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

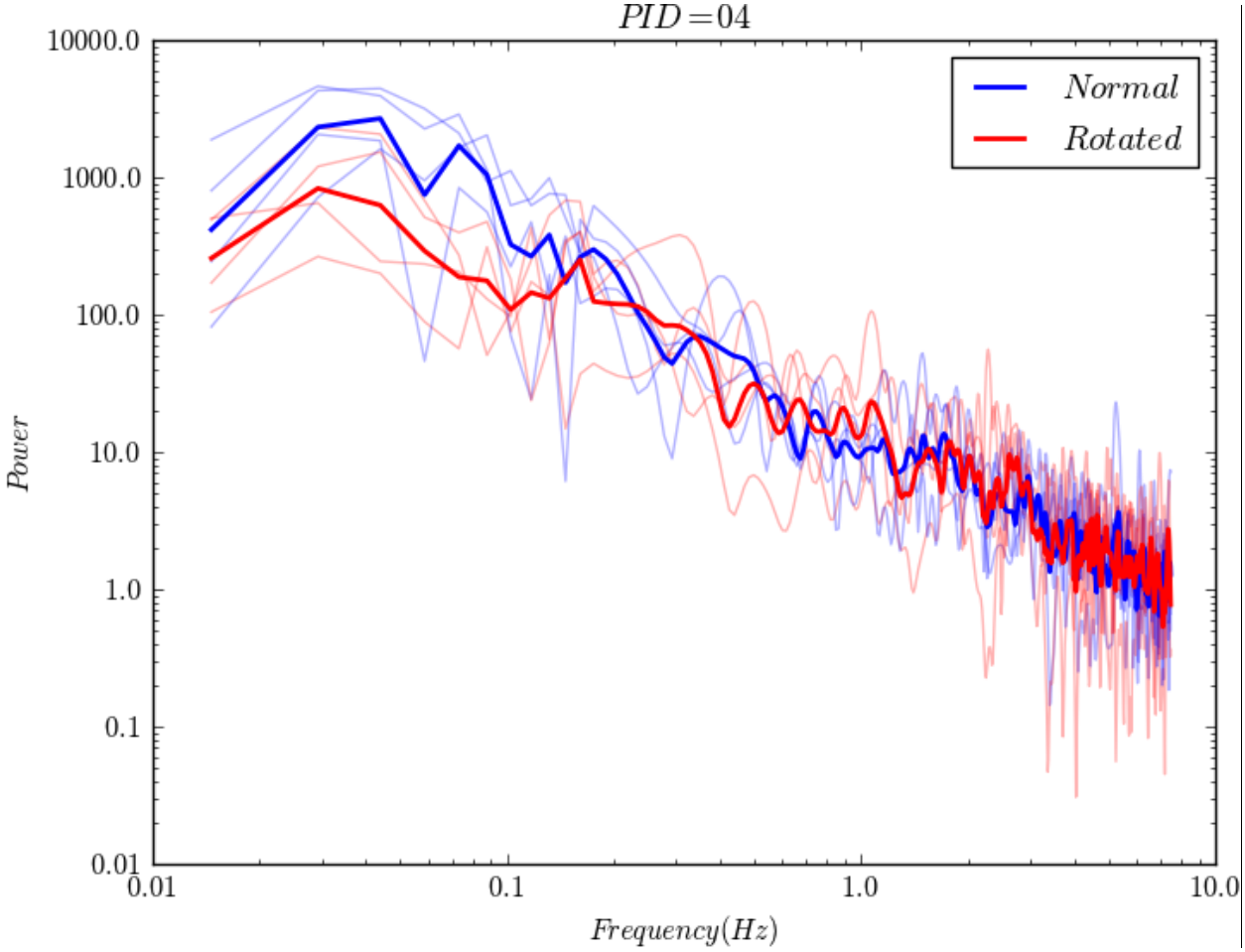


Figure 5.3.20 *Participant 5 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

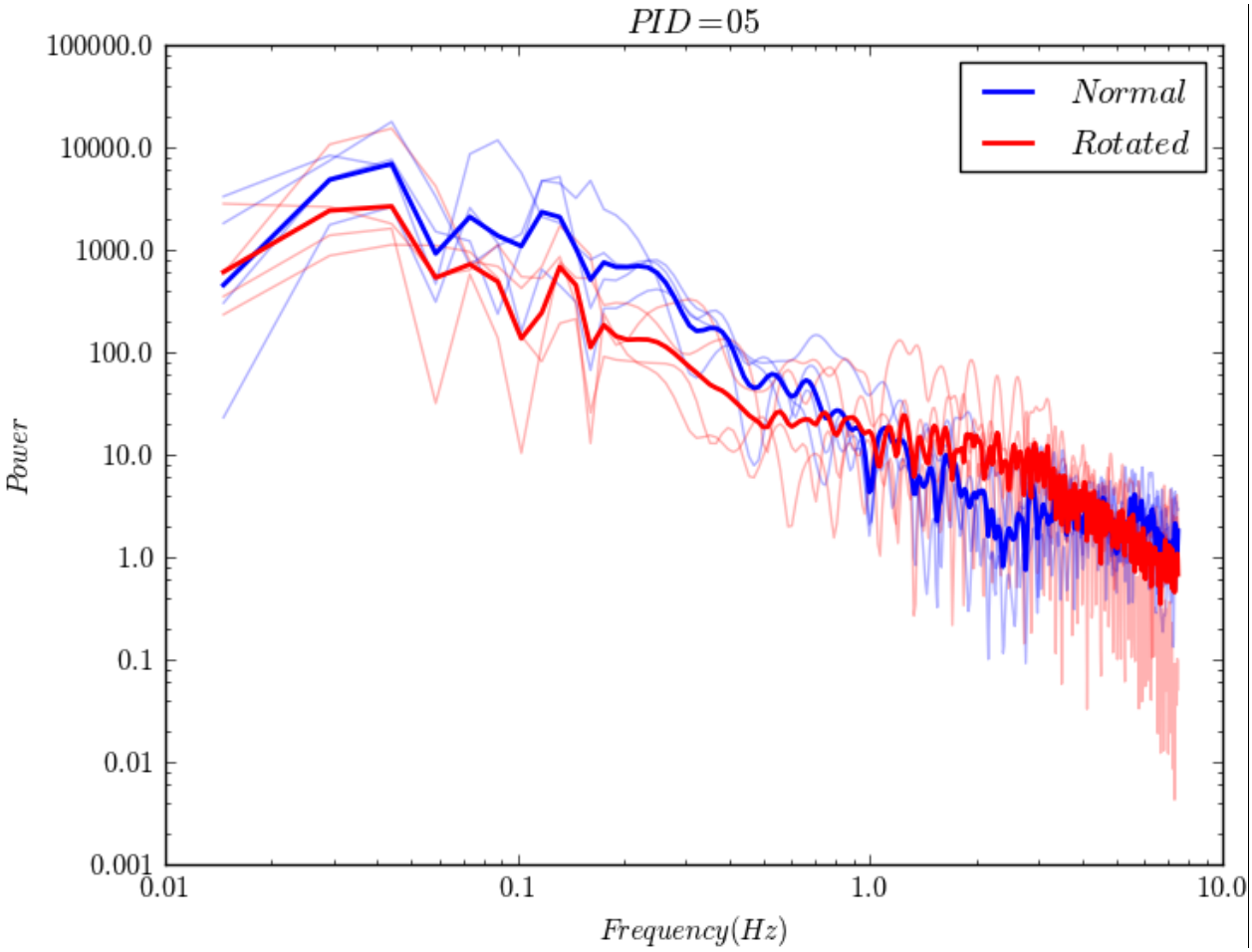


Figure 5.3.21 *Participant 6 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

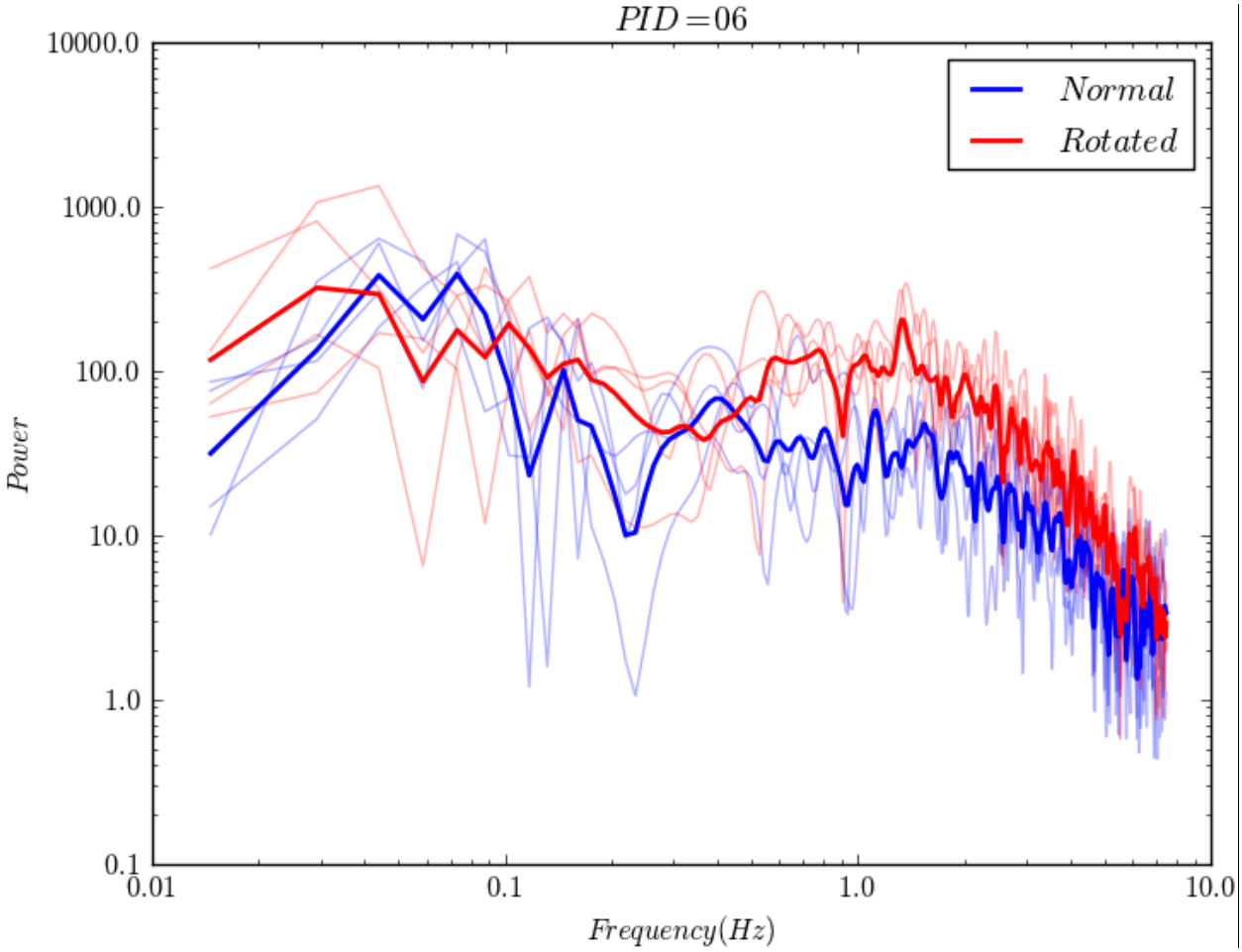


Figure 5.3.22 *Participant 7 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

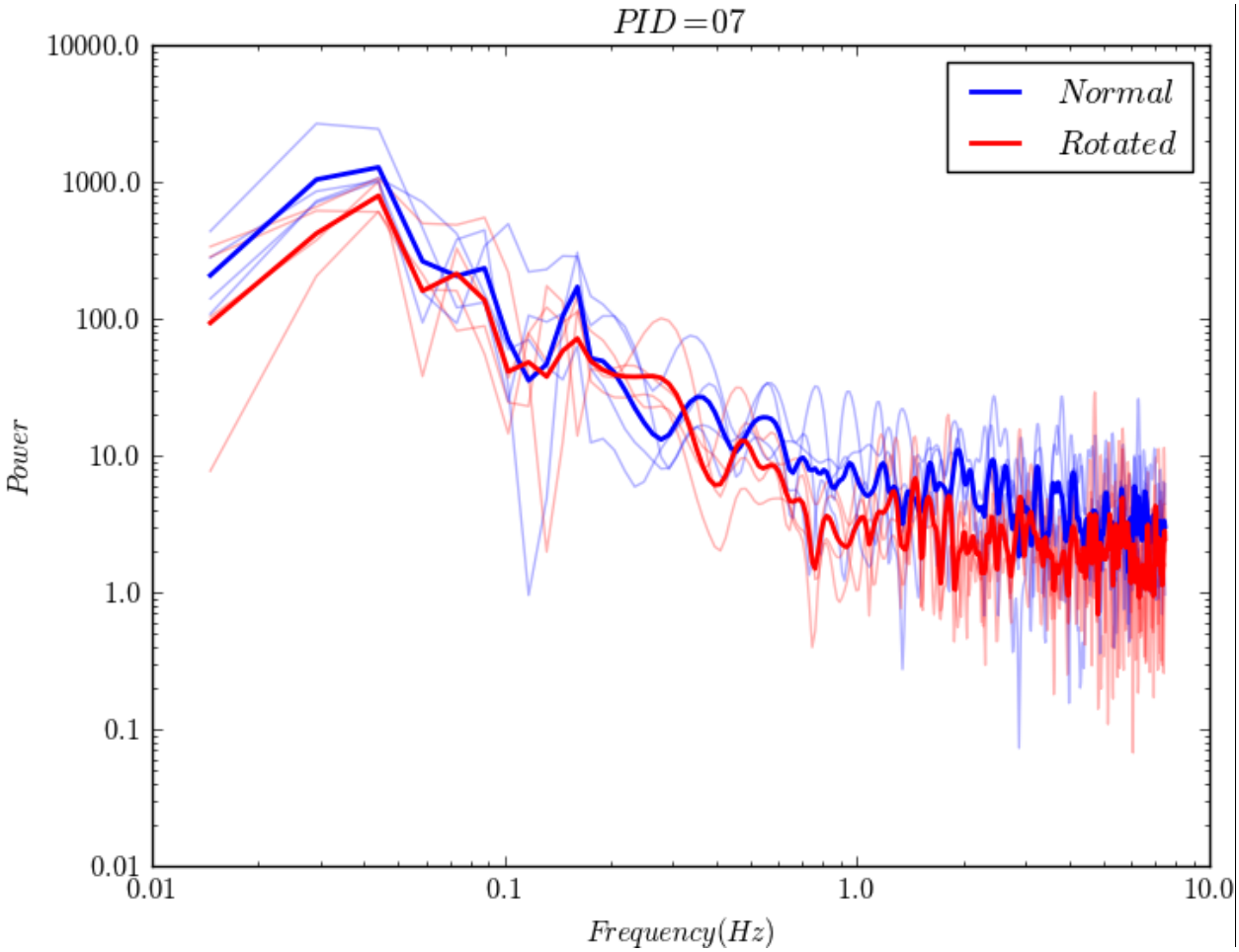


Figure 5.3.23 *Participant 8 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

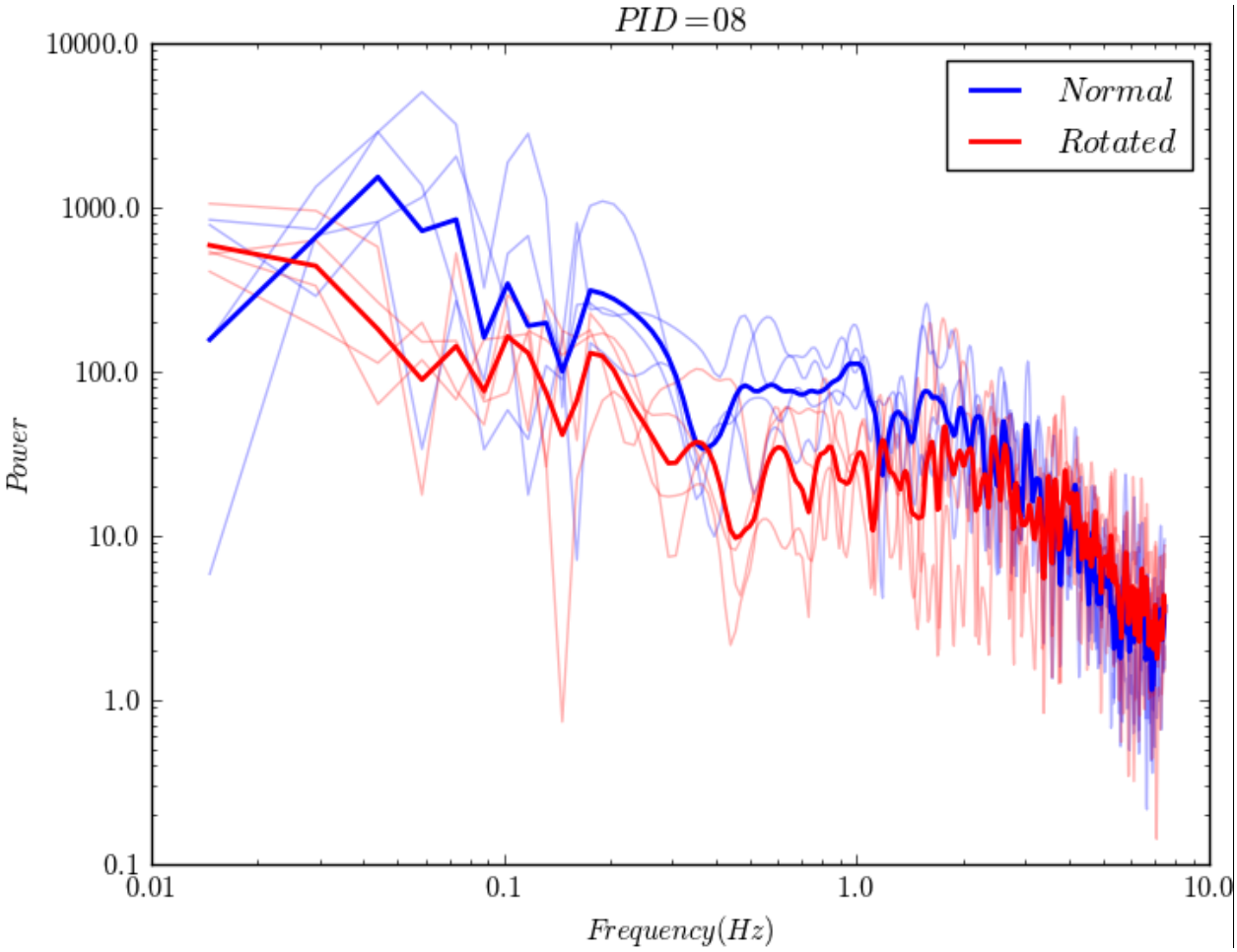


Figure 5.3.24 *Participant 9 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

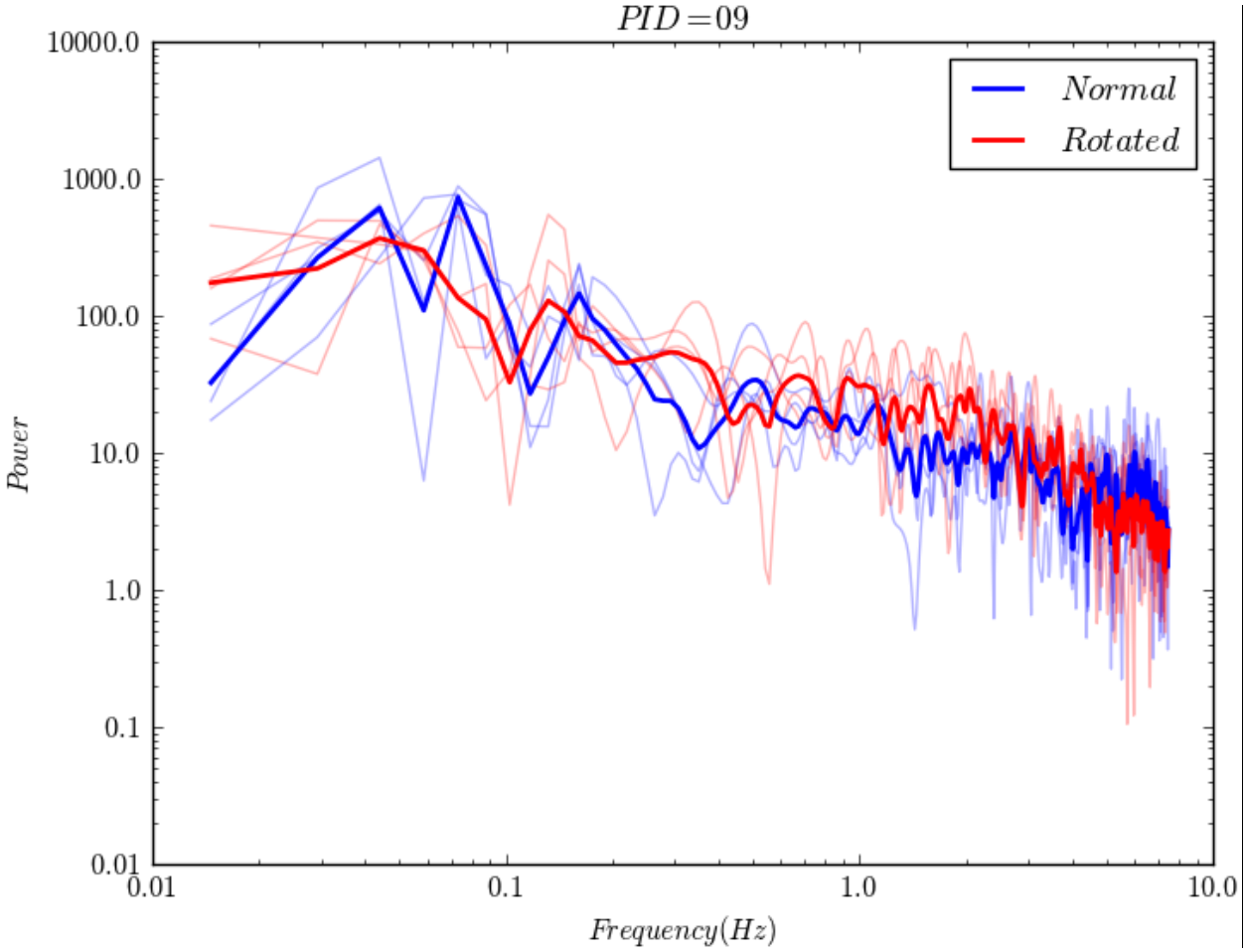


Figure 5.3.25 *Participant 10 normal versus rotated power spectral density. Bold traces represent averages across four blocks. The lighter traces represent spectra in each of the four blocks.*

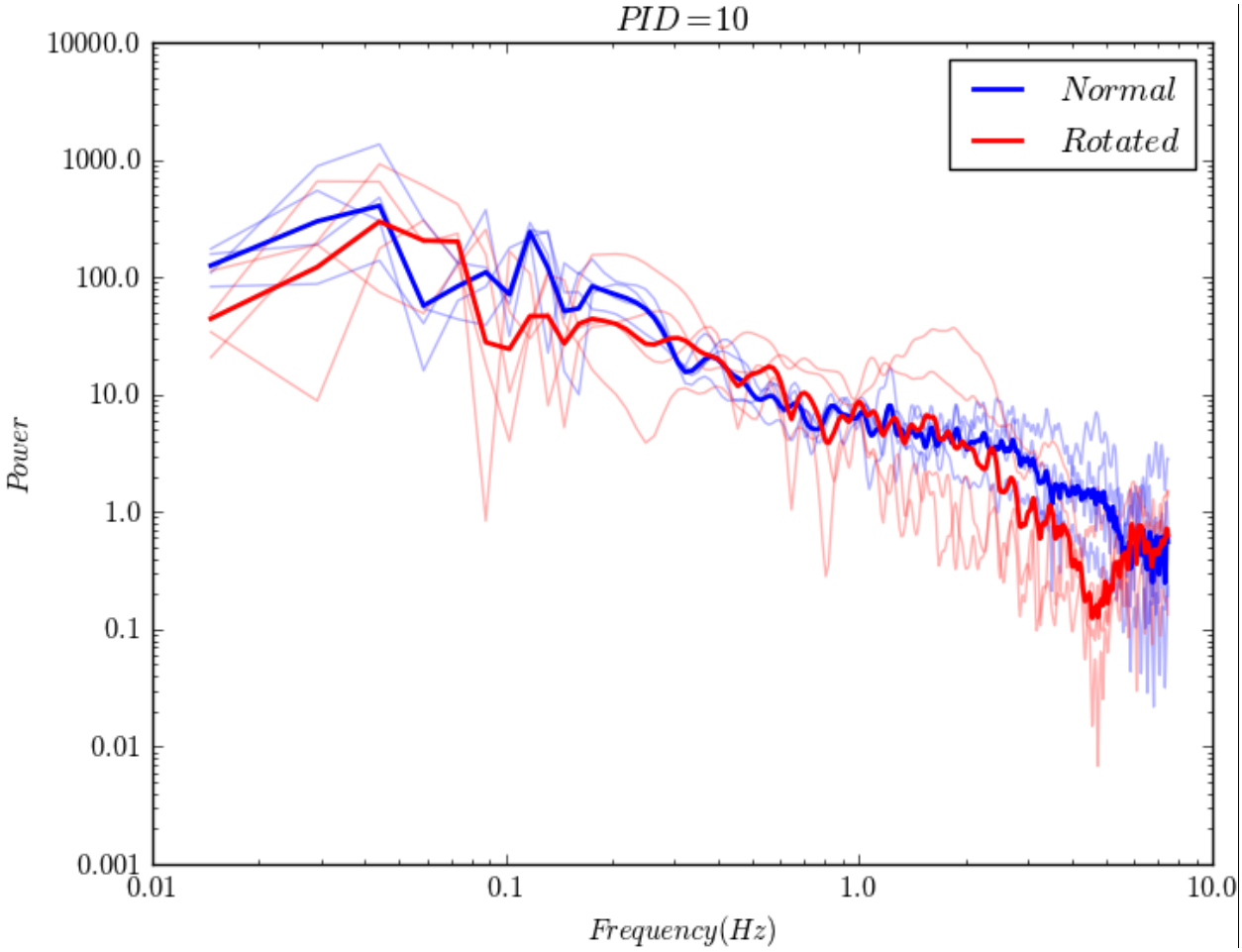
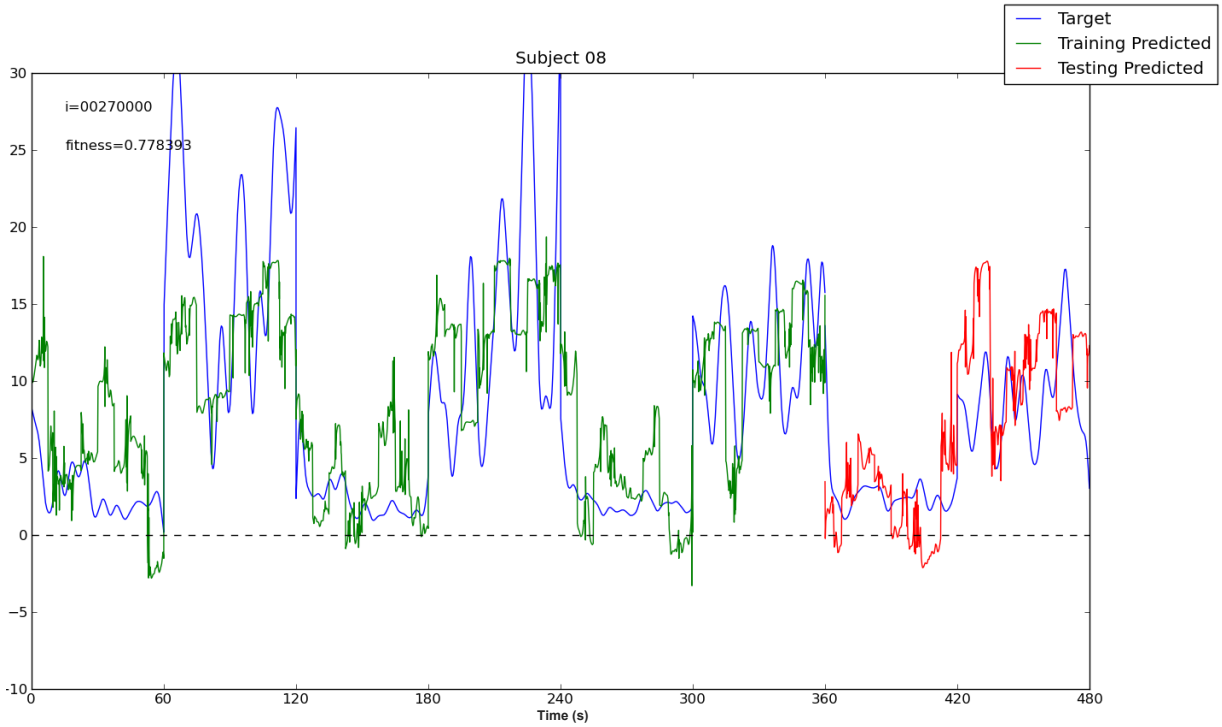


Figure 5.3.26 *Tracking error predictions for Participant 8. Run 6 for participant 8 had a r^2 for the training data of 0.491 and a r^2 for the test data of 0.443 using only 2 physiological parameters and 13 nodes. While this solution is atypically good it does illustrate the potential of symbolic regression.*

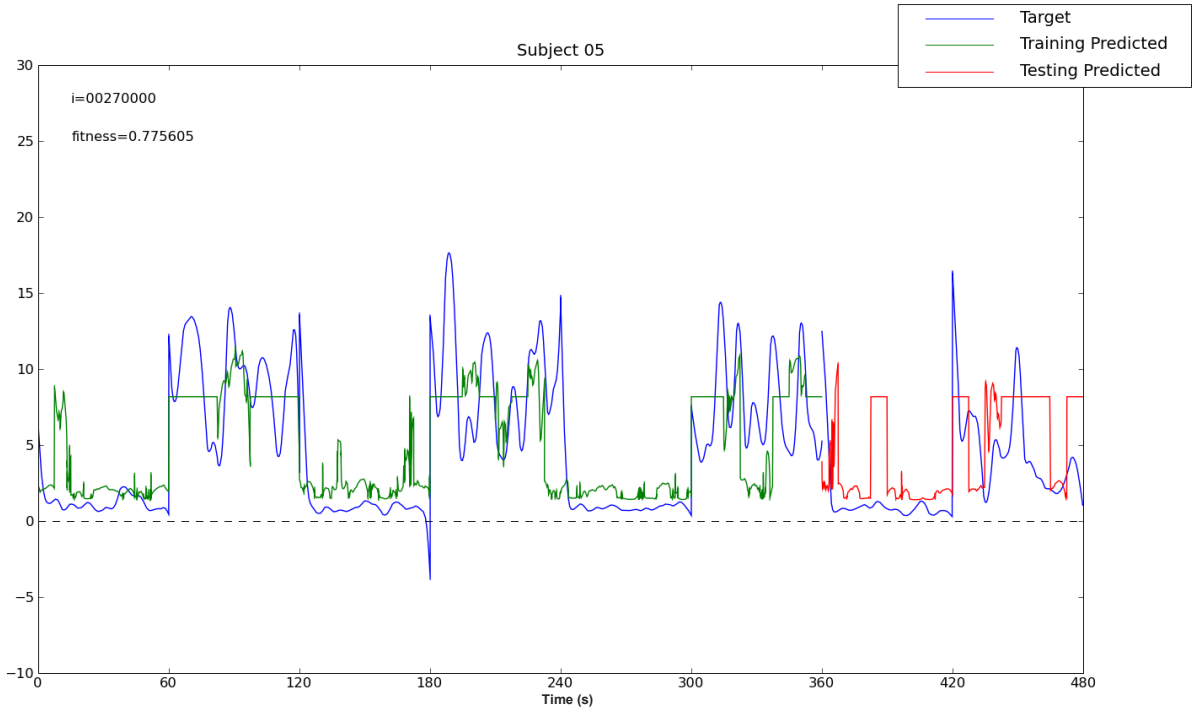


The r-squared values between the training and test data for GP suggest that the model might still be over fitting the test data. Future work should examine using smaller subsets or possible shorter runs to hedge over fitting. The fact that the multiple r-squared values for the test data is also lower (by about .1) than the training data also suggests that perhaps there are discrepancies in the physiological markers between 0 and 360 seconds and 360 and 480 seconds (see Figure 5.3.27). Having more data to sample from would help remedy this problem. Future fitting should also try to have the participants go through an initial task training period before training data is collected. In this experiment participant's 2, 4, and 10 have large transient in tracking error after the control mappings switch the first time at the 60 second mark but do not have errors of equal magnitude latter throughout the rest of the trial. Because the Keizer's Scaled Error reflects the squared tracking error the model is severely penalized if it doesn't predict from 60 to 65 seconds, yet that time period only accounts for slightly more than 1% of the trial.

5.3.6.1 Limitations. While this study demonstrates that GP modeling of PD and SC may be useful for estimating primary task performance and task difficulty, one important caveat is that my task, with its abrupt changes in control mappings, caused large changes in primary task performance coincident with changes in the physiological measures. The purpose of AC is to be able to measure increases mental workload with physiological measures before primary task performance is affected. The following proposed experiment is designed to assess the predictive utility of GP modeling of PD and SC with tasks in which cognitive demand changes more subtly so that changes in the physiological measures occur prior to changes in primary task performance.

5.3.6.2 Methodological Considerations for Future Research. Here I am making the assumption that I can manipulate workload by manipulating the control dynamics of the tracking task and that those manipulations subsequently affect tracking performance. Experiment 3 serves as a proof-of-concept that the relevant information is contained

Figure 5.3.27 *Tracking error predictions for Participant 5. The best performing prediction on the test data over 8 runs for participant 5 is shown above. The green line is showing the predicted tracking error from a single run values over the training data. The red line extrapolates the model to the test data. The training data fits well but does not generalize to the test data as evidenced by the portion between 120 and 180 seconds and the portion between 420 and 450 seconds.*



within the physiological measures and that wavelet analysis with symbolic regression can resolve some of that information. While task performance is likely correlated with workload obtaining direct estimates of workload would be more useful to the implementation of augmented cognition systems. I hypothesize that physiological indications of workload might arise before decrements in task performance.

5.3.6.3 Modeling Considerations for Future Research. Based on increased knowledge of the spectral analysis techniques discussed in Chapter 2 I have identified some analytic processes that may be improved upon. Marshalls's (2000, 2002, 2007) ICA uses non-redundant discrete wavelet transformation with Daubechies 4 real-valued wavelets. Because the resulting coefficients are real valued estimating the amount of power in the signal requires some sort of rectification of the resulting coefficients. Marshall's ICA accomplishes this using thresholding. When complex valued wavelets are used power can be obtained much more simply by taking the complex modulus of the coefficients. Here I proposed using non-redundant discrete wavelet transformation with complex-valued Morlet wavelets to obtain power spectral estimates.

Secondly, I hypothesize that redundant transforms whose scales (bandwidths) are optimized by an evolutionary algorithm (EA) may improve the ability of GP to classify the resulting wavelet components. Previous work has shown large amounts of variability between individuals in the power spectrum distinguishing low and high workload conditions. Allowing evolutionary algorithms to manipulate the scale of the wavelets might provide a means of tailoring filter banks to specific individuals.

I also hypothesize that incorporating additional physiological measures may increase the efficacy of workload estimates obtained via GP. Empirically observed physiological measures often have low correlations with one another (Kahneman, 1973). This suggests that they may provide non-redundant information pertaining to mental workload. I am specifically interested in examining respiration and electrocardiogram measures.

Table 5.3.1
Overview of GP Model Parameters

Initial Individuals:	"Full Method" with depth 4, 5, or 6
Number of Layers:	10
Individuals / Layer:	100
Total Iterations:	270,000
Age Calculation:	$1 + (i_{completed} - i_{created})/popsize_{total}$
Max ages:	Fibonacci: [5, 8, 13, 21, 34, 55, 89, 144, 233, ∞]
Fitness:	Scaled Symbolic Regression
Crossover:	90/10 rule with "Standard" swapping
Mutation:	Point mutation, 10% for scalar constants,1% of non-scalar constants and non-terminals
Parsimony Pressure:	Fixed at $parsimony\_penalty * size(ind)$ where: $parsimony\_penalty = .0005$

Table 5.3.2
LDA vs. Symbolic Regression on Predicting Tracking Error (r^2).

	LDA		*Symbolic Regression*	
Participant	*Best Training*	*Best Test*	*Best Training*	*Best Test*
1	0.273	0.115	0.680	0.119
2	0.399	0.179	0.722	0.314
3	0.415	0.069	0.617	0.187
4	0.240	0.177	0.568	0.189
5	0.397	0.019	0.567	0.181
6	0.449	0.314	0.579	0.332
7	0.307	0.089	0.705	0.104
8	0.588	0.481	0.636	0.443
9	0.276	0.175	0.527	0.160
10	0.411	0.115	0.760	0.121
Averages:	0.376	0.173	0.636	0.215

Table 5.3.3
LDA vs. Symbolic Regression on Classifying Mapping State (classification accuracies)

	LDA		*Symbolic Regression*	
Participant	*Best Training*	*Best Test*	*Best Training*	*Best Test*
1	67.6%	80.4%	80.3%	79.3%
2	74.3%	62.0%	70.8%	62.5%
3	80.9%	49.1%	86.1%	68.8%
4	56.2%	53.3%	69.6%	73.4%
5	65.6%	69.8%	81.0%	81.3%
6	50.0%	49.8%	87.6%	86.0%
7	50.1%	50.0%	70.7%	72.0%
8	92.5%	97.5%	93.1%	93.1%
9	50.1%	50.2%	84.0%	80.7%
10	83.4%	63.2%	81.1%	75.0%
Averages:	67.1%	62.5%	80.6%	77.2%

Appendix 5.3.A Consent Form

CONSENT FORM

Idaho Visual Performance Laboratory
Department of Psychology and Communication Studies
College of Liberal Arts and Social Sciences
University of Idaho
Control of speed during altitude changes

During this experiment you will be presented a display in a virtual environment. Various parameters of this display will be manipulated to examine stress and mental workload. In this experiment you will be asked to control movement in the virtual world using an input device such as a joystick.

The data you provide will be kept anonymous. There will be absolutely no link between your identity and your particular set of data.

Your participation will help increase knowledge of stress and mental workload. Subsequent to your participation the purpose and methods of the study will be described to you and questions about the study will be answered. It is our sincere hope that you will learn something interesting about your visual system from this debriefing.

The risks in this study are minimal, however displays simulating movement may on rare occasion cause motion sickness or eye fatigue in sensitive individuals. If at any time during the experiment you feel eye fatigue, dizziness, headache or nausea, please let the experimenter know immediately so that you can take a break before these symptoms become too intense. We endeavor to design our displays to minimize eye fatigue and motion sickness, and schedule periodic breaks to further reduce their occurrence. As a result, these phenomena have not been a common problem in previous similar studies.

Your participation will require **1** session of approximately **30** minutes. You may withdraw from this study at anytime without penalty. You will receive partial credit for your time spent. However, please be aware that your data is useful to us only if you complete the experiment in its entirety.
This research project has been approved by the University of Idaho Human Assurance Committee. As such, new information developed during the course of the research which may relate to your willingness to continue participation will be provided to you.

Thank you for your participation

Signature_____ Date _____

If you have further questions or encounter problems please contact:
Dr. Brian P. Dyre
(208) 885-6927
bdyre@uidaho.edu

Appendix 5.3.B Debriefing Form

Debriefing Form

Department of Psychology and Communication Studies

College of Letters, Arts, and Social Sciences

INL Physiological Predictors of Workload

Experiment 3

Participant:____

Date:_____

1. Did you move your left hand during the course of the trial while the GSR was still hooked up?

2. How often do you play video games?

    a. What is your video game skill? (Bad, okay or good)

    b. Are you right or left handed?

3. Did you notice that the controls changed throughout the trial?

    a. How many times?

4. How difficult was the task when you first started? (1-10)

5. How difficult were the normal vs. reversed controls? (1-10)

6. Did you feel that you had enough time to feel confident with:

    a. Normal mappings?

   b. Rotated mappings?

7. How uncomfortable was the eye-tracker when you first put it on? (1-10)

8. How uncomfortable was the eye-tracker when you finished? (1-10)

9. Did you find the eye-tracker distracting from the task at hand?

10. Do you think that fatigue played a role in your performance?

    a. How about fatigue from the eye-tracker?

11. Did you have any eye-strain, fatigue, blurred vision, problems focusing on the target, etc. ?

Any additional comments

Appendix 5.3.C Human Assurances Approval

To: Brian Dyre
Psychology & Communication Studies Department
University of Idaho
Moscow, Idaho 83844-3043

From: Traci Craig
Chair, University of Idaho Institutional Review Board
University Research Office
Moscow, Idaho 83844-3010

IRB No.: IRB00000843

FWA: FWA00005639

Date: October 23, 2009January 14, 2014

Project: Second Year Extension: "Perception and Control of Locomotion in Virtual Environments" (Protocol No. 07-115) **Approved October 23, 2009**

On behalf of the Institutional Review Board at the University of Idaho, I am pleased to inform you that the second year extension of your proposal protocol for the above-named research project is approved as offering no significant risk to human subjects. This extension of approval is valid for **one year** from the approval date listed above, after which it will require a new application if you intend to continue.

Thank you for submitting your extension request.

Appendix 5.3.D Genetic Programming Goodness of Fit Development

5.3.A.1 Iteration I. The first symbolic regressor used trials 1, 2, 5, and 6 as the training data with trials 3, 4, 7, and 8 as the test data. For each subject eight separate runs were performed with 300,000 iterations. At each iteration, fitness was calculated using all 4096 data points of the training data. Eight runs were performed for each subject. Although GP did find solutions that fit the training data well (in some cases explaining over 60% of the variability of the training data), it tended to over fit models to and did not generalize well to the testing data (see Figure 5.3.28). shows the results of a typical run with over fitting.

5.3.A.2 Iteration II. In an attempt to prevent over-fitting, the tracking data was first smoothed using a Hanning window of 256 (15 seconds). Then 256 random, non-consecutive,points in time were sampled and removed to serve as test data. The remaining 7936 time points were given to the model as training data, but at iteration only a subset was selected to evaluate fitness. Iteration II used a subset size of 256 points (3.22% of the avalaible training data). This size was found to prevent over fitting, but severely hampered the dynamics of the model predictions (see Figure 5.3.29). The models basically predict the mean but do not capture any useful information. Eight runs were ran for subjects 1, 2, 3, 4, 8, and 10 which was enough to convince me it didn't work well.

5.3.A.3 Iteration III. In this Iteration the design and parameters remained the same as Iteration II. The only modification was that the subset size was increased to 2048 (25.8% of the available training data). This subset size did find solutions with good dynamic performance, but did show some signs of over fitting (see Figure 5.3.30). Internal review (with Dyre and Werner) suggested that the model should predict a prolonged period of time rather than randomly distributed points of time. The following Iteration IV model attempts to address these issues.

5.3.A.4 Iteration IV. In this Iteration trials 1, 2, 3, 4, 5, and 6 were used as training data. Trials 7 and 8 were used as testing data. At each iteration a subset of 1024 random points (16.66%

of the available training data) were used to calculate fitness (see Figure 5.3.31). Eight runs were performed for each subject. The results from Iteration IV were compared to LDA.

Figure 5.3.28 *Iteration I. Model depicts participant 5's actual tracking error as the blue line in degrees (y-axis) by time in seconds (on the x-axis). The green line is showing the predicted tracking error from a single run values over the training data. The red line extrapolates the model to the test data. The training data fits well but does not generalize to the test data as evidenced by the portion between 120 and 180 seconds and the portion between 420 and 450 seconds.*

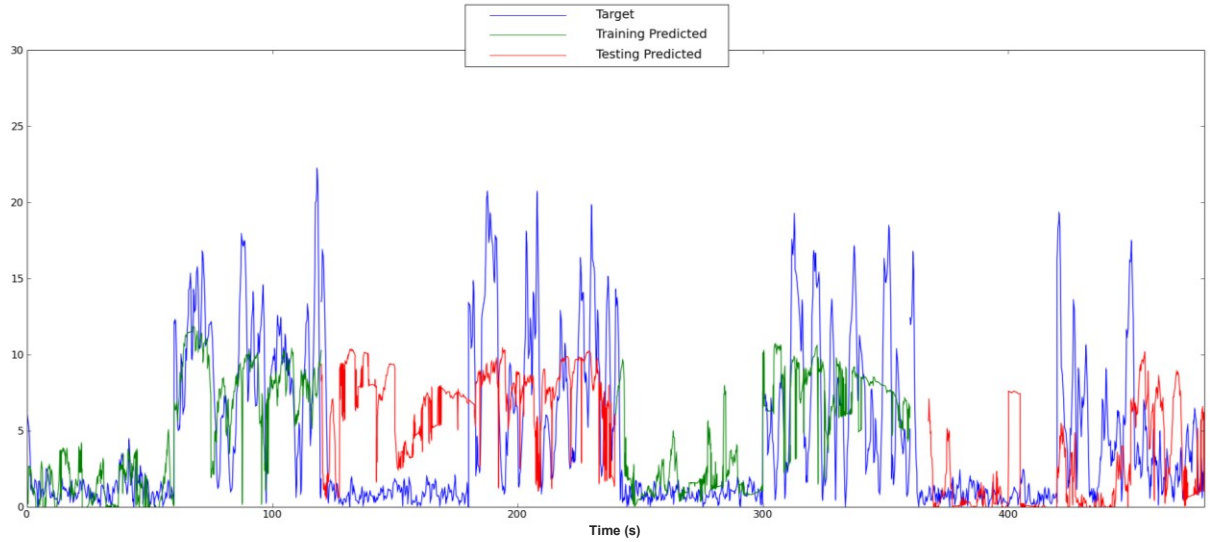


Figure 5.3.29 *Iteration II. Like Figure 3.3 the blue line is showing actual tracking error over time. The green line is showing the predicted values from a single run of the Iteration II model and the red dots are showing testing predictions from computed from values not in the training set. When the subset size is too small (256) the programs can identify the means, but fail to predict dynamic changes in tracking error. (Plot is showing actual tracking error, smoothed tracking error model was fit to smoothed tracking error)*

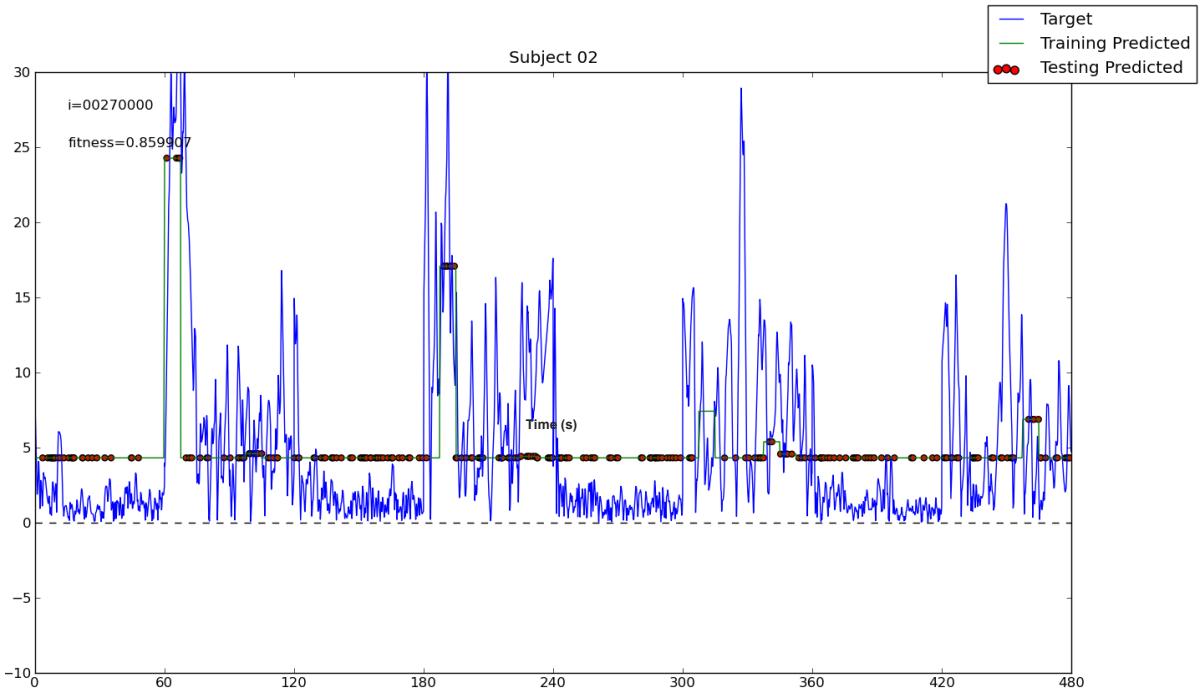


Figure 5.3.30 *Iteration III. The blue line shows actual tracking error over time. The green line shows the predicted values from a single run and the red dots are showing testing predictions from computed from values not in the training set. The Iteration III model used here had random subset of 2048 time points was chosen at each iteration to calculate fitness. Here the fit is fairly good, but the red point below the green line at ~ 205 seconds and the red point above the green line at ~ 460 seconds imply the model is slightly over fitting the training data. (Plot is showing actual tracking error, smoothed tracking error model was fit to smoothed tracking error).*

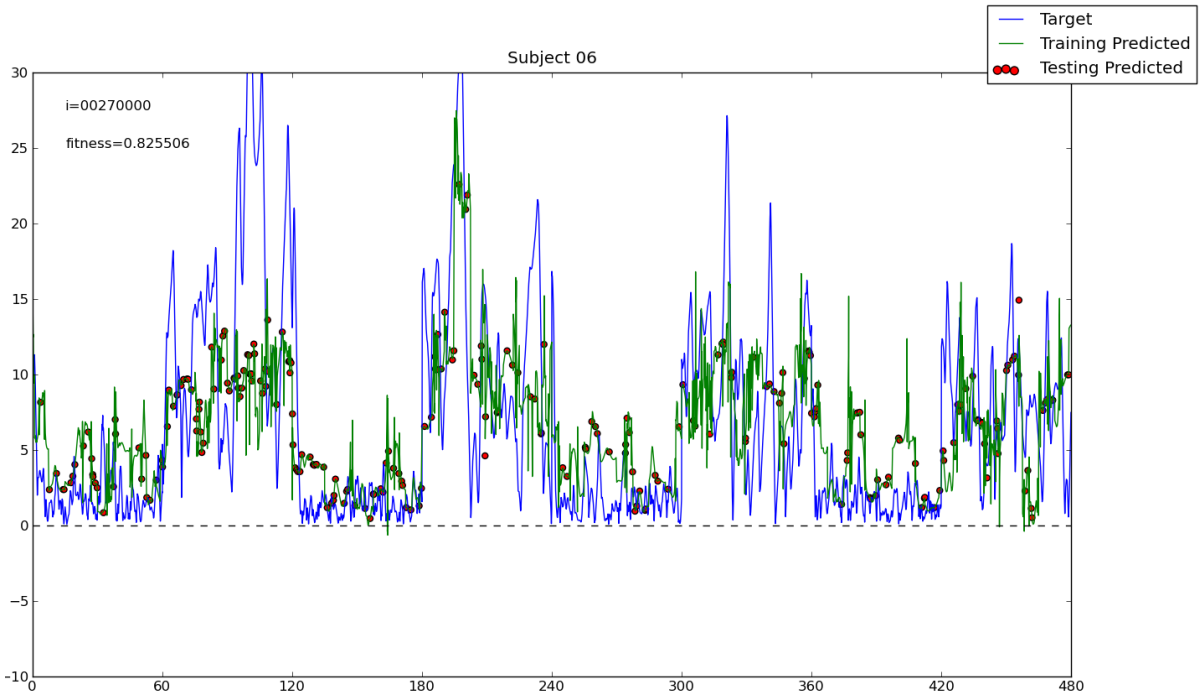
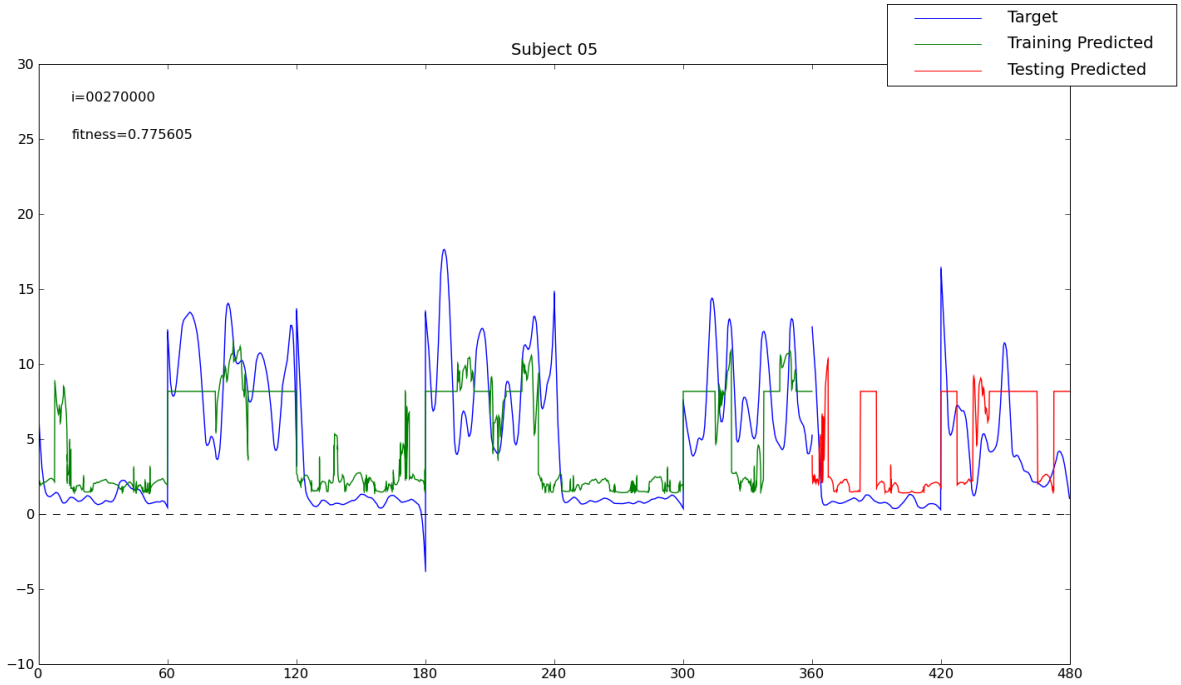


Figure 5.3.31 *Iteration IV. The Iteration IV model is the best prediction on the test data over 8 runs for participant 5's as the red line. As with the previous graphs the blue line is showing smoothed The training data fits well but does not generalize to the test data as evidenced by the portion between 120 and 180 seconds and the portion between 420 and 450 seconds.*



Appendix 5.3.E LDA Models of Tracking Error by Participant

```
Subject 1

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-5.588 -2.347 -1.317  0.938 20.457

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.4976204  0.5353479  -6.533 6.82e-11 ***
GSR_RAW      0.0039470  0.0157854   0.250  0.80256
GSR_COEFF0  -0.0472236  0.0076868  -6.143 8.45e-10 ***
GSR_COEFF1   0.2085829  0.0305603   6.825 9.40e-12 ***
GSR_COEFF2  -0.0851984  0.0278178  -3.063  0.00220 **
GSR_COEFF3  -0.2336022  0.0295385  -7.908 2.95e-15 ***
GSR_COEFF4  -0.0447442  0.0318734  -1.404  0.16041
GSR_COEFF5  -0.0047582  0.0338023  -0.141  0.88806
GSR_COEFF6  -0.0107204  0.0310983  -0.345  0.73031
GSR_COEFF7  -0.0068240  0.0280815  -0.243  0.80801
PUP_RAW      0.0513158  0.0038789  13.229  < 2e-16 ***
PUP_COEFF0   0.0062678  0.0003545  17.679  < 2e-16 ***
PUP_COEFF1  -0.0002740  0.0007403  -0.370  0.71130
PUP_COEFF2  -0.0151778  0.0012473 -12.169  < 2e-16 ***
PUP_COEFF3  -0.0111247  0.0019176  -5.801 6.82e-09 ***
PUP_COEFF4  -0.0075670  0.0025570  -2.959  0.00309 **
PUP_COEFF5   0.0025790  0.0028912   0.892  0.37240
PUP_COEFF6   0.0053456  0.0042938   1.245  0.21318
PUP_COEFF7   0.0189606  0.0095999   1.975  0.04829 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.786 on 8173 degrees of freedom
Multiple R-squared: 0.09282,    Adjusted R-squared: 0.09083
F-statistic: 46.46 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 2

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-24.980  -3.705  -1.162   1.797  31.944

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.4926906  0.8920636 -10.641  < 2e-16 ***
GSR_RAW      0.1287994  0.0220146   5.851 5.09e-09 ***
GSR_COEFF0  -0.0675660  0.0037866 -17.843  < 2e-16 ***
GSR_COEFF1   0.1534978  0.0175036   8.770  < 2e-16 ***
GSR_COEFF2   0.1074249  0.0395299   2.718  0.00659 **
GSR_COEFF3  -0.2053483  0.0471761  -4.353 1.36e-05 ***
GSR_COEFF4   0.0346978  0.0455438   0.762  0.44617
GSR_COEFF5   0.0231100  0.0485839   0.476  0.63432
GSR_COEFF6   0.0405626  0.0441097   0.920  0.35782
GSR_COEFF7  -0.0101404  0.0392371  -0.258  0.79607
PUP_RAW      0.0715792  0.0074379   9.624  < 2e-16 ***
PUP_COEFF0   0.0150976  0.0007587  19.900  < 2e-16 ***
PUP_COEFF1  -0.0148982  0.0013418 -11.103  < 2e-16 ***
PUP_COEFF2  -0.0197211  0.0020999  -9.391  < 2e-16 ***
PUP_COEFF3   0.0027143  0.0028628   0.948  0.34309
PUP_COEFF4   0.0302116  0.0047869   6.311 2.91e-10 ***
PUP_COEFF5   0.0009610  0.0059669   0.161  0.87206
PUP_COEFF6   0.0146284  0.0074879   1.954  0.05078 .
PUP_COEFF7   0.0284381  0.0149764   1.899  0.05762 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.316 on 8173 degrees of freedom
Multiple R-squared: 0.1608,     Adjusted R-squared: 0.1589
F-statistic: 86.97 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 3

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-8.163 -2.697 -1.350  1.226 25.055

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.9601763  0.4873021  -4.023 5.81e-05 ***
GSR_RAW     -0.0134113  0.0170959  -0.784 0.432787
GSR_COEFF0  -0.0702055  0.0052090 -13.478  < 2e-16 ***
GSR_COEFF1   0.1666719  0.0215943   7.718 1.32e-14 ***
GSR_COEFF2  -0.0971365  0.0299156  -3.247 0.001171 **
GSR_COEFF3  -0.1222133  0.0327593  -3.731 0.000192 ***
GSR_COEFF4  -0.0407960  0.0350796  -1.163 0.244882
GSR_COEFF5  -0.0313087  0.0370877  -0.844 0.398594
GSR_COEFF6  -0.0189753  0.0340550  -0.557 0.577409
GSR_COEFF7  -0.0013668  0.0306741  -0.045 0.964461
PUP_RAW      0.0550566  0.0038758  14.205  < 2e-16 ***
PUP_COEFF0   0.0046504  0.0003442  13.510  < 2e-16 ***
PUP_COEFF1  -0.0036930  0.0005430  -6.802 1.11e-11 ***
PUP_COEFF2  -0.0057787  0.0011365  -5.085 3.77e-07 ***
PUP_COEFF3  -0.0032540  0.0020523  -1.586 0.112875
PUP_COEFF4   0.0027213  0.0035376   0.769 0.441764
PUP_COEFF5  -0.0165504  0.0048057  -3.444 0.000576 ***
PUP_COEFF6   0.0204748  0.0061836   3.311 0.000933 ***
PUP_COEFF7   0.0248225  0.0075447   3.290 0.001006 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.155 on 8173 degrees of freedom
Multiple R-squared: 0.0746,     Adjusted R-squared: 0.07257
F-statistic:  36.6 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 4

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-9.790 -3.890 -1.699  1.786 39.953

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0795665  1.0904524  -0.990 0.322195
GSR_RAW      0.0308272  0.0252794   1.219 0.222706
GSR_COEFF0   0.0113635  0.0173577   0.655 0.512700
GSR_COEFF1  -0.1102145  0.0453485  -2.430 0.015104 *
GSR_COEFF2  -0.0933474  0.0480962  -1.941 0.052311 .
GSR_COEFF3   0.0118849  0.0481525   0.247 0.805055
GSR_COEFF4  -0.1193563  0.0515467  -2.316 0.020610 *
GSR_COEFF5   0.0458219  0.0541614   0.846 0.397564
GSR_COEFF6   0.0146020  0.0496950   0.294 0.768893
GSR_COEFF7  -0.0046208  0.0447866  -0.103 0.917827
PUP_RAW      0.0210188  0.0043795   4.799 1.62e-06 ***
PUP_COEFF0   0.0088535  0.0004727  18.728  < 2e-16 ***
PUP_COEFF1   0.0033996  0.0007129   4.769 1.88e-06 ***
PUP_COEFF2   0.0075504  0.0010324   7.313 2.86e-13 ***
PUP_COEFF3   0.0062854  0.0018135   3.466 0.000531 ***
PUP_COEFF4   0.0189718  0.0031230   6.075 1.30e-09 ***
PUP_COEFF5  -0.0008549  0.0039320  -0.217 0.827889
PUP_COEFF6   0.0028020  0.0067800   0.413 0.679414
PUP_COEFF7   0.0045826  0.0097565   0.470 0.638582
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.033 on 8173 degrees of freedom
Multiple R-squared: 0.06078,    Adjusted R-squared: 0.05871
F-statistic: 29.38 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 5

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-8.554 -3.262 -1.175  2.079 19.970

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.8931571  0.5458973  18.123  < 2e-16 ***
GSR_RAW     -0.0149224  0.0183622  -0.813 0.416429
GSR_COEFF0  -0.2708924  0.0122889 -22.044  < 2e-16 ***
GSR_COEFF1   0.0941819  0.0282364   3.335 0.000855 ***
GSR_COEFF2   0.0412410  0.0319700   1.290 0.197090
GSR_COEFF3  -0.0484786  0.0345131  -1.405 0.160166
GSR_COEFF4  -0.0906730  0.0371827  -2.439 0.014766 *
GSR_COEFF5  -0.0380633  0.0392759  -0.969 0.332510
GSR_COEFF6  -0.0241001  0.0360194  -0.669 0.503460
GSR_COEFF7  -0.0023054  0.0324444  -0.071 0.943354
PUP_RAW      0.0554238  0.0019806  27.984  < 2e-16 ***
PUP_COEFF0   0.0042311  0.0001658  25.523  < 2e-16 ***
PUP_COEFF1  -0.0028109  0.0003121  -9.006  < 2e-16 ***
PUP_COEFF2  -0.0051059  0.0004899 -10.422  < 2e-16 ***
PUP_COEFF3   0.0033019  0.0011252   2.935 0.003350 **
PUP_COEFF4  -0.0061709  0.0018981  -3.251 0.001155 **
PUP_COEFF5   0.0112832  0.0029548   3.819 0.000135 ***
PUP_COEFF6  -0.0006015  0.0046704  -0.129 0.897532
PUP_COEFF7   0.0294532  0.0069916   4.213 2.55e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.388 on 8173 degrees of freedom
Multiple R-squared: 0.1989,     Adjusted R-squared: 0.1971
F-statistic: 112.7 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 6

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-27.586  -4.107  -2.227   2.486  31.230

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.534064   0.995229 -20.633  < 2e-16 ***
GSR_RAW       0.113220   0.026914   4.207 2.62e-05 ***
GSR_COEFF0    0.131372   0.008744  15.024  < 2e-16 ***
GSR_COEFF1    0.054608   0.024986   2.186   0.0289 *
GSR_COEFF2    0.006776   0.045717   0.148   0.8822
GSR_COEFF3   -0.018762   0.048315  -0.388   0.6978
GSR_COEFF4   -0.046417   0.055438  -0.837   0.4025
GSR_COEFF5    0.007745   0.058675   0.132   0.8950
GSR_COEFF6    0.005190   0.053750   0.097   0.9231
GSR_COEFF7   -0.029444   0.048398  -0.608   0.5430
PUP_RAW       0.024624   0.004375   5.628 1.88e-08 ***
PUP_COEFF0    0.015077   0.001017  14.823  < 2e-16 ***
PUP_COEFF1   -0.009712   0.001372  -7.079 1.57e-12 ***
PUP_COEFF2    0.004281   0.001824   2.347   0.0189 *
PUP_COEFF3    0.008307   0.002134   3.893 9.99e-05 ***
PUP_COEFF4   -0.002801   0.001894  -1.479   0.1393
PUP_COEFF5   -0.004612   0.002069  -2.229   0.0258 *
PUP_COEFF6   -0.004151   0.003288  -1.262   0.2068
PUP_COEFF7    0.015364   0.006864   2.238   0.0252 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.467 on 8173 degrees of freedom
Multiple R-squared: 0.09664,    Adjusted R-squared: 0.09465
F-statistic: 48.57 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 7

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-6.984 -3.201 -1.953  1.340 28.629

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.180616   1.443957  -4.280 1.89e-05 ***
GSR_RAW      0.024171   0.021062   1.148 0.251166
GSR_COEFF0  -0.017692   0.016676  -1.061 0.288755
GSR_COEFF1   0.122427   0.029902   4.094 4.28e-05 ***
GSR_COEFF2   0.024022   0.036379   0.660 0.509073
GSR_COEFF3  -0.139272   0.039181  -3.555 0.000381 ***
GSR_COEFF4  -0.062012   0.042852  -1.447 0.147894
GSR_COEFF5  -0.011003   0.045304  -0.243 0.808116
GSR_COEFF6  -0.021822   0.041443  -0.527 0.598513
GSR_COEFF7  -0.012050   0.037254  -0.323 0.746349
PUP_RAW      0.052768   0.004649  11.349  < 2e-16 ***
PUP_COEFF0   0.009329   0.000536  17.404  < 2e-16 ***
PUP_COEFF1  -0.001557   0.001130  -1.378 0.168191
PUP_COEFF2  -0.015645   0.001618  -9.672  < 2e-16 ***
PUP_COEFF3  -0.012388   0.002898  -4.274 1.94e-05 ***
PUP_COEFF4   0.008589   0.003780   2.272 0.023108 *
PUP_COEFF5  -0.006216   0.004694  -1.324 0.185483
PUP_COEFF6  -0.008524   0.005589  -1.525 0.127268
PUP_COEFF7   0.019539   0.006109   3.199 0.001386 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.004 on 8173 degrees of freedom
Multiple R-squared: 0.06306,    Adjusted R-squared: 0.06099
F-statistic: 30.56 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 8

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-14.046  -4.729  -1.079   2.552  25.442

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.777e+01  7.010e-01 -25.341  < 2e-16 ***
GSR_RAW      4.211e-02  2.726e-02   1.545  0.12249
GSR_COEFF0  -1.563e-01  9.421e-03 -16.588  < 2e-16 ***
GSR_COEFF1  -3.769e-02  4.201e-02  -0.897  0.36964
GSR_COEFF2  -1.459e-01  4.740e-02  -3.078  0.00209 **
GSR_COEFF3   2.337e-01  5.051e-02   4.626 3.78e-06 ***
GSR_COEFF4   1.973e-02  5.570e-02   0.354  0.72316
GSR_COEFF5   1.704e-02  5.869e-02   0.290  0.77160
GSR_COEFF6   4.890e-04  5.365e-02   0.009  0.99273
GSR_COEFF7  -9.481e-03  4.816e-02  -0.197  0.84394
PUP_RAW      6.201e-02  4.114e-03  15.071  < 2e-16 ***
PUP_COEFF0   2.463e-02  5.013e-04  49.126  < 2e-16 ***
PUP_COEFF1   1.416e-02  7.822e-04  18.100  < 2e-16 ***
PUP_COEFF2  -2.040e-03  1.201e-03  -1.698  0.08950 .
PUP_COEFF3  -5.088e-03  1.718e-03  -2.962  0.00307 **
PUP_COEFF4  -9.224e-03  2.182e-03  -4.228 2.38e-05 ***
PUP_COEFF5  -2.458e-03  2.404e-03  -1.022  0.30670
PUP_COEFF6   1.437e-02  3.489e-03   4.119 3.85e-05 ***
PUP_COEFF7   2.350e-02  7.218e-03   3.256  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.524 on 8173 degrees of freedom
Multiple R-squared: 0.3195,     Adjusted R-squared: 0.318
F-statistic: 213.2 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 9

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-7.346 -3.184 -1.356  2.017 23.083

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.9483964  0.7801850  -8.906  < 2e-16 ***
GSR_RAW      0.0270299  0.0192516   1.404 0.160348
GSR_COEFF0  -0.0487356  0.0052611  -9.263  < 2e-16 ***
GSR_COEFF1   0.3437227  0.0241079  14.258  < 2e-16 ***
GSR_COEFF2  -0.1286393  0.0359124  -3.582 0.000343 ***
GSR_COEFF3  -0.1630407  0.0381332  -4.276 1.93e-05 ***
GSR_COEFF4  -0.0521932  0.0398376  -1.310 0.190182
GSR_COEFF5  -0.0301926  0.0418407  -0.722 0.470556
GSR_COEFF6   0.0040968  0.0380832   0.108 0.914336
GSR_COEFF7   0.0013190  0.0340449   0.039 0.969097
PUP_RAW      0.0346666  0.0039832   8.703  < 2e-16 ***
PUP_COEFF0   0.0124888  0.0005751  21.715  < 2e-16 ***
PUP_COEFF1   0.0074621  0.0008634   8.643  < 2e-16 ***
PUP_COEFF2   0.0035784  0.0012862   2.782 0.005413 **
PUP_COEFF3   0.0165425  0.0021181   7.810 6.42e-15 ***
PUP_COEFF4  -0.0182888  0.0023270  -7.859 4.35e-15 ***
PUP_COEFF5   0.0014283  0.0030807   0.464 0.642923
PUP_COEFF6  -0.0016439  0.0039018  -0.421 0.673542
PUP_COEFF7  -0.0018306  0.0054879  -0.334 0.738720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.587 on 8173 degrees of freedom
Multiple R-squared: 0.1323,     Adjusted R-squared: 0.1304
F-statistic: 69.21 on 18 and 8173 DF,  p-value: < 2.2e-16
```

```
Subject 10

Call:
lm(formula = TRK_ERR ~ GSR_RAW + GSR_COEFF0 + GSR_COEFF1 + GSR_COEFF2 +
    GSR_COEFF3 + GSR_COEFF4 + GSR_COEFF5 + GSR_COEFF6 + GSR_COEFF7 +
    PUP_RAW + PUP_COEFF0 + PUP_COEFF1 + PUP_COEFF2 + PUP_COEFF3 +
    PUP_COEFF4 + PUP_COEFF5 + PUP_COEFF6 + PUP_COEFF7, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-11.631  -4.102  -1.767   1.796  36.920

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.759e+01  1.259e+00 -29.848  < 2e-16 ***
GSR_RAW     -2.860e-03  2.798e-02  -0.102   0.9186
GSR_COEFF0   2.028e-01  1.722e-02  11.778  < 2e-16 ***
GSR_COEFF1   3.207e-01  4.887e-02   6.561 5.65e-11 ***
GSR_COEFF2  -1.306e-01  5.029e-02  -2.598   0.0094 **
GSR_COEFF3  -2.771e-01  5.274e-02  -5.255 1.52e-07 ***
GSR_COEFF4  -2.789e-01  5.740e-02  -4.859 1.20e-06 ***
GSR_COEFF5   6.842e-03  6.036e-02   0.113   0.9097
GSR_COEFF6  -2.608e-02  5.501e-02  -0.474   0.6355
GSR_COEFF7  -2.168e-02  4.943e-02  -0.439   0.6610
PUP_RAW      1.183e-01  8.495e-03  13.926  < 2e-16 ***
PUP_COEFF0   3.209e-02  9.575e-04  33.514  < 2e-16 ***
PUP_COEFF1  -5.862e-03  1.429e-03  -4.102 4.14e-05 ***
PUP_COEFF2  -2.423e-03  2.120e-03  -1.142   0.2533
PUP_COEFF3   2.968e-03  2.791e-03   1.064   0.2875
PUP_COEFF4   3.717e-02  4.560e-03   8.152 4.11e-16 ***
PUP_COEFF5  -4.271e-02  7.291e-03  -5.858 4.86e-09 ***
PUP_COEFF6   2.300e-02  1.054e-02   2.182   0.0291 *
PUP_COEFF7   3.607e-02  1.844e-02   1.957   0.0504 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.682 on 8173 degrees of freedom
Multiple R-squared: 0.1699,     Adjusted R-squared: 0.1681
F-statistic: 92.93 on 18 and 8173 DF,  p-value: < 2.2e-16
```

5.4 Experiment 4: Compensatory Tracking (Subjective Workload Validation)

The previous experiment manipulated task difficulty in a manner that was overt to the participants. In real world settings workload often increases subtly. In such scenarios additional cognitive resources can be applied to compensate and task performance may remain high. Participants may not even be conscious of the task becoming more difficult. I speculate that under such circumstances physiological indicators may still reliable predict workload. To examine whether physiological measures can be used to detect increased workload before task performance degrades a method of subtly manipulating task difficulty is needed. This experiment will examine whether using a tracking task based on the critical instability task (McRuer & Graham, 1965; McDonnell & Jex, 1967) may fulfill this need.

The critical-instability tracking task is a compensatory tracking task originally designed to study aircraft handling and psychomotor tracking. Pilots at the time needed to control unstable high speed aircraft and booster rockets and is based off of McRuer's previous analytical work in this area. McRuer was first to describe a point of instability based on an *effective time delay*. The effective time delay reflects nervous system processing delays, muscular response lags, as well as system lags and high frequency leads. This allowed for the limits of human control to be precisely quantified. Here the intention is not to quantify the limits of human control, but to use the compensatory tracking task as a means of manipulating task difficulty in a manner that is less obvious to the participants.

The key concept with a compensatory tracking task is the incorporation of a positive feedback loop from the system's output to the controller's input. The gain of the feedback is increased as the target deviates from its ideal position. The compensatory task should allow workload to increase in a subtle manner unlike the previous experiments where the control mappings abruptly switched. Put simply, the task can be conceptualized as a task in which a participant balances an inverted pendulum. As the pendulum swings from vertical gravity provides

positive feedback pulling the pendulum from the desired position. Assuming the mass of the pendulum stays constant shorter lengths result in greater instability. Here the goal is to examine how the pendulum length relates to subjective difficulty. This experiment will also allow for some quantification of learning effects (how stable performance is over time), as well as inter-participant variability tracking performance. This experiment found that task performance and perceived difficulty are influenced by pendulum length in a predictable fashion, that satisfactory performance occurs nearly instantaneously, and task performance is highly correlated amongst participants over a wide range of difficulty settings.

5.4.1 *Method*

5.4.1.1 Participants. Ten participants with normal or corrected to normal Snellen visual acuity of 20/30 participated in this study. All were naïve to the hypotheses of the experiment. All participants were ethically treated in accordance with experimental protocols approved by the University of Idaho's Human Assurance Committee (see Appendices 5.4.A – 5.4.C).

5.4.1.2 Stimuli and Apparatus. Participants performed a series of short 30 second single axis compensatory tracking task trials similar to the Critical Tracking Task (McRuer & Graham, 1965; McDonnell & Jex, 1967). The internal dynamics of the simulation modeled an inverted pendulum fixed to a stationary pivot point. User inputs applied torque to the arm of the pendulum. The difficulty of the tracking task was manipulated across trials by setting the pendulum's length at eight equidistant levels in logspace between 0.060 and 2 meters (0.060, 0.099, 0.163, 0.270, 0.445, 0.734, 1.212, 2.000 m). The visually representation of the pendulum model emulated the visual appearance of the previous tracking experiments, even though the simulation dynamics differed substantially. Participants saw the same balanced dot on a grey screen as they saw before. A fixed square annulus in the center of the screen replaced the user controlled crosshair cursor used in the previous experiment. The balanced dot's position on the screen mapped to the angle of an inverted

pendulum such that when the pendulum was perfectly vertical the dot was centered in the square annulus. The pendulum's angular deviations from center mapped linearly to the horizontal position on the screen such that a deviation of 90° corresponded to the edge of the screen. When the dot moved off the screen it was automatically reset at a randomly chosen position close to vertical. The dot's position was fixed along the vertical axis, and the visual appearance of the dot remained invariant to the length of the pendulum.

5.4.1.3 Procedure. As previously mentioned, the stability of an inverted pendulum is inversely related to its length (assuming fixed mass). Future experiments need to be able to precisely control the perceived difficulty of the task manipulating the length of the pendulum. To do this it is first necessary to relate pendulum length to subjective difficulty. This can and was accomplished through application of the Steven's (1953; 1970) free-modulus method of magnitude estimation. Participants received ten block randomized presentations of the eight previously specified pendulum lengths for a total of 80 trials. After each trial they reported a numerical estimate of the perceived difficulty. The *free-modulus* component refers an open ended response scale. That is participants were free to choose any positive value to relate the magnitude of their perceived difficulty. In addition to collecting subjective difficulty ratings state variables were recorded such that root mean squared error (RMSE) of time series tracking performance could be calculated.

5.4.2 *Results*

5.4.2.1 Effects of length and block on subjective ratings and RMSE. Data analysis proceeded conducting omnibus ANOVAs for the subjective ratings and RMSEs across the eight levels of length and ten levels of block. The ANOVAs used Greenhouse and Geisser's correction for violations of sphericity (1958; 1959). Additionally, the subjective ratings exhibited such gross violations in sphericity that a log10 transform was warranted (a constant of 1 was added before the log10 transform; some participants reported the an occasional rating of zero). As anticipated, the

ANOVAs found that length was negatively correlated with both measures. The subjective rating measure had a $F(7,63) = 20.368$, $p = 0.001$, $\varepsilon = 0.152$ and RMSE measure had a $F(7,63) = 333.536$, $p < 0.001$, $\varepsilon = 0.148$. These 8 x 10 (length x block) ANOVAs found no reliable main effects of block and no reliable interactions of length by block (see Table 5.4.1).

To test for learning effects more directly 8 x 2 ANOVAs were conducted across all eight levels of length but only the first and last blocks. If learning is occurring the greatest disparities would be expected between the first and last blocks (discounting the role of fatigue, see discussion). Besides replicating the main effects of length, a moderate main effect of block was found for the subjective ratings measure ($F(1,9) = 7.145$, $p = 0.025$). The main effect of block on RMSE was not reliable nor were any length by block interactions for both measures (see Table 5.4.2). The marginal means of subjective ratings suggest that the task becomes subjectively easier as familiarity develops (see Figure 5.4.1). The RMSE results come with the caveat that observed power is grossly insufficient to eliminate the possibility of Type II error but given that effect size is small, as measured by η^2 at only 0.01, the sample size would have to be increased to over 780 participants to obtain a power of 0.80. Such an endeavor would likely be a poor use of a finite subject pool when overall RMSE marginal means dropped from 18.212° (1.456) across the first block to 17.107° (1.321) across the last block. Across the different lengths the largest difference between blocks occurred at the second shortest length (0.099) with an observed improvement of 3.612° (32.658 to 29.046°, see Figure 5.4.2). These results suggest that while the perceived difficulty of the task decreases with familiarity the actual proficiency is remarkable intuitive and stable. Participants seem to perform at asymptotic levels within moments of encountering the task. Reasons why this might be will be elucidated in the discussion.

Figure 5.4.1 *The log transformed rating data. Found main effects of length and block. Error bars reflect 95% confidence intervals.*

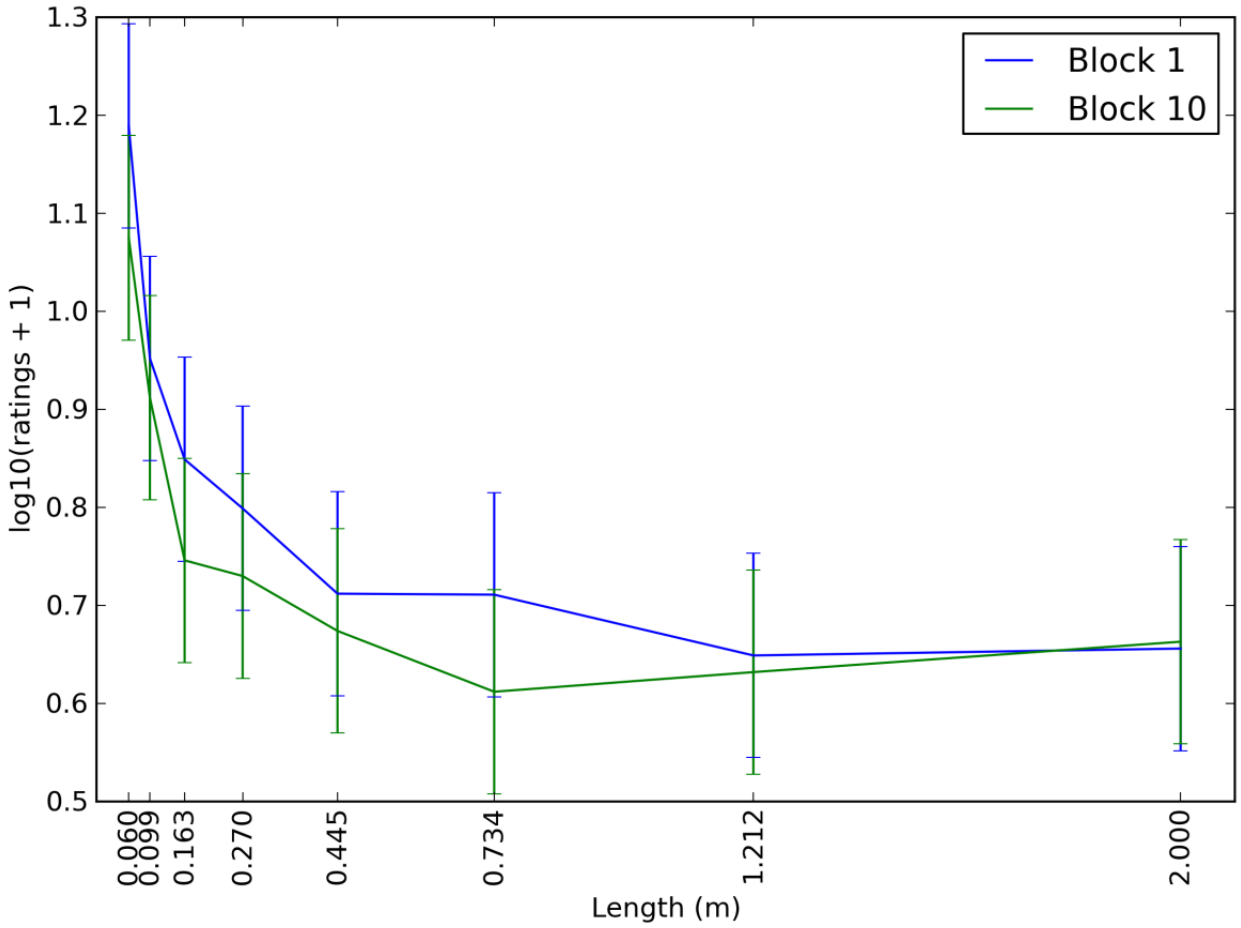
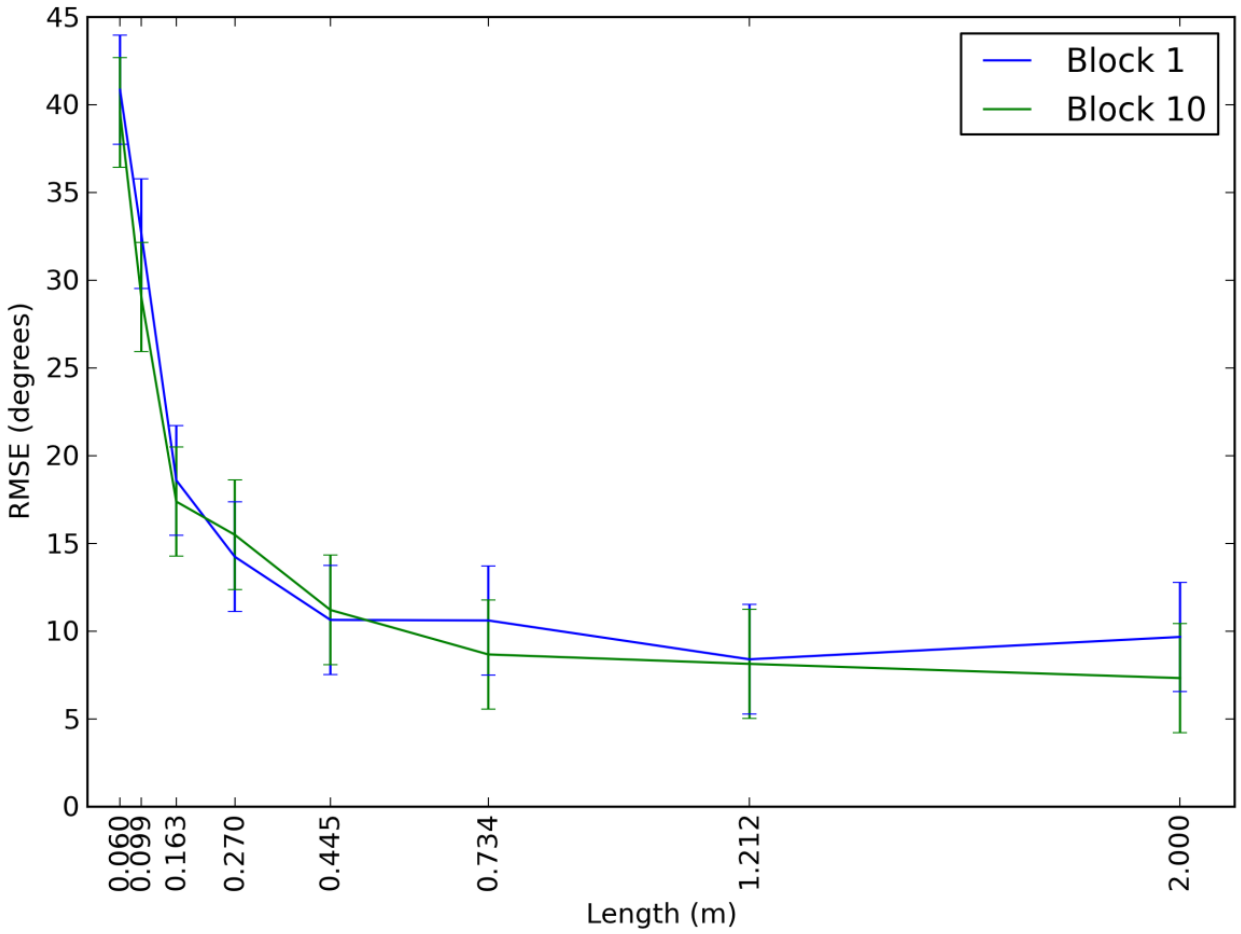


Figure 5.4.2 *The RMSE data found a main effect of length but not block. Error bars reflect 95% confidence intervals.*



5.4.2.2 Magnitude Estimation. Pendulum length can be correlated to subjective ratings using magnitude estimation. This can be accomplished by minimizing the psychophysical transfer functions:

$$\psi(I) = kI^a$$

where,

I is the magnitude of the pendulum's length

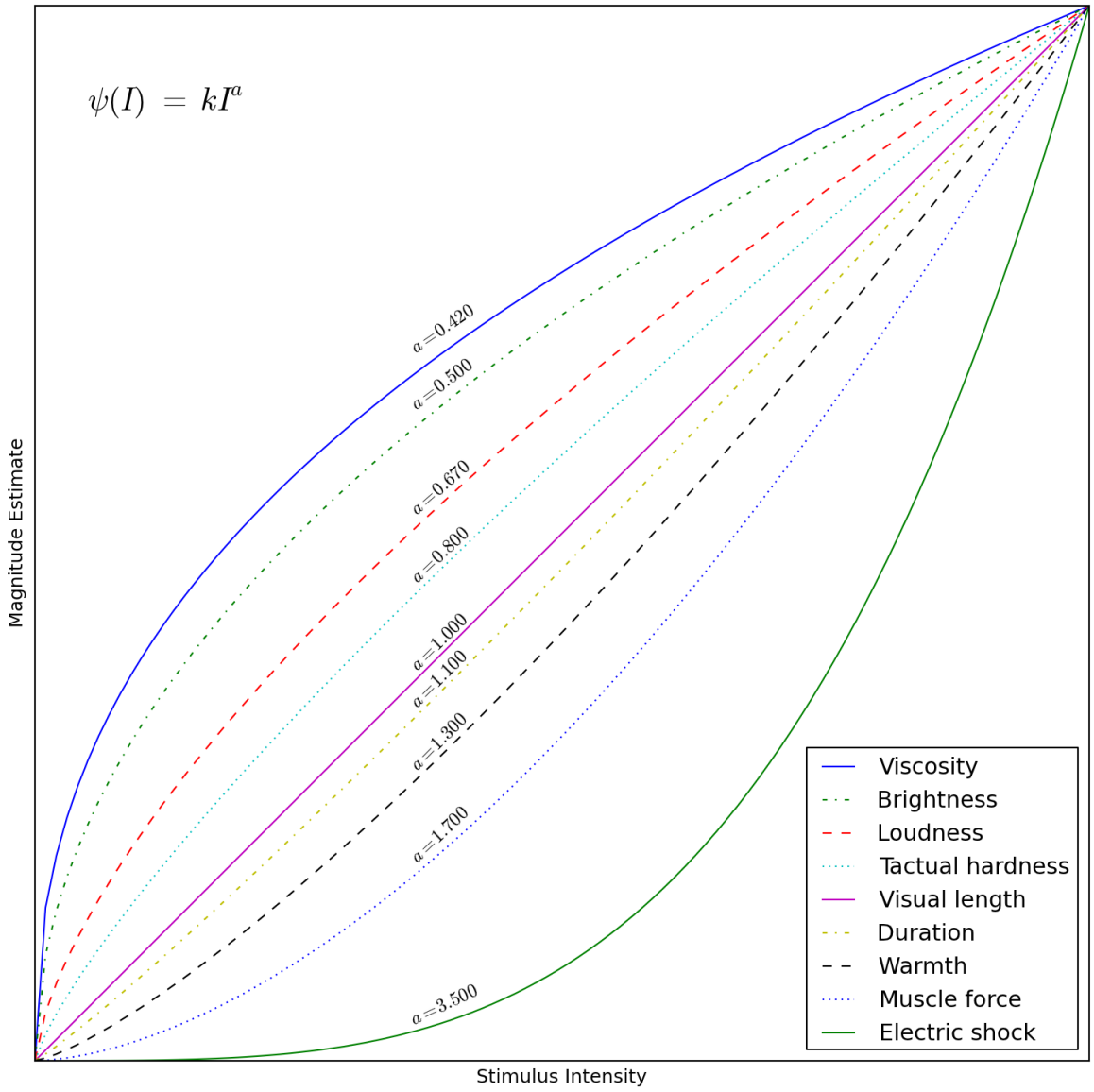
a is the power exponent dependent on the transfer function

k is an arbitrary proportionality constant related to the scale of I

Reader's may recognize the above equation as Steven's power law (1953; 1970). By assuming that each participant maintained a consistent internal rating system throughout the course of the experiment it is possible to assess intrasubject variability by finding k and a parameters for each block and looking at the dispersion across blocks. The exponential parameter a is the more important of the two. The k parameter is somewhat arbitrary as it relates to the particular metrics used in the simulation as well as the range chosen by individual participants. The a expresses the extent to which a stimulus is compressed or expanded. If $a = 1$ a stimulus is perceived linearly. For example, visual lengths are perceived linearly. When $a > 1$ response expansion occurs; that is the ratio of perceived change exceeds the ratio of physical change. Electric shocks for instance induce response expansion. When $a < 1$ response compression occurs; that is the ratio of perceived change is less than the ratio of physical change. Brightness exhibits response compression. **Figure 5.4.3** depicts how the a parameter affects magnitude responses to a variety of physical stimuli.

Typically with magnitude estimation the measurement scale of the physical stimulus and the reported magnitude estimates are positively correlated which makes the a values always positive. In this instance, pendulum length negatively correlates with subjective ratings and RMSE. This in turn makes the a values negative. Values with magnitudes < 1 still reflect response compression and values with magnitudes greater than 1 reflect response expansion. Table 5.4.3

Figure 5.4.3 *Steven's power law. Describes how the magnitude physical stimuli are perceived by our senses.*



reports the average *a* values for each participant across the ten blocks. The magnitudes across all participants are less than one indicating response compression. Less compression (values closer to -1) indicate that a participant is more sensitive at detecting differences in pendulum length. Across participants the average power exponent is -0.338 and participant's subjective ratings can account for 66% of the variability in pendulum length changes. A similar approach can be applied to predict RMSEs by pendulum length. The results of this endeavor are shown in Table 5.4.4 and are depicted in Figure 5.4.4 through Figure 5.4.13. This data shows that task performance as measured by RMSE shows less response compression in predicting pendulum lengths (M = -0.467, SD = 0.058) compared to subjective ratings (M = -0.338, SD = 0.202). An unpaired one-way t-test assuming unequal variance (Welch–Satterthwaite) comparing the average subjective rating power exponents to the average RMSE power exponents shows significance at $p = .04$, $t(10) = 1.95$. This suggests that task performance is a better predictor of the difficulty manipulation than subjective reports. The R^2 values obtained during the least-squares optimization procedure for calculating the power functions adds further support to this claim ($t(11)=4.65$, $p < .001$). RMSE could account for nearly 82% of length variability compared to the 66% predicted by subjective ratings.

5.4.2.3 Intrasubject and intersubject variability. Part of the reason RMSE does so much better at predicting length is because the measure exhibits much less intrasubject dispersity across blocks. Table 5.4.3 and Table 5.4.4 list standard deviations of *a* as well as coefficient of variations of *a*. Standard deviation expresses dispersity in units of the original measure whereas the coefficient of variation is a normalized measured of dispersion obtained by taking the ratio of the standard deviation by the mean. The coefficients of variation across the *a* parameters for RMSE are on average less than half that of those observed for the subjective rating estimates (M=.129 compared to M=.273). A t-test also shows these to be reliably different ($t(18)=4.58$, $p < .001$). These measures of intrasubject variability reflect learning, fatigue, familiarity, adaptation, vigilance, and other cognitive processes as well as noise. In addition to quantifying intrasubject variability it is

Figure 5.4.4 *Magnitude estimates for subjective ratings and RMSE for Participant 1.*

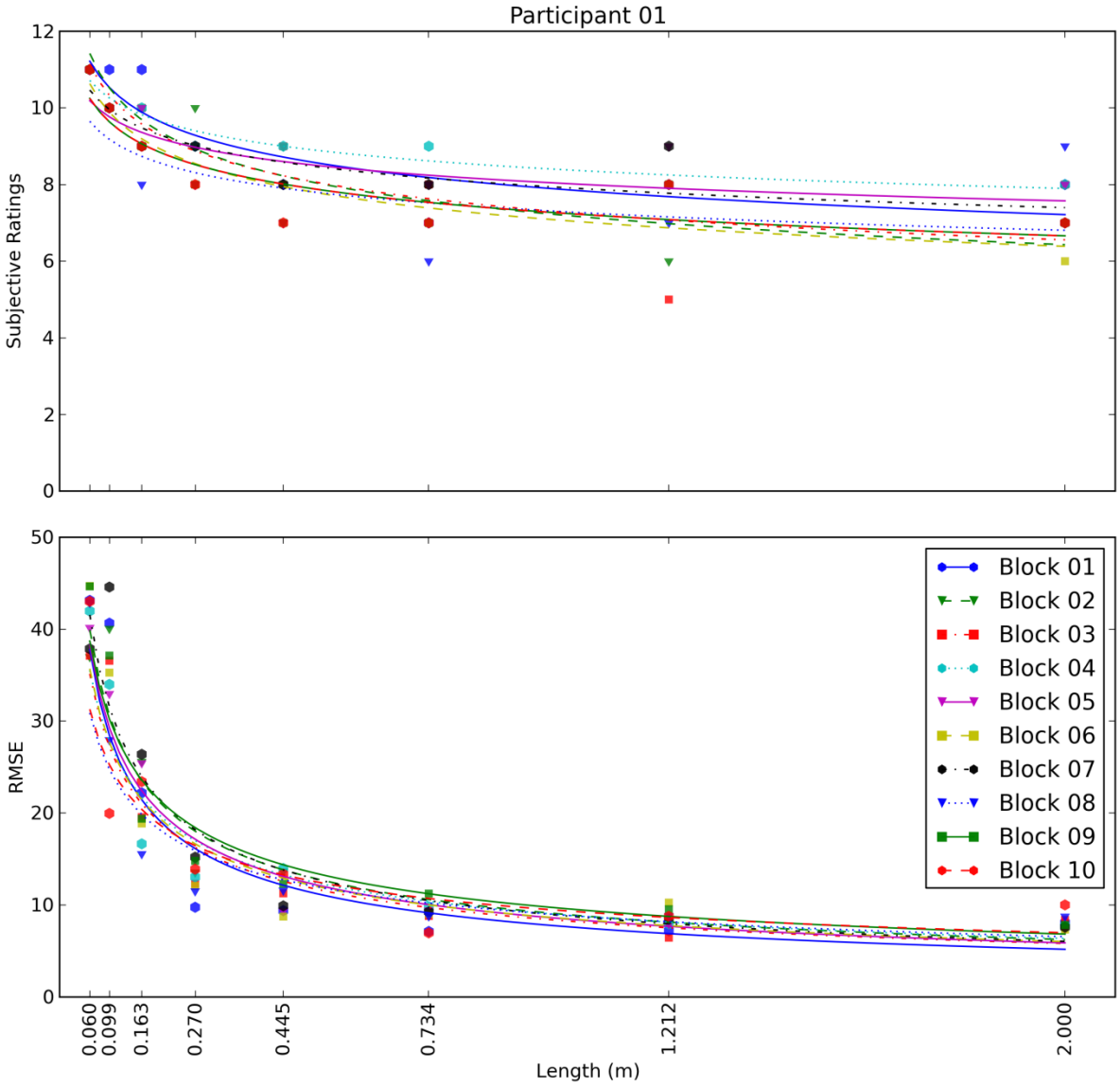


Figure 5.4.5 *Magnitude estimates for subjective ratings and RMSE for Participant 2.*

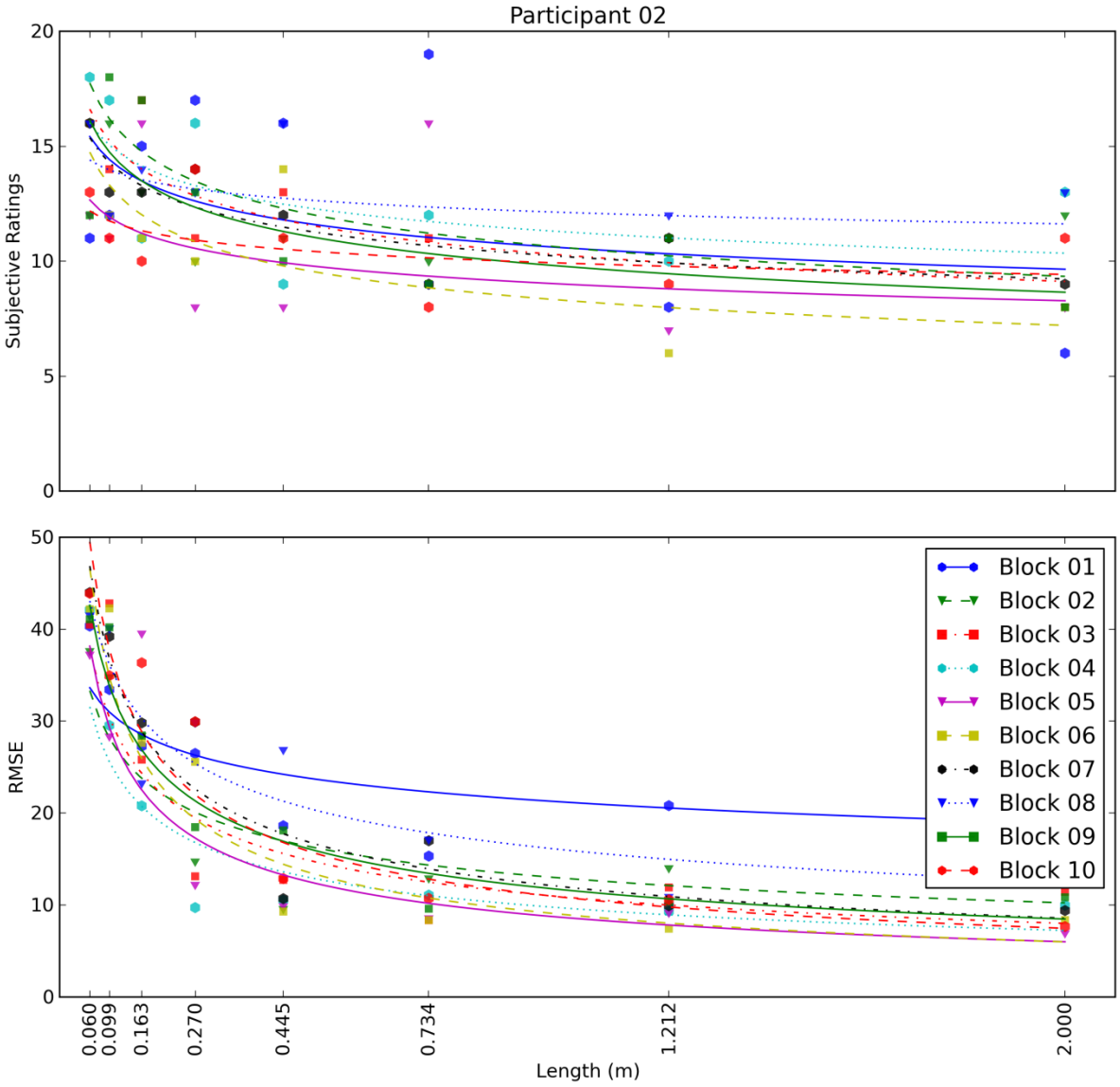


Figure 5.4.6 *Magnitude estimates for subjective ratings and RMSE for Participant 3.*

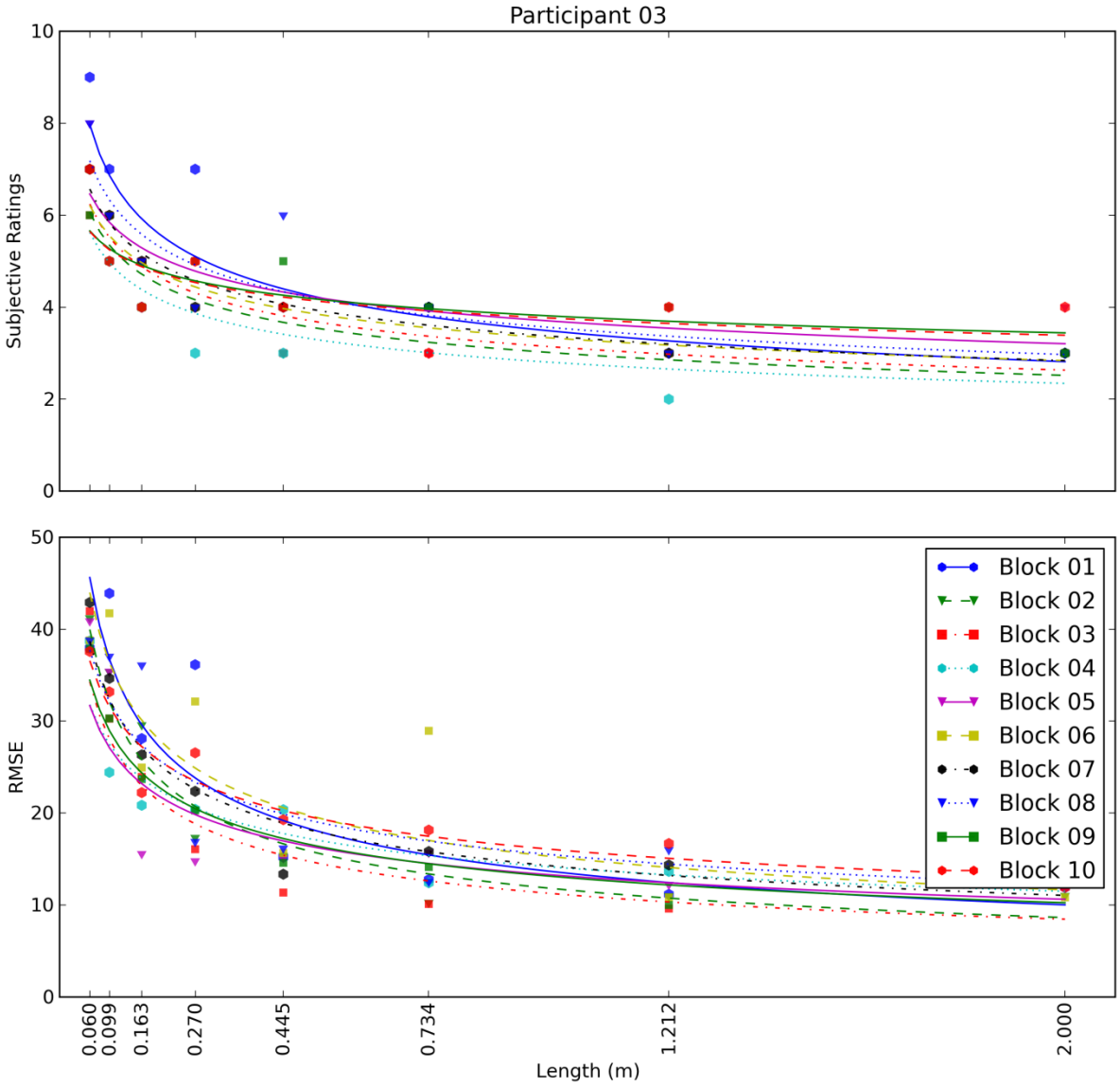


Figure 5.4.7 *Magnitude estimates for subjective ratings and RMSE for Participant 4.*

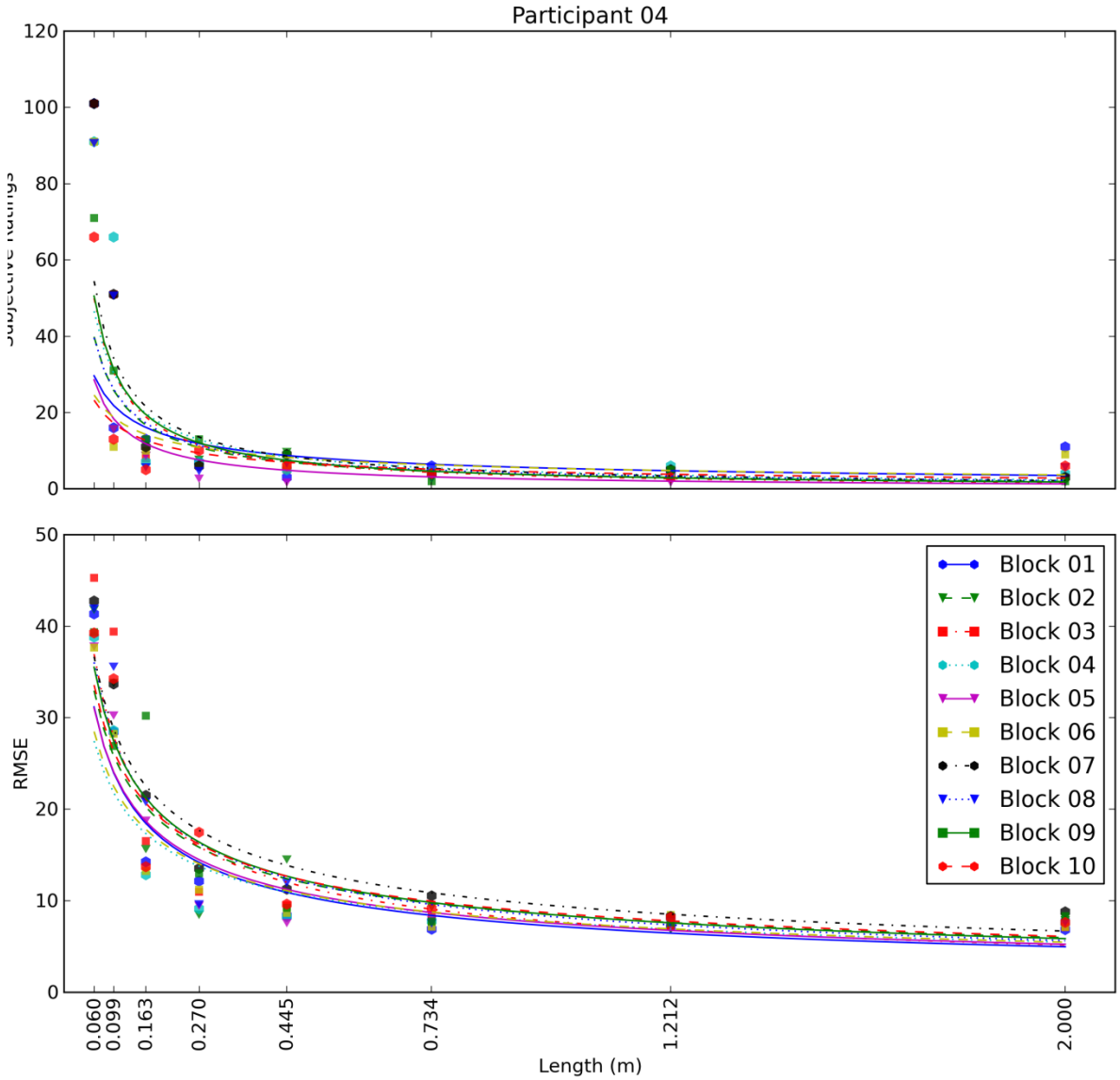


Figure 5.4.8 *Magnitude estimates for subjective ratings and RMSE for Participant 5.*

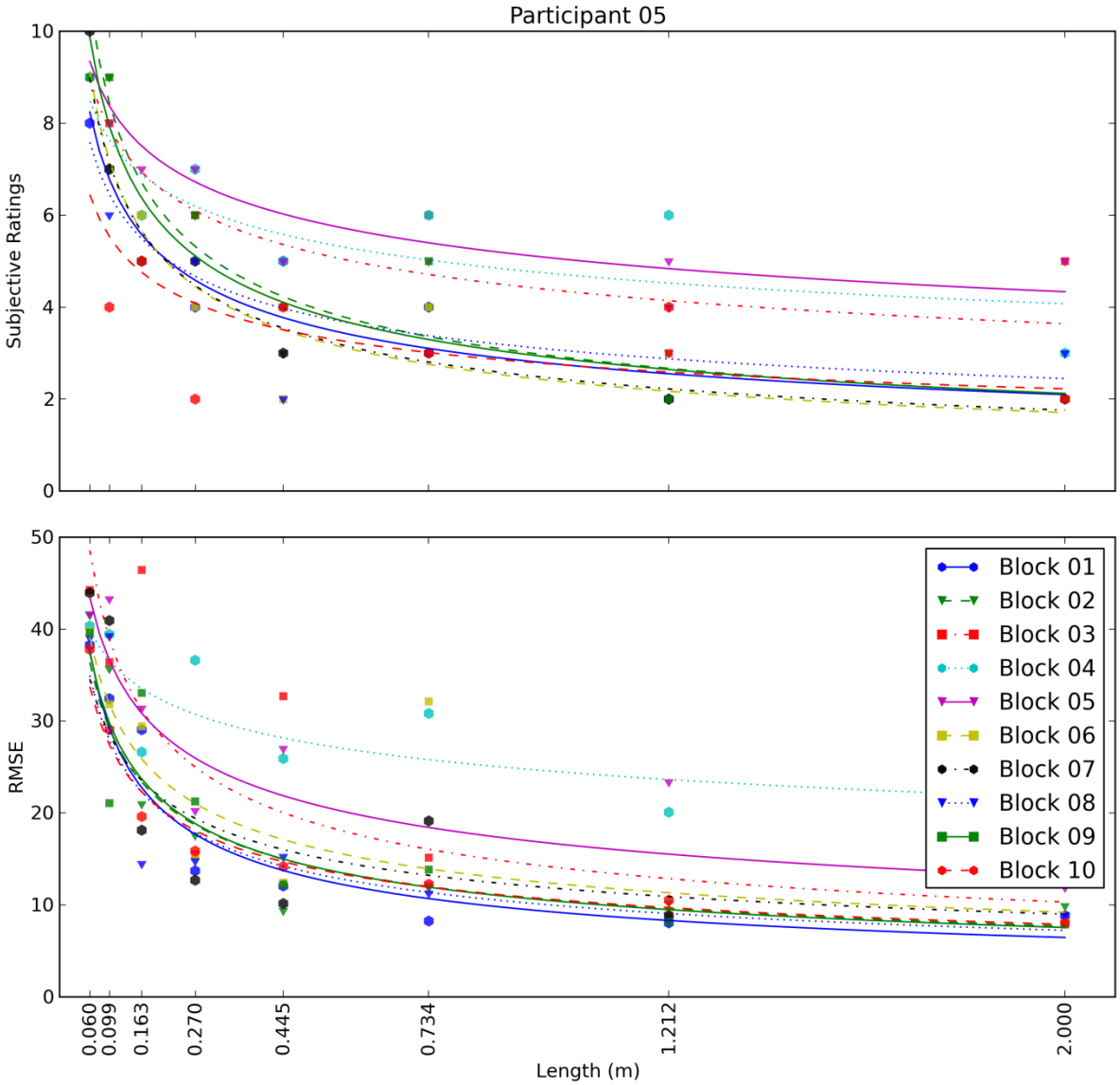


Figure 5.4.9 *Magnitude estimates for subjective ratings and RMSE for Participant 6.*

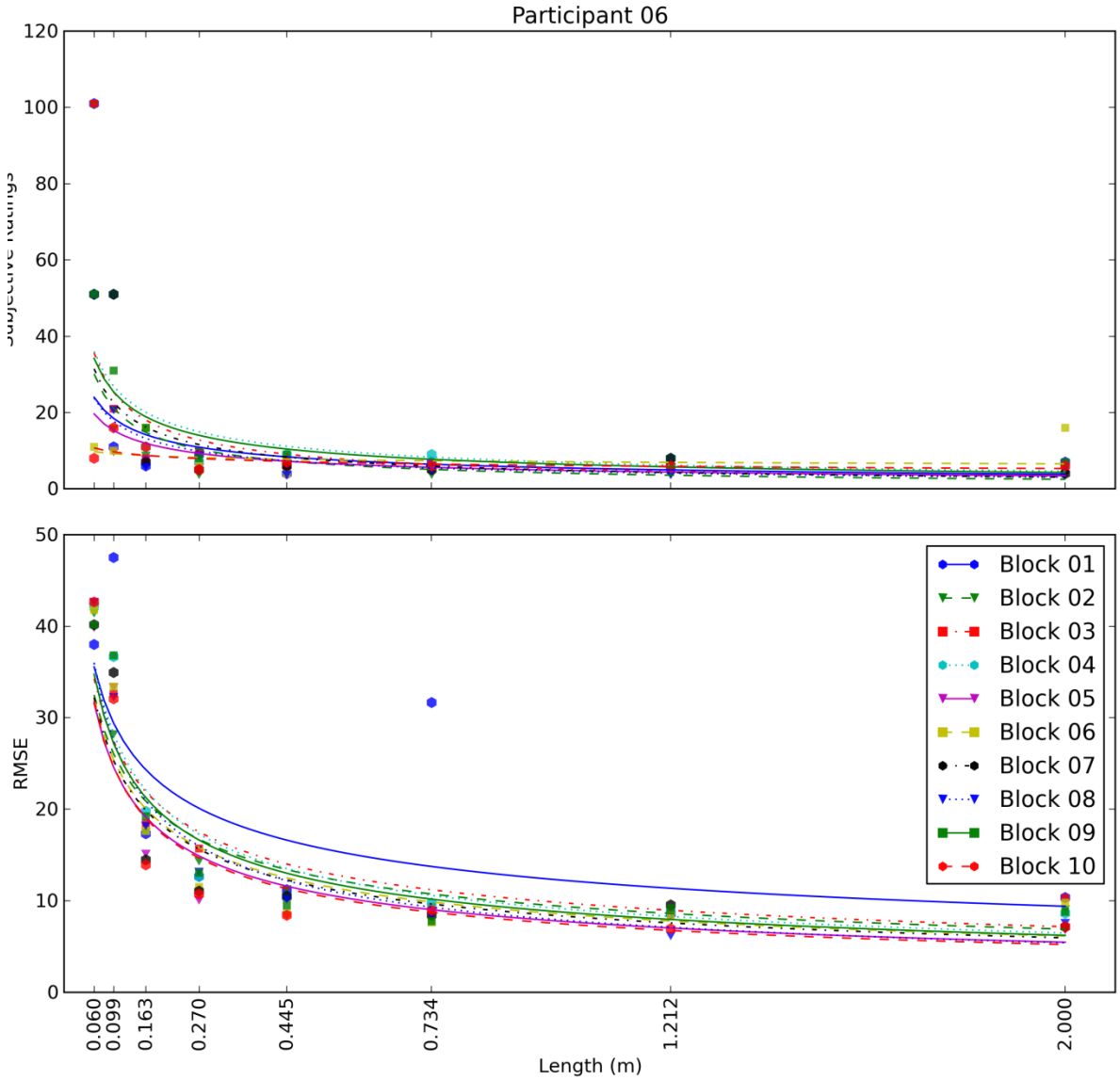


Figure 5.4.10 *Magnitude estimates for subjective ratings and RMSE for Participant 7.*

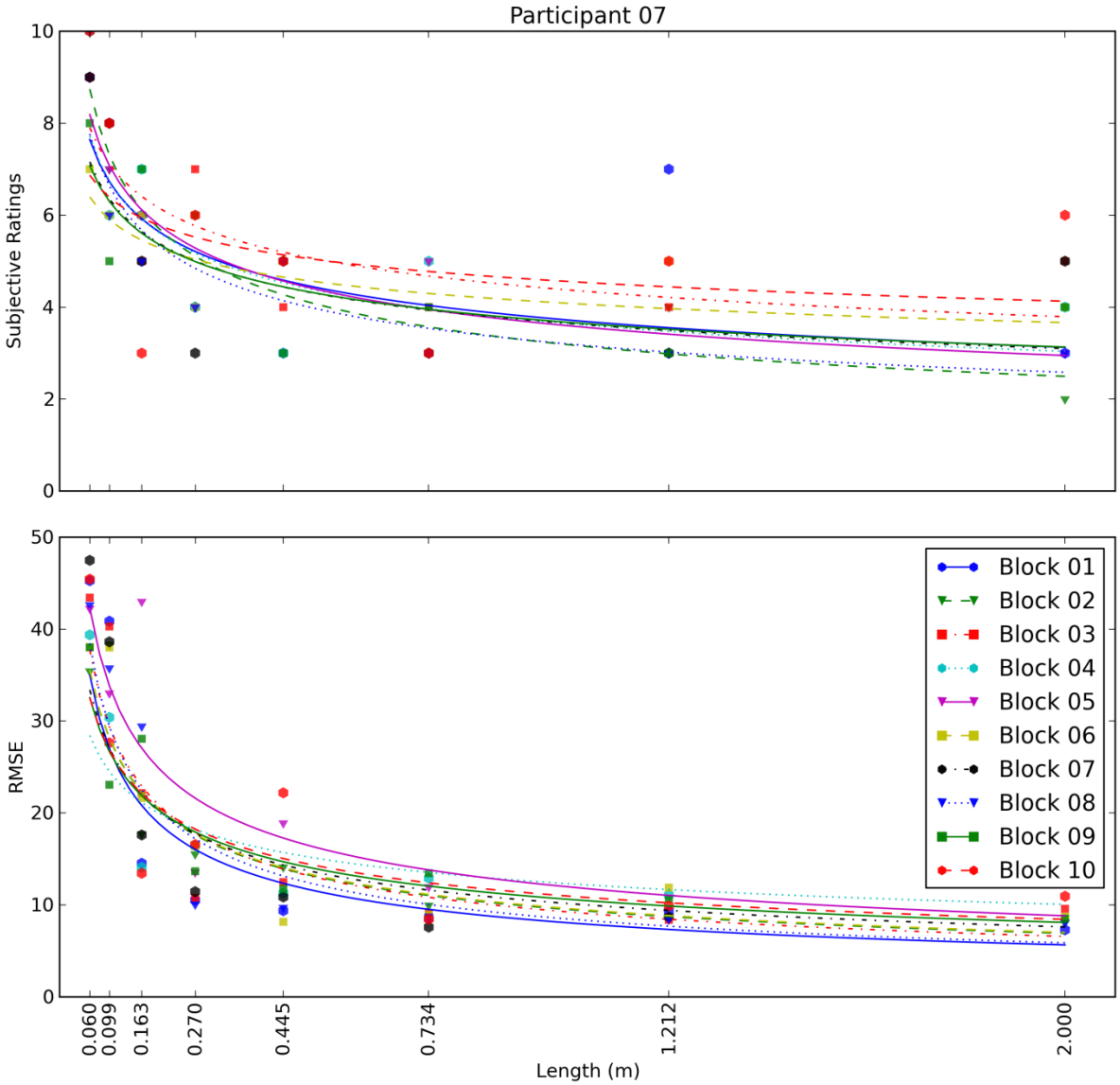


Figure 5.4.11 *Magnitude estimates for subjective ratings and RMSE for Participant 8.*

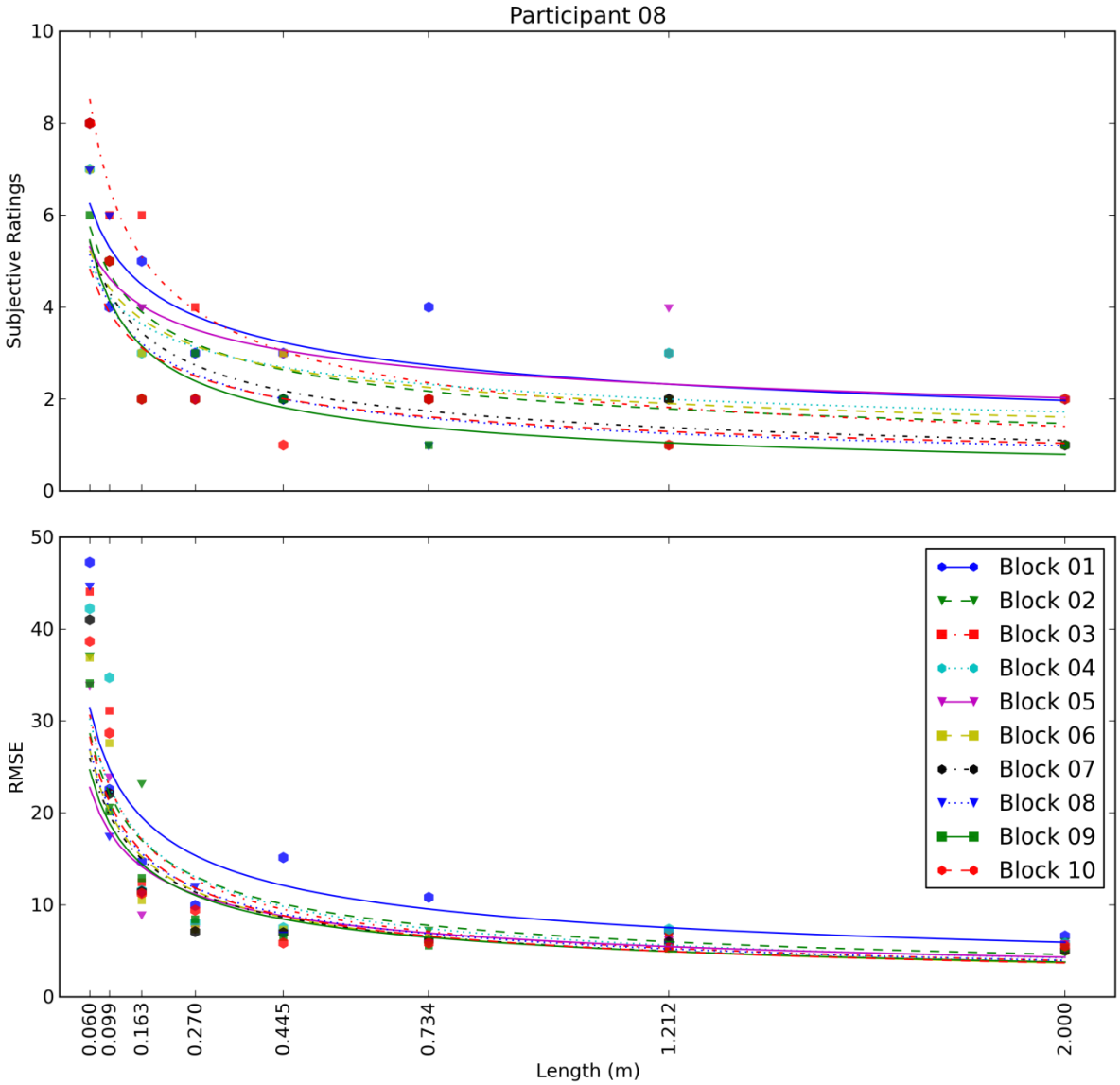


Figure 5.4.12 *Magnitude estimates for subjective ratings and RMSE for Participant 10.*

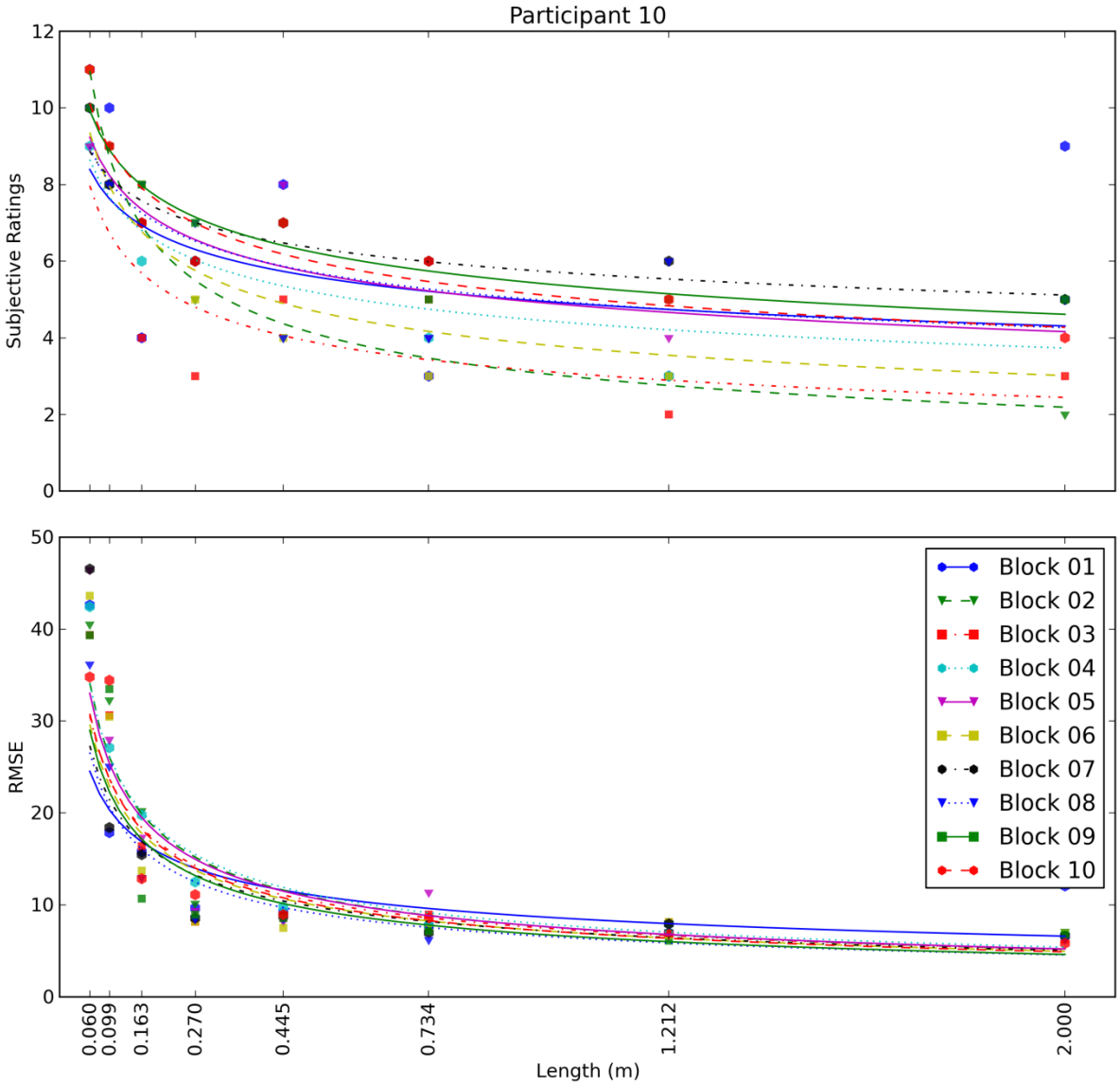
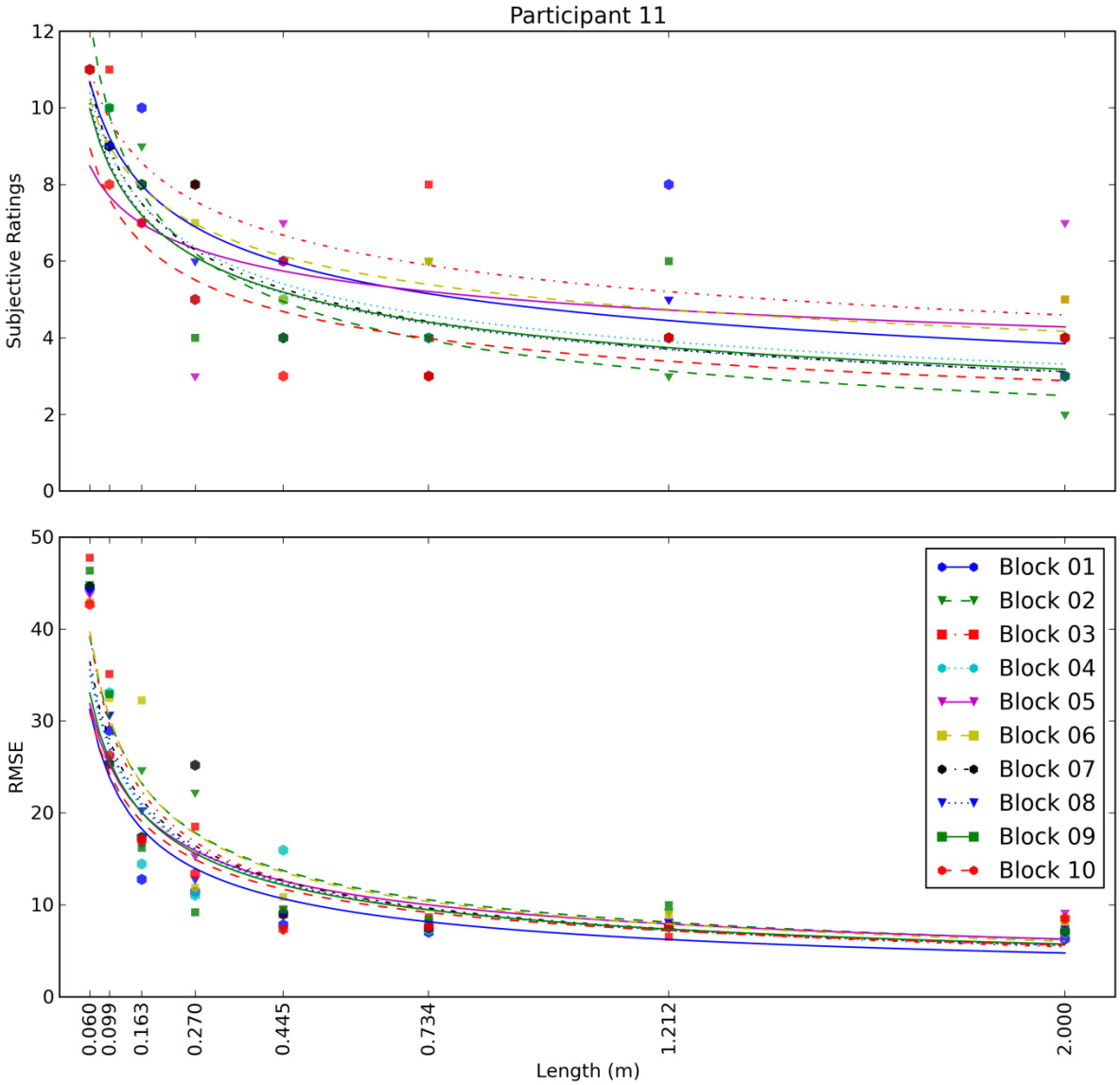


Figure 5.4.13 *Magnitude estimates for subjective ratings and RMSE for Participant 11.*

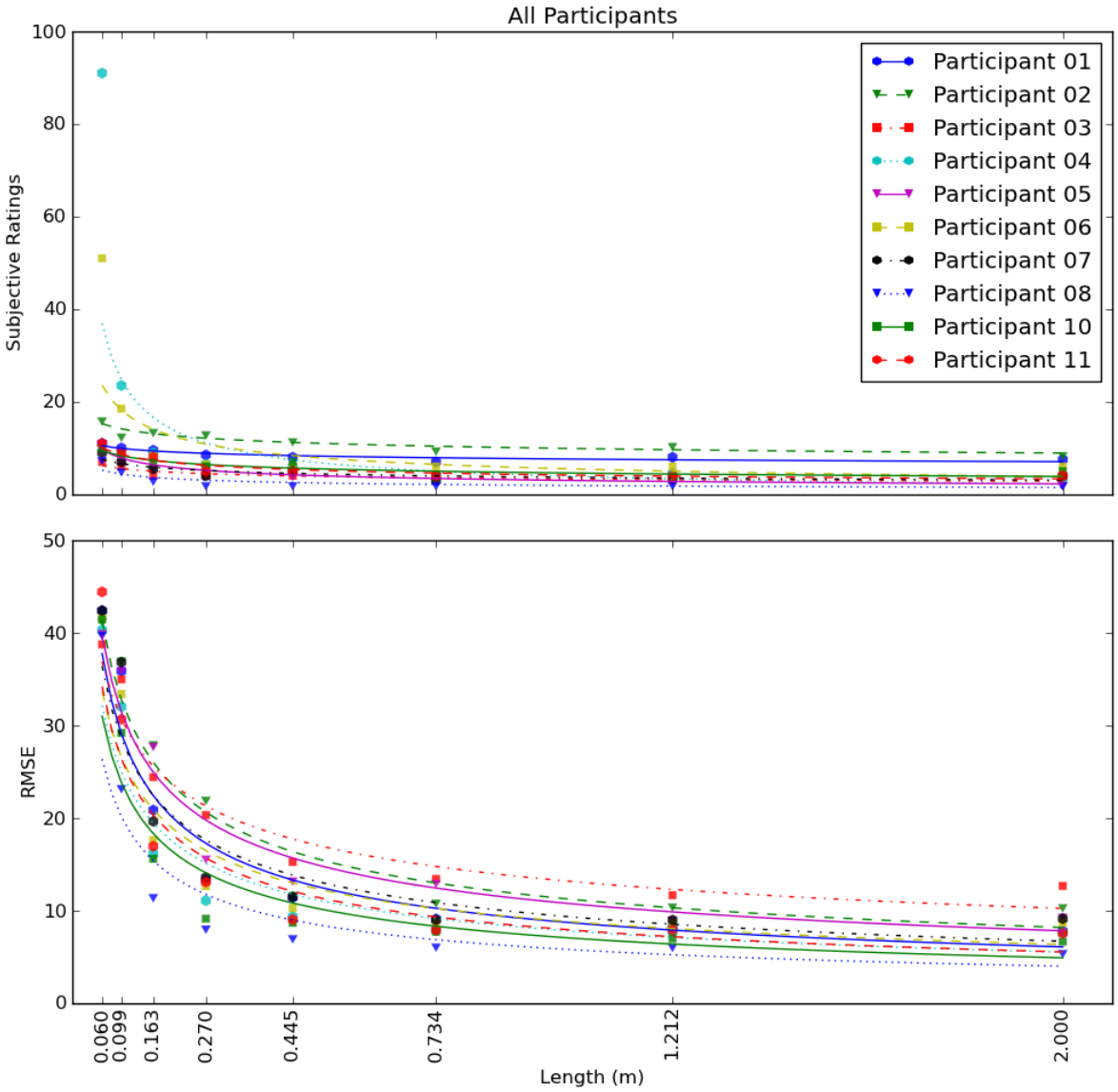


also important to quantify variability between participants since human ensembles are often non-ergodic. This claim suggests that the variability in a single subject over time may not be equal to the variability of a group of subjects at a single moment in time.

Assessing intersubject variability is slightly more complicated because the participant's internal rating scales are non-equivalent. Some restricted their responses between 1 and 10, while others choose values as high as 100. These differences in range should affect the *k* parameter but not the *a* parameter. Therefore, it is the variability of *a* that is of interest. The second concern is that the cognitive components that factor into intrasubject variability should be accounted for otherwise the measure would reflect total variability and not merely intersubject variability. The solution here is to collapse across blocks and build a unified model for each participant. The average *a* values reported in Table 5.4.5 could be interpreted as a block averaged estimate for each participant, but in the strictest sense it is more appropriate to first calculate medians for each participant at each length over the ten blocks and then use these medians to conduct a magnitude estimation for each participant (see Figure 5.4.14). The rationale being, participants may not necessarily have internalized their ratings as having an interval measurement scale. However, the instructions should ensure that they are at least using an ordinal scale where higher values reflect more workload. Medians are also more robust to the positive skew exhibited in the subjective rating responses.

The *a* parameters based on the median ratings are reported in Table 5.4.5. In many cases they are close to those found through averaging (in Table 5.4.4) although there are cases that differ by more than 10%. From Table 5.4.5 it is evident that the coefficient of variation across participants for the *a* parameter is 0.589. This is more than double the average intrasubject variability found across blocks (0.273). RMSE intersubject variability was also assessed based on median values across blocks. This data suggests that intersubject and intrasubject variability do not substantially

Figure 5.4.14 *Magnitude estimates collaborated across participants. Participant magnitude estimates are based on median subjective ratings and median RMSE.*

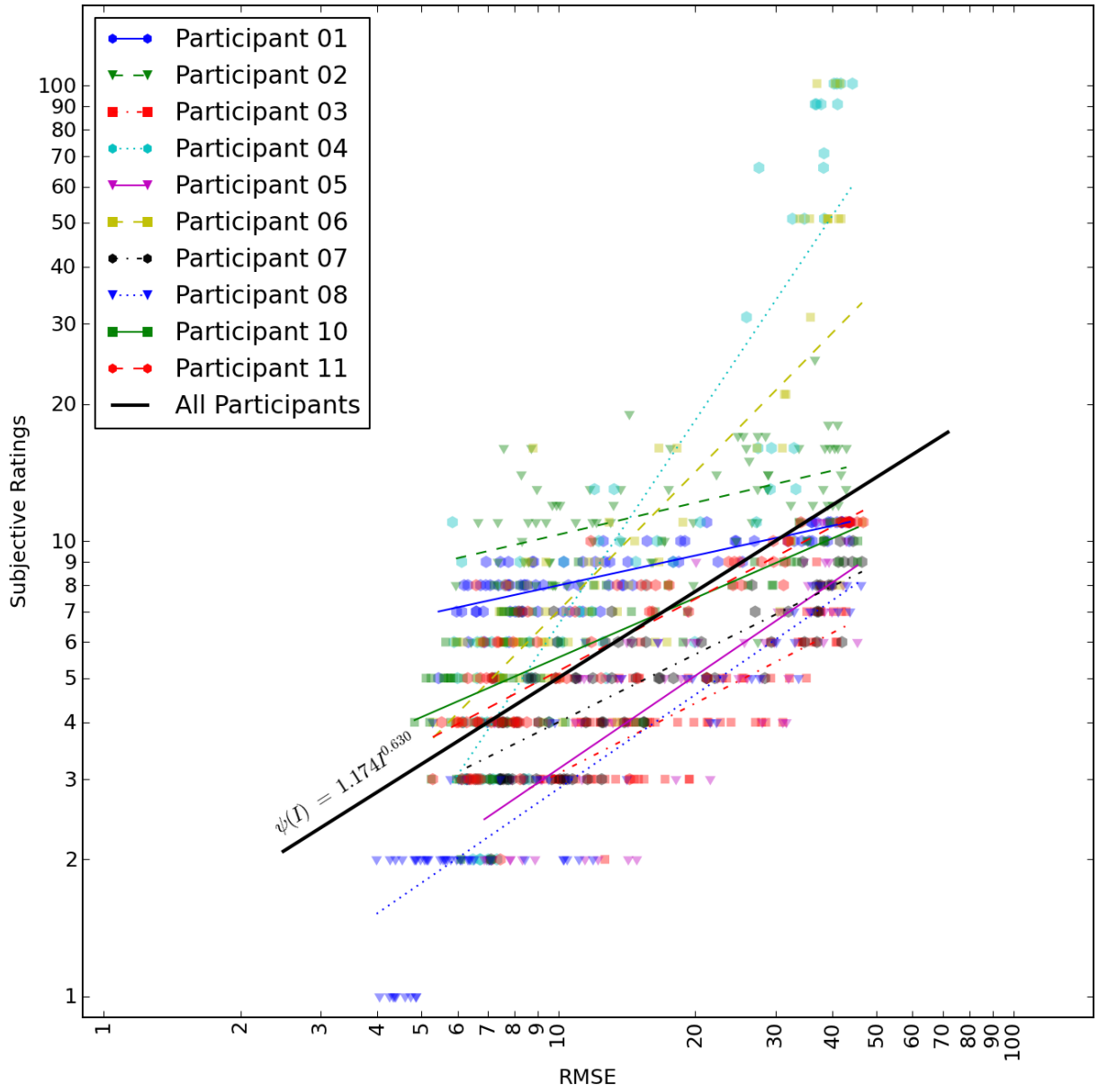


differ (see Figure 5.4.14). The average intrasubject coefficient of variability across blocks is 0.129 and the intersubject measures at 0.102. This suggests that while the internalized ratings scales vary more greatly between participants than within any given participant the variability tracking performance is surprisingly consistent between participants.

5.4.2.4 Intersubject magnitude estimation model. The end goal of this experiment is to identify a psychophysical transfer function that can be applied to a naïve subject pool. Although intersubject variability is more than double intrasubject variability a unified model should still be able to account for a large portion of subjective difficulty variability. A unified model was fitted by taking the geometric mean of the participant's medians across block. The geometric mean is used to combine data between participants since they have different internalized scales. Optimum k and a values were found to be 4.300 and -0.340 respectively. The model fit with an $R^2 = 0.865$ and was statistically reliable with a $F(1,6) = 9.274$ and $p = 0.023$.

5.4.2.5 Correlation between subjective difficulty and task performance. In the previous discussion of workload it was noted that subjective, physiological, and performance measures of workload may not be highly correlated. To indulge scientific curiosity, magnitude estimation can be applied to assess how changes in RMSEs relate to subjective ratings. Figure 5.4.15 depicts the results by participant as well as an overall fit across participants. The coefficients and accompanying inferential statistics for these regressions are listed in Table 5.4.6. These results show that RMSE could only predict 28% of the variability for participant 2, but could predict almost 80% for participant 4. On average, the individual models could account for 63% of the rating variability. Despite accounting for a good deal of variability of the ten individual models none were statistically reliable. An overall fit across participants was found by first normalizing the responses of each participant between 1 and 10 and then performing the least squares regression. The overall model could only account for 38% of the variability and was not statistically reliable ($F(1, 798) = 0.622$, $p = 0.431$).

Figure 5.4.15 *Correlations between subjective ratings and RMSE. In logspace power functions appear linear.*



5.4.3 *Conclusions and Discussion*

Task performance requires almost no training to reach asymptotic performance and the variability between participants was observed to be about equivalent to the variability within participants. The perceived difficulty over the first hour decreases slightly (from the first block to the tenth block) but gross changes in subjective difficulty were not observed at any given length. These results also suggest that in this context task performance is more sensitive than subjective reports at assessing task difficulty. The consistency in performance suggests that tracking ability is fairly uniform amongst the participant pool of young adults with normal vision and motor function. While these results are encouraging further work is needed before the physiological algorithms are applied to the compensatory tracking task.

As the results indicate performance and subjective ratings are fairly uniform until a critical instability is reached and performance quickly degrades. Workload theory would explain this by postulating that up until the point of instability cognitive resources are not being fully utilized. This results in the subjective ratings of difficulty and task performance having low sensitivity through a wide range of lengths and is evident as the stimulus compression exhibited in the magnitude estimation analyses. This line of research is postulating that physiological measures may be more sensitive than performance and subjective measures through this range. Before examining this hypothesis it is logical to examine whether incorporating a secondary task is sensitive to changes in pendulum length where tracking performance and subjective ratings are rather insensitive.

Table 5.4.1
Length (8) x Block (10) ANOVA results on subjective difficulty ratings and RMSE

Source	df_{source}, df_{error}	*F*	*p*	*MSE*	ε	η^2	*Obs. Power*
$\log_{10}$(ratings + 1)							
Length	7, 63	20.368	.001	.183	.152	0.532	0.850
Block	9, 81	1.395	.260	.026	.396	0.007	0.128
Length x Block	63, 567	1.389	.244	.015	.083	0.027	1.000
RMSE							
Length	7, 63	333.536	<.001	43.619	.149	5.094	1.000
Block	9, 81	0.319	.757	34.343	.252	0.005	0.003
Length x Block	63, 567	1.042	.410	15.380	.105	0.051	1.000

Table 5.4.2
Length (8) x Block (2) ANOVA results on subjective difficulty ratings and RMSE

Source	df_{source}, df_{error}	*F*	*p*	*MSE*	ε	η^2	*Obs. Power*
$\log_{10}$(ratings + 1)							
Length	7, 63	14.402	.001	.040	0.199	.429	0.957
Block	1, 9	7.145	.025	.020	1.000	.015	0.208
Length x Block	7, 63	0.373	.787	.026	0.460	.007	0.998
RMSE							
Length	7, 63	106.397	<.001	26.104	0.164	3.973	1.000
Block	1, 9	2.347	.160	12.854	1.000	0.010	0.080
Length x Block	7, 63	0.545	.656	23.095	0.430	0.018	1.000

Table 5.4.3
Magnitude estimation power exponents (a) for rating

Participant	*Average(a)*	*SD(a)*	*CV(a)*	R^2
1	-0.120	0.027	0.229	0.702
2	-0.139	0.047	0.335	0.438
3	-0.224	0.049	0.219	0.746
4	-0.803	0.158	0.197	0.680
5	-0.354	0.105	0.297	0.756
6	-0.515	0.201	0.391	0.553
7	-0.247	0.066	0.266	0.588
8	-0.405	0.094	0.232	0.700
10	-0.261	0.089	0.340	0.703
11	-0.311	0.070	0.226	0.721
Average:	-0.338	0.091	0.273	0.659

Table 5.4.4
Magnitude estimation results of the power exponent a for predictions of RMSE

Participant	*Average(a)*	*SD(a)*	*CV(a)*	R^2
1	-0.341	3.029	0.387	0.632
2	-0.486	2.142	0.705	2.387

Table 5.4.5
Magnitude estimation results based on the median ratings values collapsed across block

Participant	*k*	*a*	R^2	*F*	*p*
1	7.623	-0.113	0.822	4.628	0.075
2	9.907	-0.154	0.857	5.990	0.050*
3	3.314	-0.223	0.842	5.335	0.060
4	3.907	-0.798	0.783	3.610	0.106
5	3.005	-0.415	0.946	17.569	0.006**
6	5.507	-0.517	0.659	1.930	0.214
7	3.677	-0.255	0.726	2.655	0.154
8	1.905	-0.358	0.718	2.547	0.162
10	4.592	-0.257	0.758	3.139	0.127
11	4.152	-0.314	0.893	8.331	0.028*
Average:	4.759	-0.340	0.800		
Std. Dev.	2.377	0.200			
CV:	0.499	0.589			

*- reliable at $p < .05$, ** - reliable at $p < .01$. All F-tests had 1 and 6 degrees of freedom.

Table 5.4.6
Magnitude estimation results based on the median RMSE values collapsed across block

Participant	*k*	*a*	R^2	*F*	*p*
1	8.745	-0.520	0.925	12.401	0.012*
2	11.283	-0.462	0.909	9.936	0.020*
3	13.209	-0.366	0.932	13.640	0.010**
4	7.838	-0.502	0.862	6.243	0.047*
5	10.799	-0.463	0.936	14.600	0.009**
6	8.850	-0.475	0.862	6.243	0.047*
7	9.376	-0.482	0.878	7.180	0.037*
8	5.826	-0.537	0.834	5.032	0.066
10	7.087	-0.525	0.855	5.906	0.051
11	7.960	-0.518	0.884	7.588	0.033*
Average:	9.097	-0.485	0.888		
Std. Dev.	2.172	0.050			
CV:	0.239	0.102			

*- reliable at $p < .05$, ** - reliable at $p < .01$. All F-tests had 1 and 6 degrees of freedom.

Table 5.4.7
Magnitude estimation results predicting subjective ratings based on RMSE

Participant	*k*	*a*	R^2	*F*	*p*
1	4.841	0.218	0.681	2.135	0.148
2	6.038	0.233	0.280	0.390	0.534
3	0.948	0.513	0.646	1.824	0.181
4	0.212	1.491	0.797	3.919	0.051
5	0.661	0.680	0.646	1.825	0.181
6	0.681	1.014	0.637	1.754	0.189
7	1.280	0.495	0.655	1.900	0.172
8	0.588	0.687	0.762	3.197	0.078
10	2.041	0.435	0.537	1.162	0.284
11	1.545	0.526	0.693	2.253	0.137
Average:	1.884	0.629	0.633		
Std. Dev.	1.966	0.380	0.143		

*- reliable at $p < .05$, ** - reliable at $p < .01$. All F-tests had 1 and 78 degrees of freedom.

Appendix 5.4.A Consent Form

CONSENT FORM

Idaho Visual Performance Laboratory
Department of Psychology and Communication Studies
College of Liberal Arts and Social Sciences
University of Idaho
Control of speed during altitude changes

During this experiment you will be presented a display in a virtual environment. Various parameters of this display will be manipulated to examine stress and mental workload. In this experiment you will be asked to control movement in the virtual world using an input device such as a joystick.

The data you provide will be kept anonymous. There will be absolutely no link between your identity and your particular set of data.

Your participation will help increase knowledge of stress and mental workload. Subsequent to your participation the purpose and methods of the study will be described to you and questions about the study will be answered. It is our sincere hope that you will learn something interesting about your visual system from this debriefing.

The risks in this study are minimal, however displays simulating movement may on rare occasion cause motion sickness or eye fatigue in sensitive individuals. If at any time during the experiment you feel eye fatigue, dizziness, headache or nausea, please let the experimenter know immediately so that you can take a break before these symptoms become too intense. We endeavor to design our displays to minimize eye fatigue and motion sickness, and schedule periodic breaks to further reduce their occurrence. As a result, these phenomena have not been a common problem in previous similar studies.

Your participation will require **1** session of approximately **60** minutes. You may withdraw from this study at anytime without penalty. You will receive partial credit for your time spent. However, please be aware that your data is useful to us only if you complete the experiment in its entirety.
This research project has been approved by the University of Idaho Human Assurance Committee. As such, new information developed during the course of the research which may relate to your willingness to continue participation will be provided to you.

Thank you for your participation

Signature_____ Date _____

If you have further questions or encounter problems please contact:
Dr. Brian P. Dyre
(208) 885-6927
bdyre@uidaho.edu

Appendix 5.4.B Debriefing Form

Debriefing Form

Department of Psychology and Communication Studies

College of Letters, Arts, and Social Sciences

Physiological Workload Measures

Experiment 4a

Participant:____

Date:_____

1. How often do you play video games?
    a. What is your video game skill? (Bad, okay or good)
    b. Are you right or left handed?
2. Are you male or female?
3. Did you notice that some trials were more difficult than others?
4. Did you feel fatigued by the end of the experiment?
    a. If yes: Did you feel like fatigue influenced your performance?
5. Do you feel like your performance overall got better?
6. Did you have any eye-strain, fatigue, blurred vision, problems focusing on the target, etc. ?

Any additional comments

This experiment examines how varying parameters of the internal model of the dynamic system influences how difficult it is to control. This experiment also looks at how much your performance increases over time. These results are intended to help us manipulate task difficulty in future experiments.

Appendix 5.4.C Human Assurances Approval

University of Idaho

Office of Research Assurances
Institutional Review Board
PO Box 443010
Moscow ID 83844-3010

Phone: 208-885-6162
Fax: 208-885-5752
irb@uidaho.edu

To: Brian Dyre

From: Traci Craig, PhD
Chair, University of Idaho Institutional Review Board
University Research Office
Moscow, ID 83844-3010

IRB No.: IRB00000843

FWA: FWA00005639

Date: August 29, 2011

Title: 'Human Cognitive Workload and Perceptual Performance in Virtual Environments'

Project: 10-037
Approved: 09/28/11
Expires: 09/27/12

On behalf of the Institutional Review Board at the University of Idaho, I am pleased to inform you that the **first-year extension** of your proposal is approved as offering no significant risk to human subjects as no changes in protocol have been made on this project.

This extension of approval is valid until the date stated above at which time a second extension will need to be requested if you are still working on this project. If not, please advise the IRB committee when the project is completed.

Thank you for submitting your extension request.

Traci Craig

Traci Craig

5.5 Experiment 5: Compensatory Tracking with Random Number Generation

The previous experiment demonstrated that task difficulty in a compensatory tracking task can be manipulated by changing the length of the pendulum in the underlying physical model. These changes affect subjective ratings and tracking performance and subjective ratings and tracking performance covary with one another, but subjective ratings and tracking performance have low sensitivity when workload is underloaded. In this study a secondary task is incorporated as a means of assessing residual cognitive resources at varying pendulum lengths. In theory, a properly designed and implemented secondary task should be able to measure changes in workload by requiring participants to work at capacity throughout the duration of the experiment. Secondary task performance should decrease as residual capacity from performing the primary task decreases. The secondary task is also likely to interfere with the primary task and result in lower primary task performance.

In this experiment an externally guided (externally paced) verbal random number generation (RNG) task was incorporated. The RNG task requires participants to produce a random list of numbers from a finite set of digits. Superficially the task seems simple, but producing *random* sequences of digits or letters has been found to be taxing on cognitive resources. The rationale behind the task lies in Baddeley and Hitch's (1974) functional model of working memory. The model postulates a *central executive* controls and regulates cognitive processes. The central executive interacts with modality specific short-term memory stores known as the phonological loop and the visual-spatial sketchpad (Beech, 1984; Logie, 1995). Baddeley (1996) points to generalized impairments in both long-term and working memory as well as verbal and spatial reasoning in patients with Alzheimer's disease as evidence for the construct of a central executive. Using random sequence generation tasks are theorized to load the central executive. Robbins and others (1996) found that random letter generation interfered with selecting the appropriate chess move. Koike and other (2011) have used random number generation as a diagnostic for measuring

prefrontal cortex dysfunction. Generating random sequences requires storing recently elicited responses and suppressing automated sequences (Spatt, 1996).

Other studies have incorporated random tapping to load central executive resources (Zelanznik, Spencer, & Ivry, 2002; Noordzij, van der Lubbe, Neggers, & Postma, 2004). Compared to truly random sequences human responses are often more serialized. That is they show a tendency to count in ascending or descending order (e.g.: 6, 7, 8; 3, 2, 1). Responses also exhibit repetition avoidance. Computer generated random sequences generally show far more repetitions compared to their artificial counterparts. Lastly, humans have tend to cycle through the set of responses (e.g. drawing numbers out of a hat without replacement until that hat is empty, then putting all the numbers back in the hat and repeat). Norman and Shallice (1980) have proposed that a supervisory attentional system (part and parcel to the central executive) acts to suppress such habitual tendencies. In addition to increasing the non-randomness of sequences, increased workload also decreases the maximum generation rate (Baddeley, 1996). Baddeley also reports that the cardinality of the response set is negatively correlated with the maximum generation rate. For example, generating random sequences with letters (A-Z) should yield a slower maximum generation rate compared to digits (0-9).

Over the years a variety of secondary tasks have been developed to segregate hypothetical constructs. A task known as articulatory suppression in which participants repetitively vocalize an irrelevant phrase or word is thought to disrupt the phonological loop and severely disrupts the processing of visually presented words (Besner, 1987). Spatial tapping (manual tapping) tasks in which participants repetitively tap a fixed sequence of targets (usually arranged in a grid) are thought to disrupt the visual spatial sketchpad (Witt, Laird, & Meyerand, 2008). The critical tracking task used in the previous experiment likely has a visual spatial component, but it is also likely the task has central executive components. The system exhibits fairly complex dynamics not entirely unlike those found in process control tasks or driving. Petzoldt, Bär, and Krems (2009)

found that the critical tracking task could serve as a stand in for a "complex driving task" in the context of driver distraction. If the RNG does disrupt central executive resources and the RNG is found to interfere with the critical tracking task it would provide some evidence that the tracking task may generalize to process control settings (the ultimate aim) and the physiologically based measures of workload may also generalize. A secondary task that disrupts the visual-spatial sketchpad, such as spatial tapping, would likely interfere with the critical tracking task but would be speak less to the external validity of using the physiological workload measures in low-probability high-consequence settings.

5.5.1 *Method*

5.5.1.1 Participants. Twenty six participants with normal or corrected to normal Snellen visual acuity of 20/30 participated in this study. All were naïve to the hypotheses of the experiment. All participants were ethically treated in accordance with experimental protocols approved by the University of Idaho's Human Assurance Committee (see Appendices 5.5.A – 5.5.C).

5.5.1.2 Stimuli and Apparatus. The stimulus and apparatus remained largely identical to the previous experiment. A secondary random number generation task was incorporated into the critical tracking task described in the previous study. This was accomplished by providing an auditory pacing mechanism (metronome). During each trial a 100 ms pure tone of 880 Hz played at a rate of 60 beats per minute through headphones. A boom mic on the headset served to record the verbal responses of the participants for later transcription.

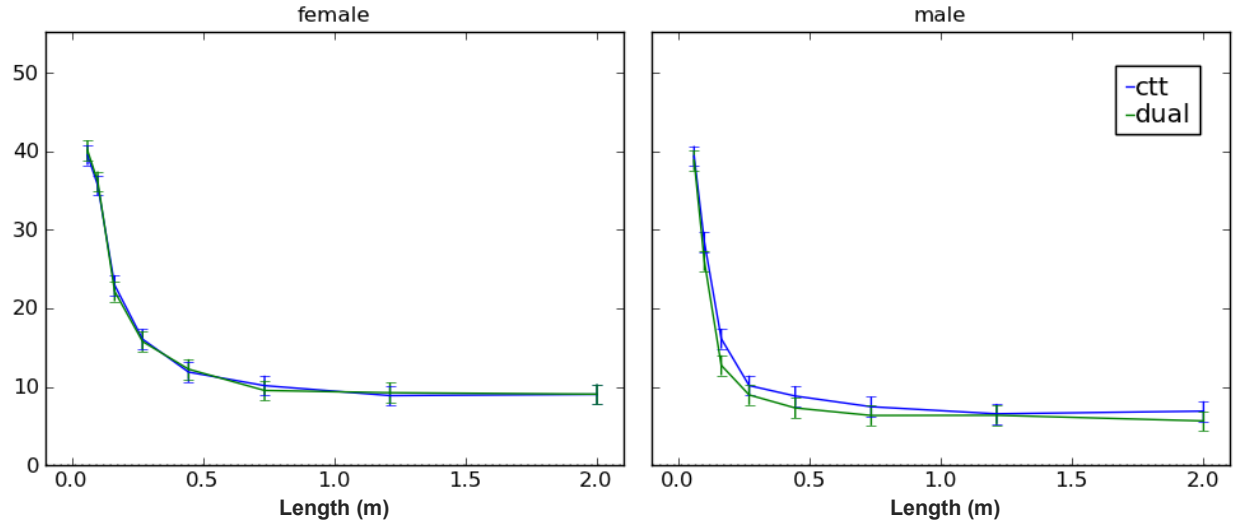
5.5.1.3 Procedure. Before simultaneously performing the random number generation task and critical tracking task participants learned these tasks independently. To avoid systematic carryover effects half of the participants first completed 24 trials of the generating random numbers followed by 24 trials of tracking. The remaining participants first trained with the tracking task. All of the trials were 30 seconds in duration. After participants were familiarized with both tasks they completed three blocks where they simultaneously performed both tasks across the

eight pendulum lengths described in the previous study. Participants reported subjective difficulty ratings using the free modulus method for trials involving the continuous control task. Participants were mandated to take short breaks between the training sessions and before the dual task session to reduce carryover effects related to fatigue.

5.5.2 *Results*

5.5.2.1 Effects of task (single vs. dual) and length on RMSE. The previous experiment examined the role of length on RMSE in fair amount of detail. The analysis presented here is primarily to examine the role of task. Contrary to my hypothesis performing both tasks simultaneously does not appear to result in a significant decrement in tracking performance. A 2 x 8 x 3 x 2 mixed analysis of variance (ANOVA) was conducted over the two levels of task (single, dual), eight pendulum lengths, three blocks, and gender (male, female; see Table 5.5.1 for summary). As anticipated, length accounts for the vast majority of the variability in RMSE. The analysis revealed a strong main effect of gender ($F(1, 24) = 20.037$, $p < .001$, $MSE = 314.461$) and a length by gender interaction ($F(7, 168) = 13.972$, $p < .001$, $\varepsilon = 0.156$, $MSE = 35.68$) indicating males overall have lower tracking error than females (16.86 vs. 17.524; see Figure 5.5.1). The pattern of means suggest that males are more affected by task than. To examine whether something systematic is at work with just the male participants a follow-up ANOVA with task, length, and block was conducted (see Table 5.5.1). This analysis found a main effect of task ($F(1, 11) = 10.925$, $p = .007$, $MSE = 27.967$) as well as a task by block interaction ($F(2, 22) = 6.325$, $p < .012$, $\varepsilon = 0.801$, $MSE = 11.362$). Closer examination indicates that tracking performance was actually worse when males performed the single task as opposed to the dual task. Interpretation must also reflect the fact that all participants encountered the single task trials before the dual task trials. With this in mind, the interaction is most likely due to a learning effect (see Figure 5.5.2). If performing both tasks simultaneously is taxing performance the effect is trivial in relation to the observed learning effect.

Figure 5.5.1 *RMS tracking error by length, gender, and task. Analysis revealed a main effect of length, a main effect of gender and a gender by length interaction. When only males are examined there is a reliable main effect of task but it is carried by a training effect. See Figure 5.6.2.2.*



Further restricting the analysis to only blocks two and three yields no reliable main effects or interactions involving task.

5.5.2.2 Random number generation dependent variables. In the literature a variety of algorithms exist for assessing the randomness of human generated sequences (Rabinowitz, 1970; Knuth, 1981; Evans, 1978; Kendall & Smith, 1938; Greenwood, 1955; Ginsburg & Wiegersma, Response bias and the generation of random sequences, 1991; Jahanshahi, Profice, Brown, Ridding, Dirnberger, & Rothwell, 1998). Ginsburg and Karpiuk (1994) conducted a factor analysis of ten of the commonly used metrics and found three underlying orthogonal factors for cycling, seriation, and repetition. To maximize power while minimizing familywise error, the varimax factor loadings published by Ginsburg and Karpiuk (1994) were used to calculate factor scores using the ordinary least squares method. These factor scores were used to assess the randomness of the participants verbalized digit sequences. Even though the random number generation task was externally paced participants sometimes did not emit precisely 30 digits or not all responses were in the digit set 0-9. Because making valid comparisons requires digit sequences of equivalent length all out of set responses were discarded, and sequences greater than 30 digits were truncated to 30 digits while sequences shorter than 30 digits were appended with digits randomly chosen with replacement from the digit set. After the factor scores were projected univariate analyses of variance were conducted.

5.5.2.3 Factor I (Cycling). The cycling scores indicate that participants were overall less random in the dual task conditions than the single task conditions ($F(1, 24) = 11.527$, $p = .002$, $MSE = 2.575$; see Figure 5.5.3; see Table 5.5.2). The analysis also found that males are less random compared to females ($F(1, 24) = 5.297$, $p < .03$, $MSE = 14.04$; see Figure 5.5.3). Although, the three way interaction between task, block, and length is not reliable the pattern of means could indicate that the factor scores for the dual task condition and block one exhibit a ceiling effect.

Figure 5.5.2 *RMS tracking error by block, task, and gender. The task main effect for male participants is trumped by a task by block interaction most likely due to improvement performance due to learning.*

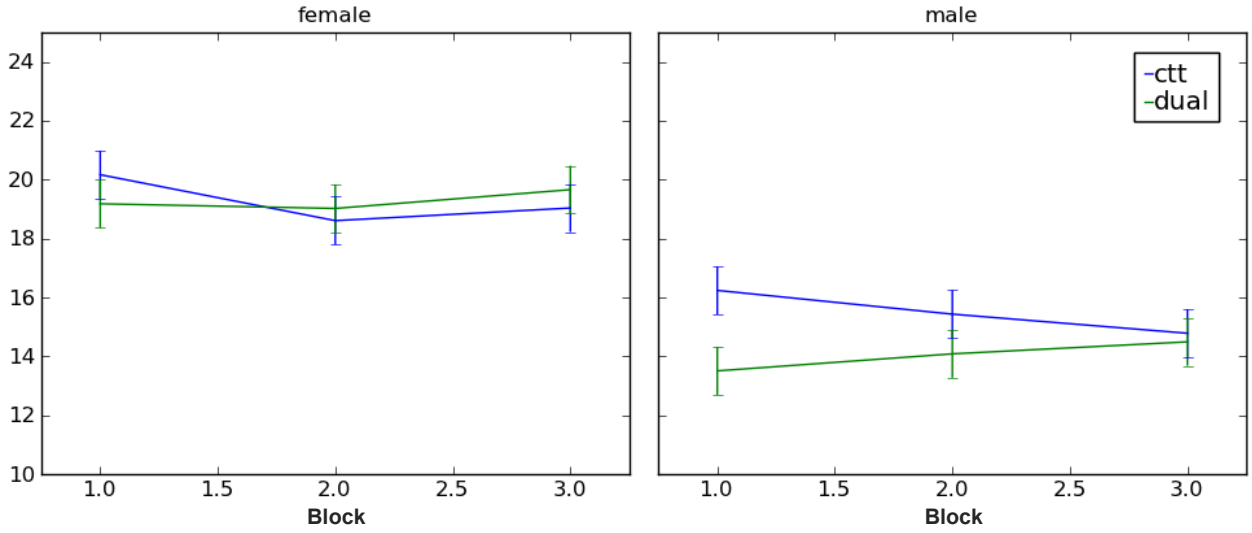
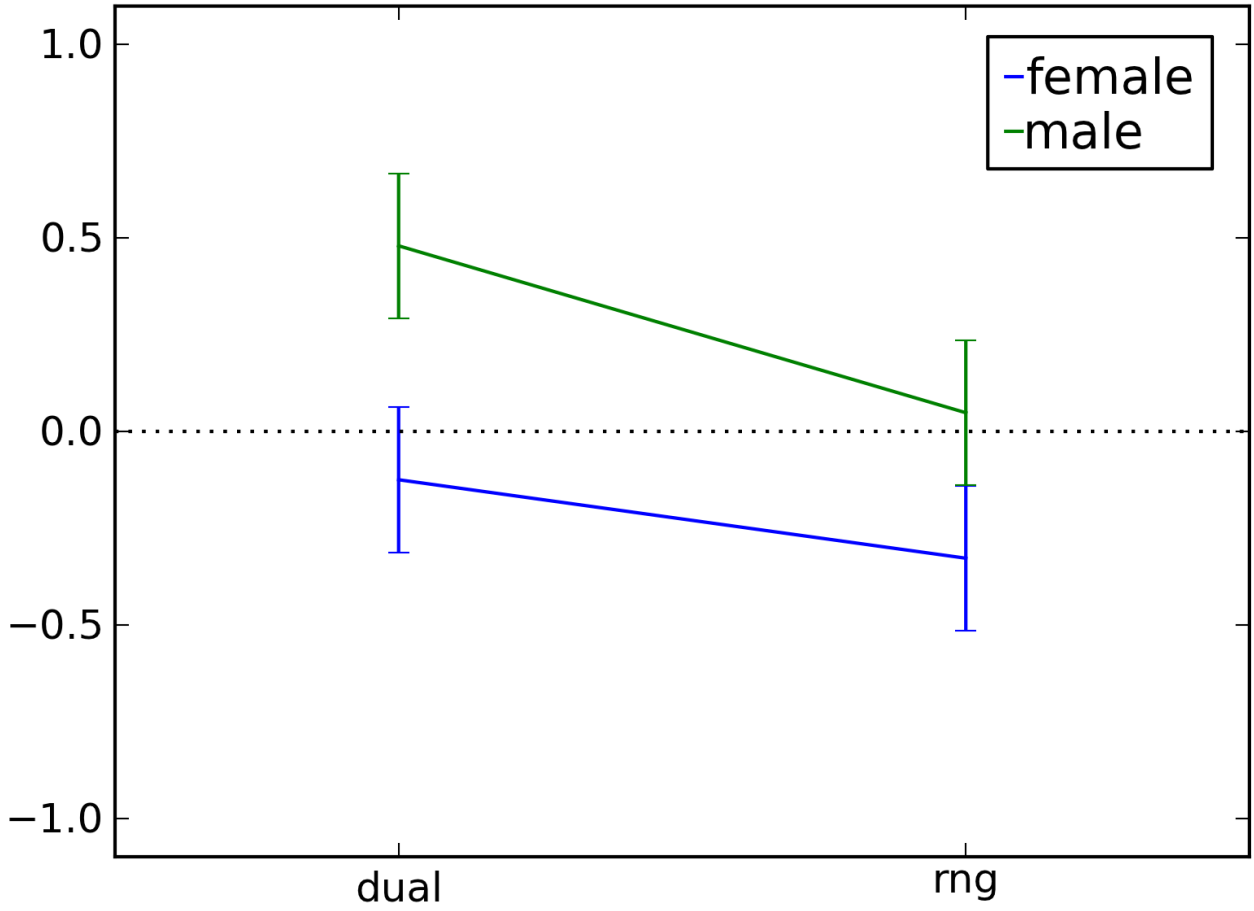


Figure 5.5.3 *Cycling by task and gender. The analysis of cycling factor scores found a reliable main effect of task as well as a main effect of gender.*



During the first block that the dual task is encountered cycling increases dramatically across all eight lengths (See Figure 5.5.4). To examine whether the length variable had a reliable effect on cycling after participants have some time to stabilize, a second ANOVA was conducted for the dual task trials over blocks 2 and 3 (see Table 5.5.2). This analysis found a reliable main effect of length ($F(7, 168) = 2.404$, $p = .043$, $\varepsilon = .687$, $MSE = .609$; see Figure 5.5.5). Although, the effect size is small (partial $\eta^2 = .091$), the pattern responses agree with the hypothesis that became responses should become less random at the more difficult tracking lengths. The follow-up analysis also found a main effect of block ($F(1, 24) = 4.691$, $p = .04$, $MSE = .850$). Responses were less random in the third block compared to the second block.

5.5.2.4 Factor II (Seriation). The omnibus analysis of the seriation factor scores found a main effect of task [$F(1, 24) = 5.836$, $p = .024$, $MSE = 1.632$], as well as a main effect of block [$F(2, 48) = 5.864$, $\varepsilon = .840$, $p = .008$, $MSE = .697$], and a task by block interaction [$F(2, 48) = 4.361$, $\varepsilon = .983$, $p = .019$, $MSE = .778$; see Figure 5.5.6]. See Table 5.5.3.The task by block interaction shows that responses were actually less random during the single task trials than the dual task trials. At first this is counterintuitive, but because all participants received the single task treatment first the most plausible explanation is that participants were self-monitoring and learned to reduce the amount of seriation in their responses by the time they encountered the dual task conditions. Self-monitoring counting behavior (a component of seriation) is more feasible than monitoring cycling behavior. Perhaps most importantly, the analysis also found a task by length interaction [$F(7, 168) = 3.151$, $p = .010$, $\varepsilon = .730$, $MSE = .505$; see Figure 5.5.7]. A follow-up ANOVA on just the dual task trials over all three blocks found a reliable main effect of length [$F(7, 168) = 3.863$, $p = .002$, $MSE = .537$]. As with the previous measure the pattern supports the hypothesis that increased tracking difficulty degrades randomness. The effect size is comparably small (partial $\eta^2 = .139$).

Figure 5.5.4 *Cycling by Task, Block, and Length.*
For the dual task conditions over the first block the cycling factor scores exhibit a ceiling effect.

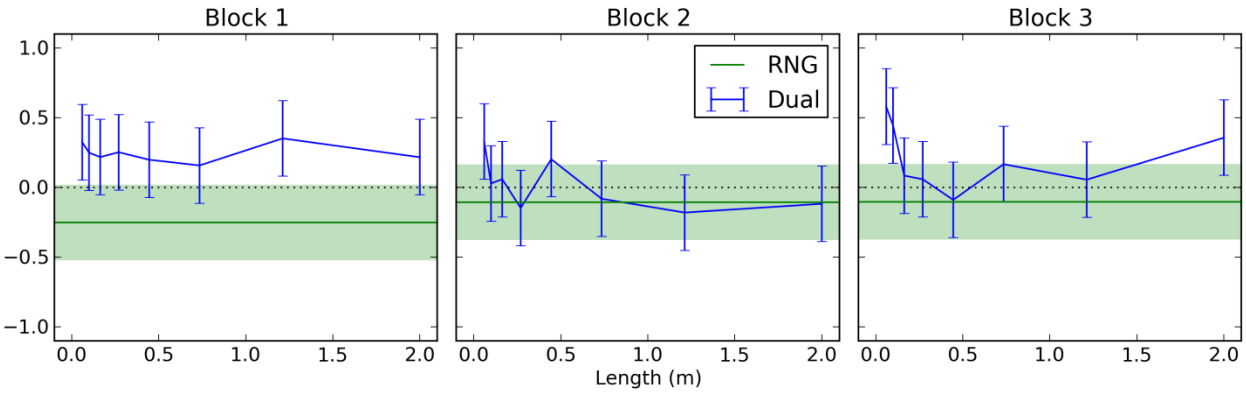


Figure 5.5.5 *Cycling by Length (Blocks 2 and 3, Dual task only). Main effect of length on cycling over blocks 2 and 3 and the dual task conditions.*

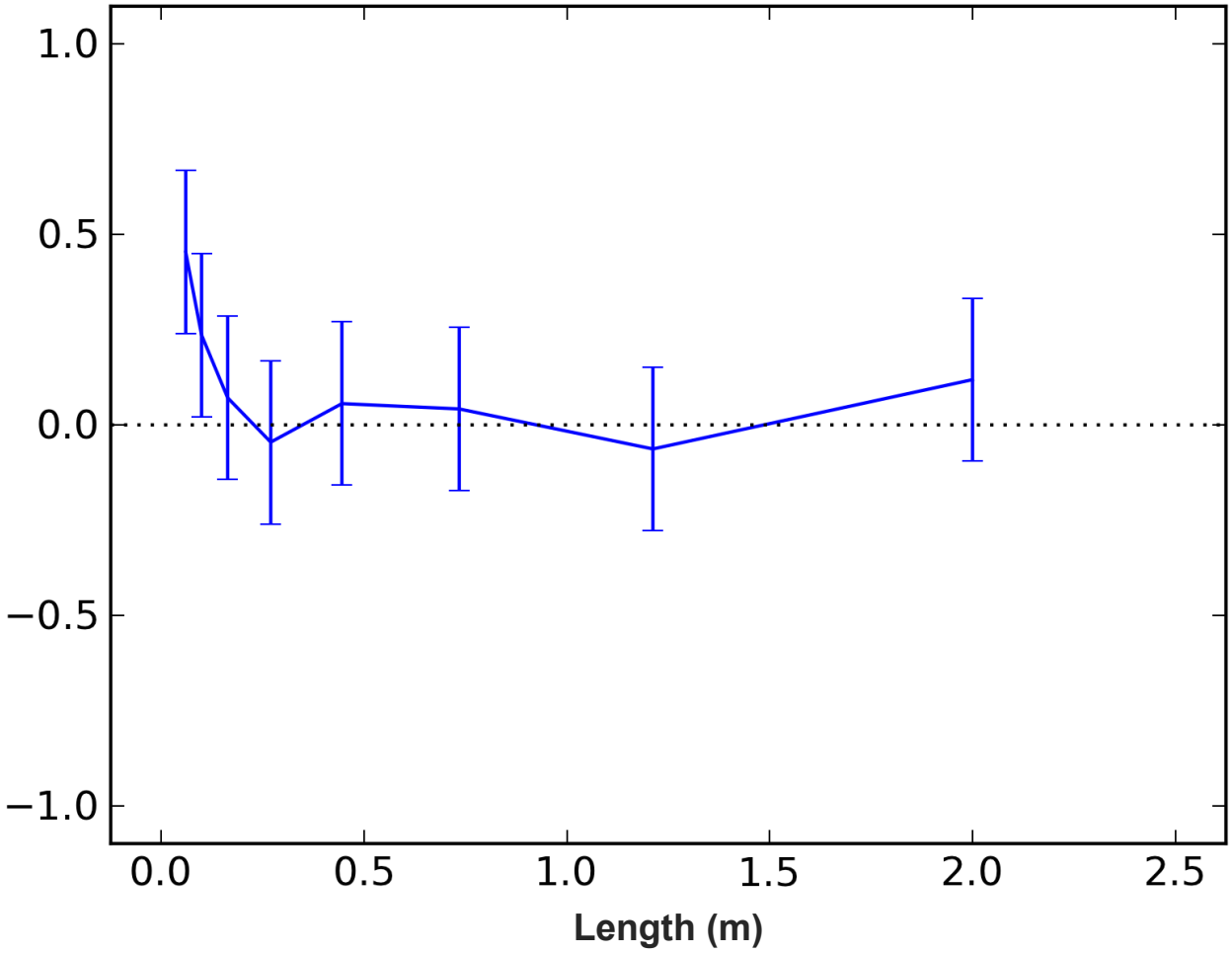


Figure 5.5.6 *Task by block interaction on the seriation factor scores.*

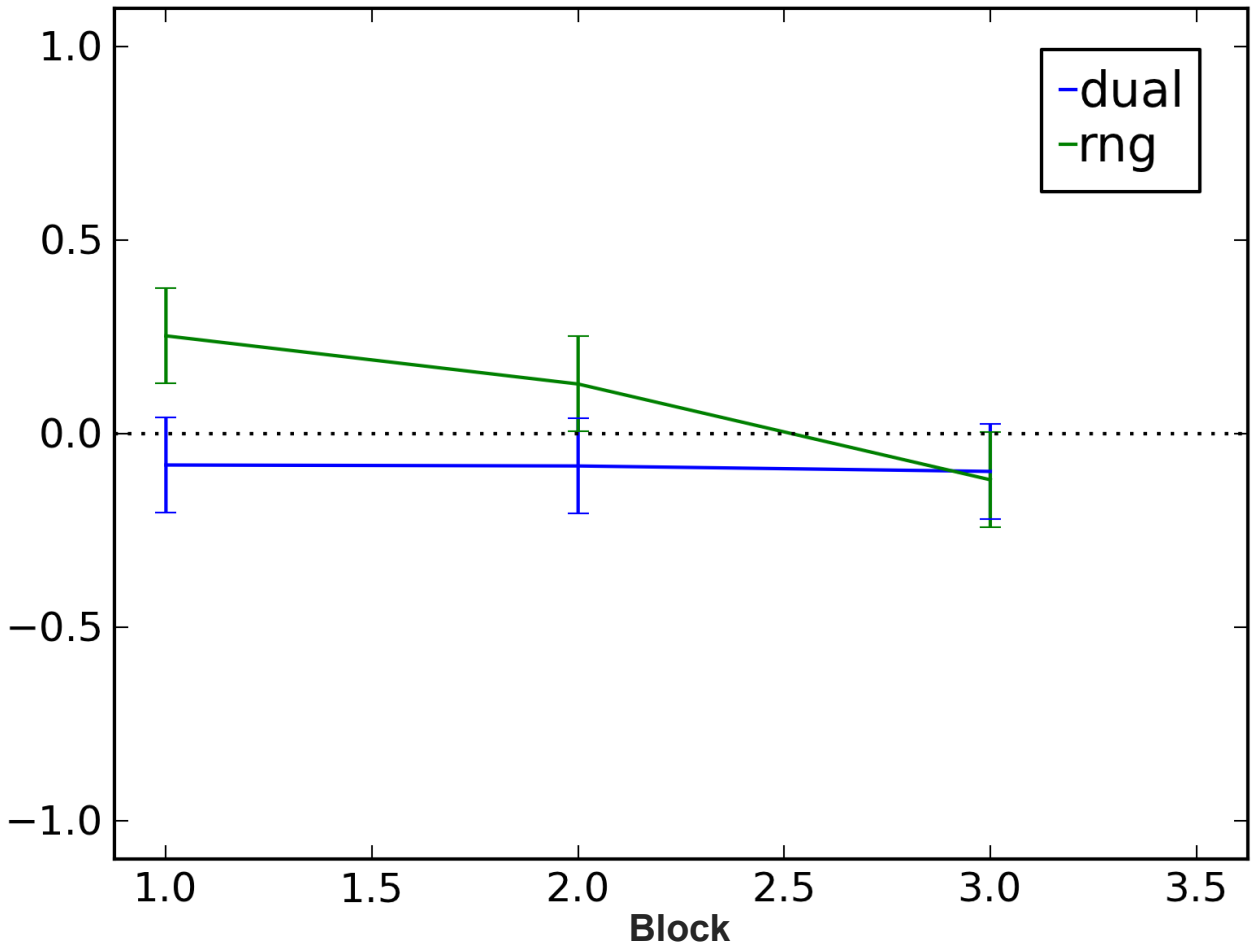
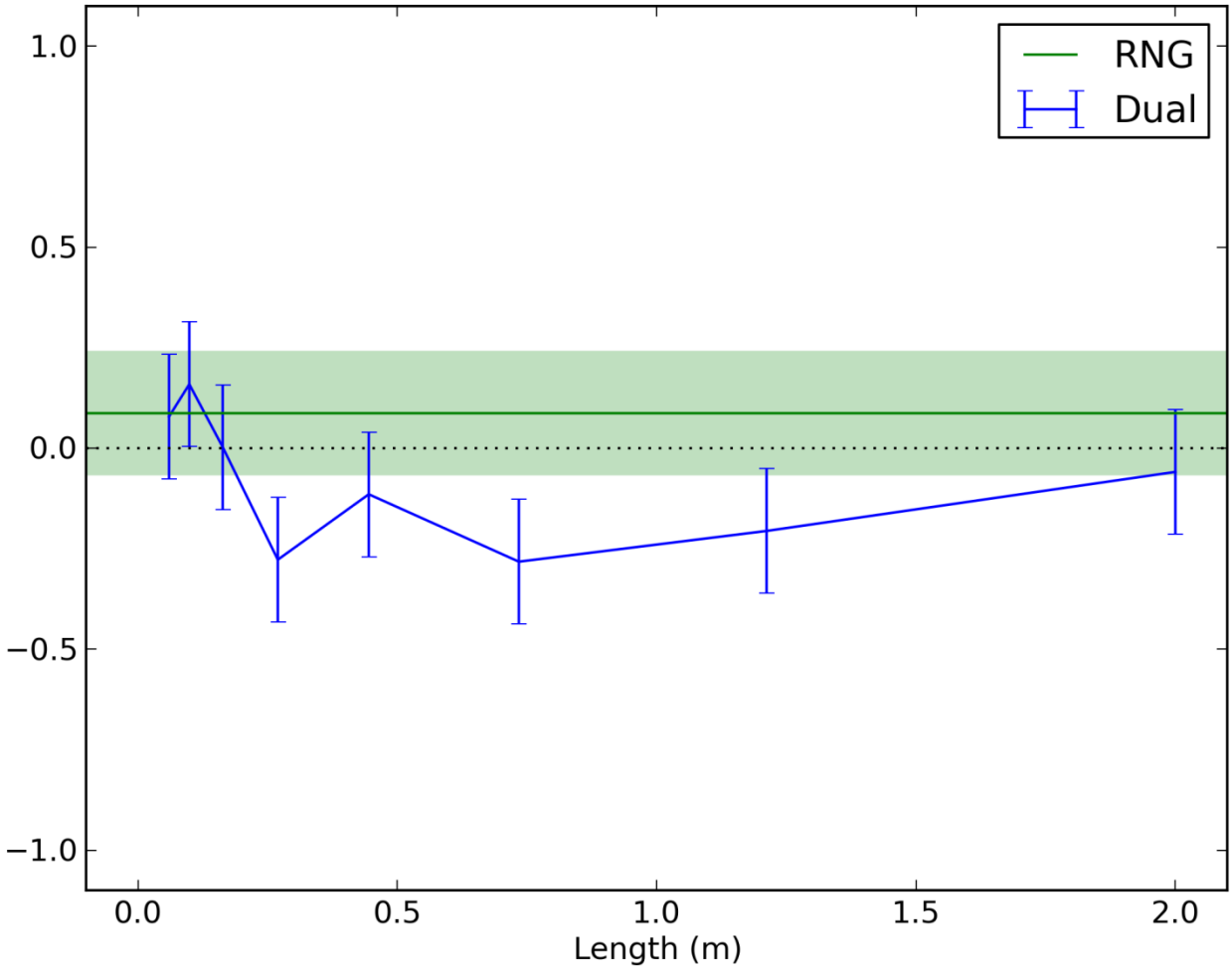


Figure 5.5.7 *Task by length interaction on the seriation factor scores.*



5.5.2.5 Factor III (Repetition). The analysis of repetition factor scores found a main effect of block [$F(2, 48) = 4.003$, $p = .030$, $\varepsilon = .874$, $MSE = .679$] and a surprising three way interaction between task, length, and gender [$F(7, 168) = 3.408$, $p = .009$, $\varepsilon = .650$, $MSE = .346$; see Figure 5.5.8]. When the analysis was restricted to just dual task trials the interaction between length and gender was also reliable [$F(2, 48) = 2.978$, $p = .025$, $\varepsilon = .733$, $MSE = .492$]. See Table 5.5.4. Because this interaction was so unexpected and difficult to interpret, a third ANOVA was conducted with only participants whose median factor scores were within 1 standard deviation of the overall median. This restriction excluded 3 participants in total. This analysis also found a reliable length by gender interaction is reliable [$F(7, 147) = 2.808$, $p = .023$, $\varepsilon = .671$, $MSE = .391$; see **Figure 5.5.9**]. The distinct peaks for males and females may be related to how the primary task is performed and perceived differently by the separate genders, but I am hesitate to speculate beyond that.

5.5.3 *Conclusions and Discussion*

In this study simultaneously performing a critical tracking task and externally paced random number generation task caused small but reliable increases in the amount of cycling and seriating behavior. Tracking performance degraded when male participants performed both tasks but no reliable differences were found when both genders were assessed. Such gender differences are not unprecedented (Petzoldt, Bär, & Krems, 2009), but the rationale behind the differences is sparse. The effect may be partially due to gender differences with video games. According to Terlecki and others (2010) males play video games more often and spend more time playing games. Males and females also show distinct preferences for the types of games they play. They also found that males were more confident in their game playing abilities. They also note that a moderate cohort of woman play video games with regularity.

Figure 5.5.8 *Repetition by task, length, and gender.*
Analysis of the repetition factor scores suggests a reliable 3 way interaction between task, length, and gender.

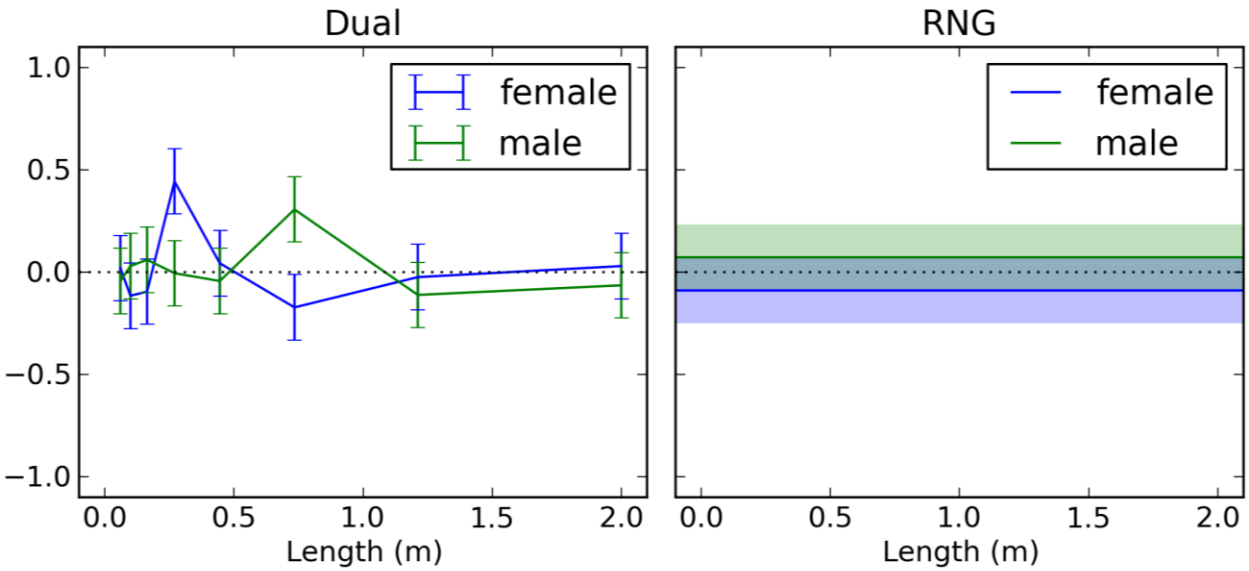


Figure 5.5.9 *Repetition by length and gender. Analysis of the repetition factor scores still found a reliable length by gender interaction even with the outlier participants removed.*

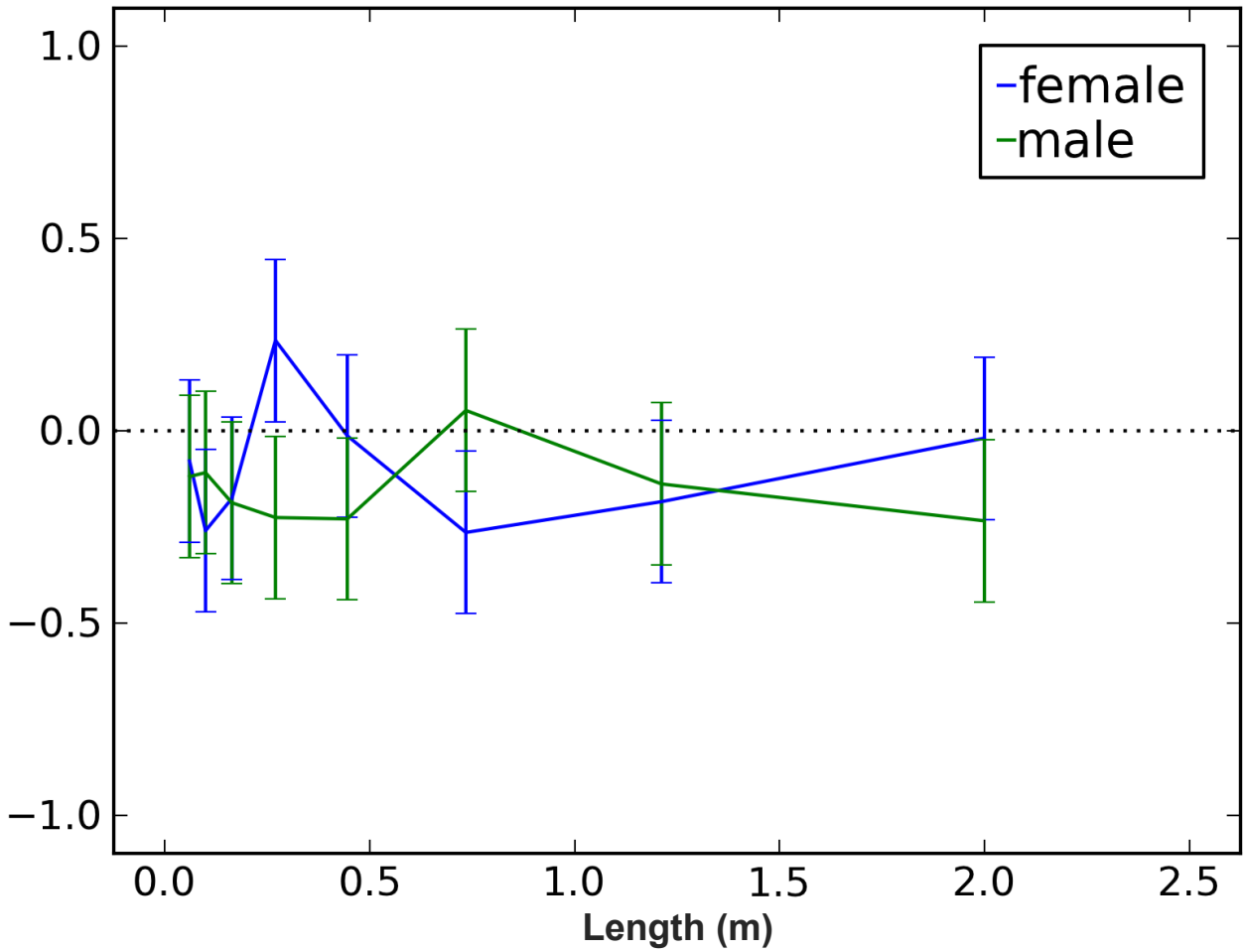


Table 5.5.1
ANOVA results on RMS tracking error

Source	df_{source}, df_{error}	*F*	*p*	*MSE*	ε	$G\,\eta^2$	*Obs. Power*
Males and Females							
Gender	1, 24	20.037	<.001	314.46	-	0.184	1.000
Length	9, 81	647.114	<.001	35.68	.156	0.853	1.000
Length x Gender	7, 168	13.972	<.001	35.68	.083	0.027	1.000
Task x Block	2, 48	6.55	.008	16.61	.717	0.008	.263
Males Only							
Task	1, 11	10.93	.007	27.97	1	0.027	0.257
Length	7, 77	241.98	<.001	43.57	.151	6.407	1.000
Task x Block	2, 22	6.33	.012	11.36	.801	0.012	.246
Length x Block	63, 567	1.042	.410	15.380	.105	0.051	1.000

Table 5.5.2
ANOVA results on Factor I (Cycling) RNG Performance

Source	df_{source}, df_{error}	*F*	*p*	*MSE*	ε	$G\,\eta^2$	*Obs. Power*
All Blocks, Both Tasks							
Gender	1, 24	5.30	.030	14.00	-	0.062	.090
Task	9, 24	11.53	.002	2.58	1	0.026	.999
Blocks 2 & 3, Dual Task Only							
Length	7, 168	2.40	.043	.609	.687	0.025	.566
Block	1, 24	4.70	.040	.850	1	0.010	.454

Table 5.5.3
ANOVA results on Factor II (Seriation) RNG Performance

Source	df_{source}, df_{error}	*F*	*p*	*MSE*	ε	$G\,\eta^2$	*Obs. Power*
Both Tasks							
Task	1, 24	5.84	.024	1.63	1	0.010	.899
Block	2, 48	5.86	.008	0.70	.840	0.009	.751
Task x Length	7, 168	3.15	.010	0.51	.730	0.012	.988
Task x Block	2, 48	4.36	.019	0.78	.983	0.007	.998
Dual Task Only							
Length	7, 168	3.86	.002	.709	.758	0.010	.856

Table 5.5.4
ANOVA results on Factor III (Repetition) RNG Performance

Source	df_{source}, df_{error}	*F*	*p*	*MSE*	ε	$G\,\eta^2$	*Obs. Power*
Both Tasks							
Block	2, 48	4.00	.030	0.679	.874	0.007	.668
Task x Length x Gender	7, 168	3.41	.009	0.346	.983	0.011	.989
Dual Task Only							
Length x Gender	7, 168	3.86	.002	.709	.758	0.010	.856
Dual Task Only, Outliers Excluded							
Length x Gender	7, 147	2.808	.023	.391	.758	0.010	.734

Appendix 5.5.A Consent Form

CONSENT FORM

Idaho Visual Performance Laboratory
Department of Psychology and Communication Studies
College of Liberal Arts and Social Sciences
University of Idaho
Control of speed during altitude changes

During this experiment you will be presented a display in a virtual environment. Various parameters of this display will be manipulated to examine stress and mental workload. In this experiment you will be asked to control movement in the virtual world using an input device such as a joystick.

The data you provide will be kept anonymous. There will be absolutely no link between your identity and your particular set of data.

Your participation will help increase knowledge of stress and mental workload. Subsequent to your participation the purpose and methods of the study will be described to you and questions about the study will be answered. It is our sincere hope that you will learn something interesting about your visual system from this debriefing.

The risks in this study are minimal, however displays simulating movement may on rare occasion cause motion sickness or eye fatigue in sensitive individuals. If at any time during the experiment you feel eye fatigue, dizziness, headache or nausea, please let the experimenter know immediately so that you can take a break before these symptoms become too intense. We endeavor to design our displays to minimize eye fatigue and motion sickness, and schedule periodic breaks to further reduce their occurrence. As a result, these phenomena have not been a common problem in previous similar studies.

Your participation will require **1** session of approximately **60** minutes. You may withdraw from this study at anytime without penalty. You will receive partial credit for your time spent. However, please be aware that your data is useful to us only if you complete the experiment in its entirety.
This research project has been approved by the University of Idaho Human Assurance Committee. As such, new information developed during the course of the research which may relate to your willingness to continue participation will be provided to you.

Thank you for your participation

Signature_____ Date _____

If you have further questions or encounter problems please contact:
Dr. Brian P. Dyre
(208) 885-6927
bdyre@uidaho.edu

Appendix 5.5.B Debriefing Form

Debriefing Form

Department of Psychology and Communication Studies

College of Letters, Arts, and Social Sciences

Physiological Workload Measures

Experiment 4b

Participant:____

Date:_____

1. How often do you play video games?
    a. What is your video game skill? (Bad, okay or good)
    b. Are you right or left handed?
2. Are you male or female?
3. Did you notice that some of the tracking trials were more difficult than others?
4. Did you use any particular strategy to generate random numbers?
5. Did you use any particular strategy to stabilize the dot?
6. When performing both tasks did you prioritize one more than the other?
7. Did you feel fatigued by the end of the experiment?
    a. If yes: Did you feel like fatigue influenced your performance?

8. Do you feel like performing both tasks was more difficult than performing the tasks independently?

9. Did you have any eye-strain, fatigue, blurred vision, problems focusing on the target, etc. ?

Any additional comments

This experiment examines how varying parameters of the internal model of the dynamic system influences how difficult it is to control. When the system is more difficult to control your verbal responses are hypothesized to become less random. These results are intended to help us manipulate task difficulty in future experiments.

Appendix 5.5.C Human Assurances Approval

University of Idaho

Office of Research Assurances
Institutional Review Board
PO Box 443010
Moscow ID 83844-3010

Phone: 208-885-6162
Fax: 208-885-5752
irb@uidaho.edu

To: Brian Dyre

From: Traci Craig, PhD
Chair, University of Idaho Institutional Review Board
University Research Office
Moscow, ID 83844-3010

IRB No.: IRB00000843

FWA: FWA00005639

Date: August 29, 2011

Title: 'Human Cognitive Workload and Perceptual Performance in Virtual Environments'

Project: 10-037
Approved: 09/28/11
Expires: 09/27/12

On behalf of the Institutional Review Board at the University of Idaho, I am pleased to inform you that the **first-year extension** of your proposal is approved as offering no significant risk to human subjects as no changes in protocol have been made on this project.

This extension of approval is valid until the date stated above at which time a second extension will need to be requested if you are still working on this project. If not, please advise the IRB committee when the project is completed.

Thank you for submitting your extension request.

Traci Craig

Traci Craig

5.6 Experiment 6: Compensatory Tracking with continuously Varied Difficulty

The previous experiment established that discretely varying the length of a simulated pendulum across short 30 second trials significantly affected tracking performance. The previous experiment also established that introducing random number generation as a secondary task produced subtle yet reliable degradations in random number generation performance. Here I examine whether these same degradations occur when pendulum length is manipulated in a continuous fashion. With real-world tasks difficulty may change in a continuous fashion without clear demarcations between transitions. An ideal measure of cognitive workload should be able measure both the absolute cognitive workload as well as its derivative as this will allow for leading measures of workload.

5.6.1 *Method*

5.6.1.1 Participants. Twenty three participants with normal or corrected to normal Snellen visual acuity of 20/30 participated in this study. All were naïve to the hypotheses of the experiment. All participants were ethically treated in accordance with experimental protocols approved by the University of Idaho's Human Assurance Committee (see Appendices 5.6.A – 5.6.C).

5.6.1.2 Stimuli and Apparatus. As with the previous experiment participants performed a critical tracking task. Here, we varied the length of the pendulum in a manner specifically designed to produce sinusoidal changes in subjective workload with a period of 120 seconds. Pendulum lengths were chosen such that the amplitude of subjective workload varied between what most people would consider a *moderate difficulty* (a rating of 6) and what most people would consider *very difficult* (a rating of 11). This is possible by using the inter-participant magnitude estimation model obtained from Experiment 5.5 describing how the length of the pendulum correlates to subjective difficulty across participants. This can be more precisely expressed as a function of time,

$$L(t) = \frac{98.3316}{\left(\left(-\cos\left(2\pi\left(\frac{t}{120}\right)0.5+0.5\right)(11-6)+6\right)\right)^{50/17}}$$

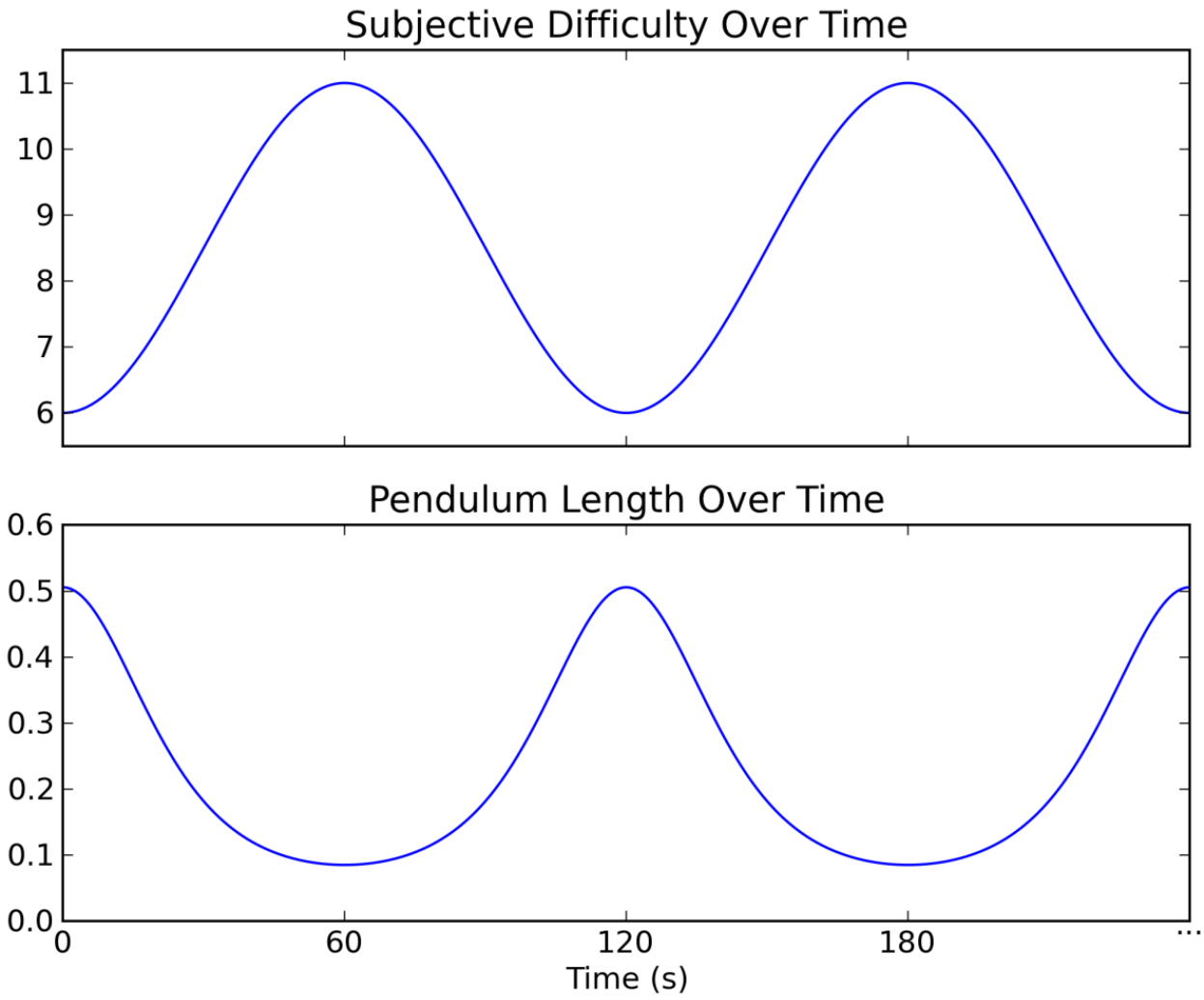
At its shortest (when $(t + 60) \bmod 120 = 0$) the simulated pendulum is 0.085 meters in length. This is between the shortest and second shortest lengths used in Experiments 5.4 and 5.5. The pendulum at its longest is 0.506, which is just beyond the 5th length used in Experiments 5.4 and 5.5. Graphically the length resembles an inverted cosine whose troughs have been widened (see bottom panel of Figure 5.6.1). Participants also performed the paced random number generation task previously described.

5.6.1.3 Procedure. The procedure was similar to the previous experiment. Before simultaneously performing the random number generation task and critical tracking task participants learned these tasks independently. To avoid systematic carryover effects half of the participants first completed a ten minute trial of the generating random numbers followed by ten minute trial of tracking. The remaining participants trained with the tracking task first and then the random number generation task. After participants were familiarized with both tasks they completed a ten minute trial where they simultaneously performed both tasks. Participants were mandated to take short breaks between trials to reduce carryover effects related to fatigue.

5.6.2 *Results*

5.6.2.1 Effects of task (single vs. dual) on tracking performance. The previous experiment conducted time series analyses comparing single and dual task tracking performance and failed to find any generalizable differences (males were better with the dual task, although participants always received the dual task treatment last). In this experiment the increased trial duration makes examining performance at the disturbance frequencies possible. In addition to countering the positive feedback effects of gravity, participants also had to manage a five component sum-of-sines disturbance with frequencies of 0.13, 0.21, 0.47, 1.13, and 1.93, and respective amplitudes if 30, 22, 18, 12, and 6. When the Fourier components of pendulum angle at these frequencies are examined the four lower frequencies are unequivocally reduced in the dual task trials compared to the single task trials (4 paired t-tests all *p-values* < 1e-9; See Figure 5.6.2 and Table 5.6.1).

Figure 5.6.1 *Subjective difficulty manipulation. Top panel illustrates how subjective difficulty should change over time. The bottom panel depicts how pendulum length varied over time.*



This suggests participants are better at nulling the disturbance, and agrees with the reduction in RMS error exhibited by the male participants in the previous study. At the highest disturbance frequency component there is a reversal. Power is reliably lower for the single task condition [$t(20) = 10.78$, $p = 4e\text{-}10$]. Examining the number of times the participants allow the target to move off-screen (a measure of instability) reveals some insight. Twenty of 23 participants had the target move off screen more times in the dual task than the single task. With the single task the target moves off screen an average of 32.5 ($SD = 45.7$) times per trial, with the dual task the target moves off an average of 75.5 (70.6) times. When these frequency counts are subjected to a two-tailed independent samples t-test this difference is found to be reliable [$t(34) = -2.35$, $p < 0.025$]. Taken together this suggests that participants improve on the *pursuit* aspect of the task but are worse on the *compensatory* aspect of the task. Performing the dual task essentially limits the band that they can maintain stability. The fact that performance improves in one regard while it degrades in another may also explain the insensitivity of RMSE with task differences.

5.6.2.2 Calculating running random number generation factor scores. To examine whether randomness correlates with the continuously changing difficulty I computed running measures of randomness. This process first requires replacing the participants' out of set responses (e.g. 10, 11, …) and filing in non-responses. For each participant this was accomplished by first removing out of set responses and non-responses and building a first order Markov model based on their remaining response. The Markov model was then used to fill in their remaining responses. Running measures of cycling, seriation, and repetition were then calculated using a moving window of 30 digits and the algorithms and factor loadings presented in the previous experiment. The measures that feed the factor loadings can be heavily influenced by small differences between sequences. To eliminate such biases, the running factor scores used in the subsequent analysis are averaged over 20 of runs. The resulting time-series measures are presented in Figure 5.6.3 and Figure 5.6.4.

Figure 5.6.2 *Power spectral densities of pendulum angle by task. PSDs were obtained with a NFFT of 2^15 and a Blackman-Harris 4 window. Lower indicates better performance.*

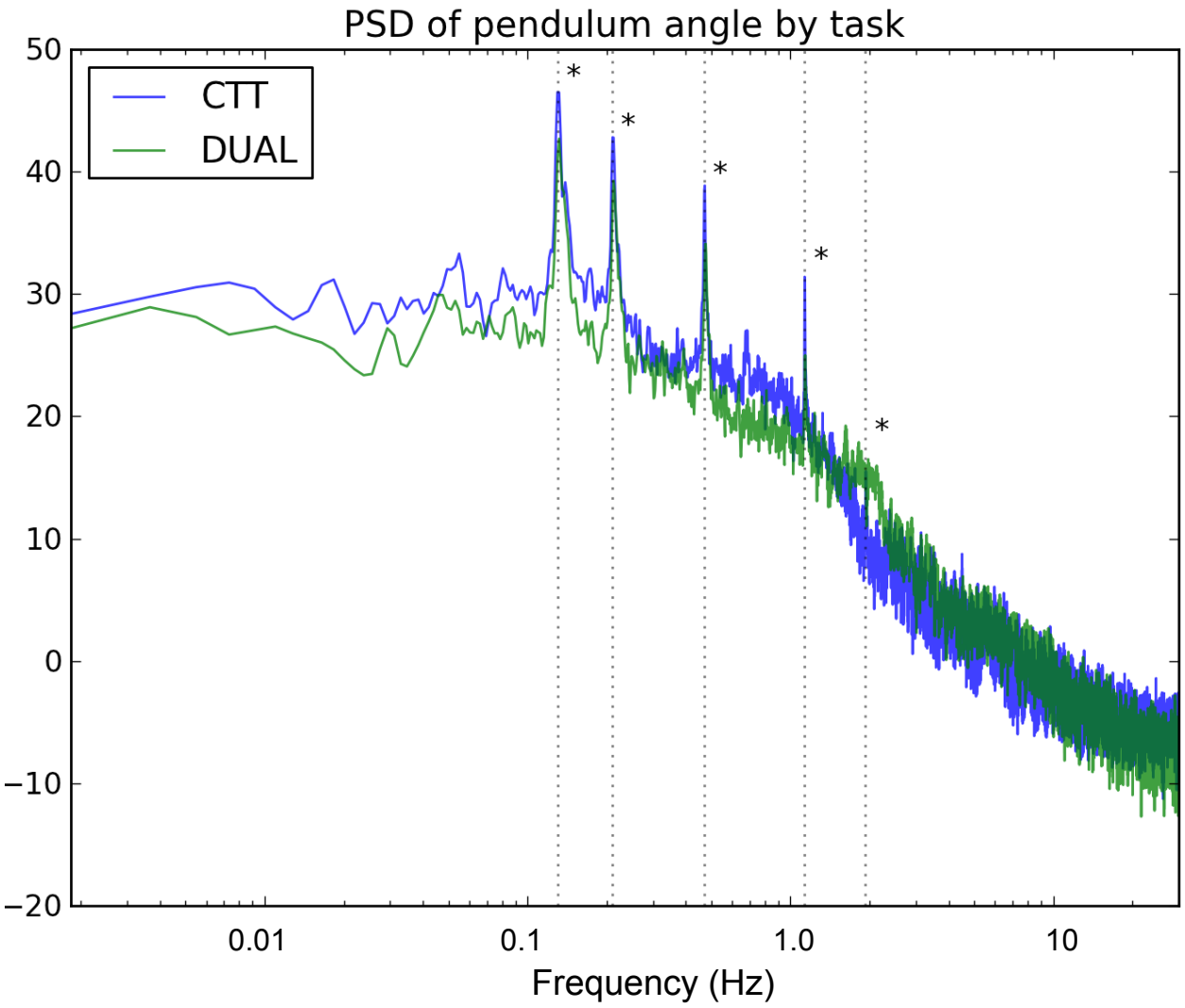


Figure 5.6.3 *RNG performance over time for RNG only.The panels reflect RNG performance during the RNG only trial. The top panel reflects cycling performance, the middle panel reflects seriation, and the bottom panel reflects repetition. The green trace depicts how difficulty changed during the CTT only and dual task only trials. The factor scores should be uncorrelated to the green trace.*

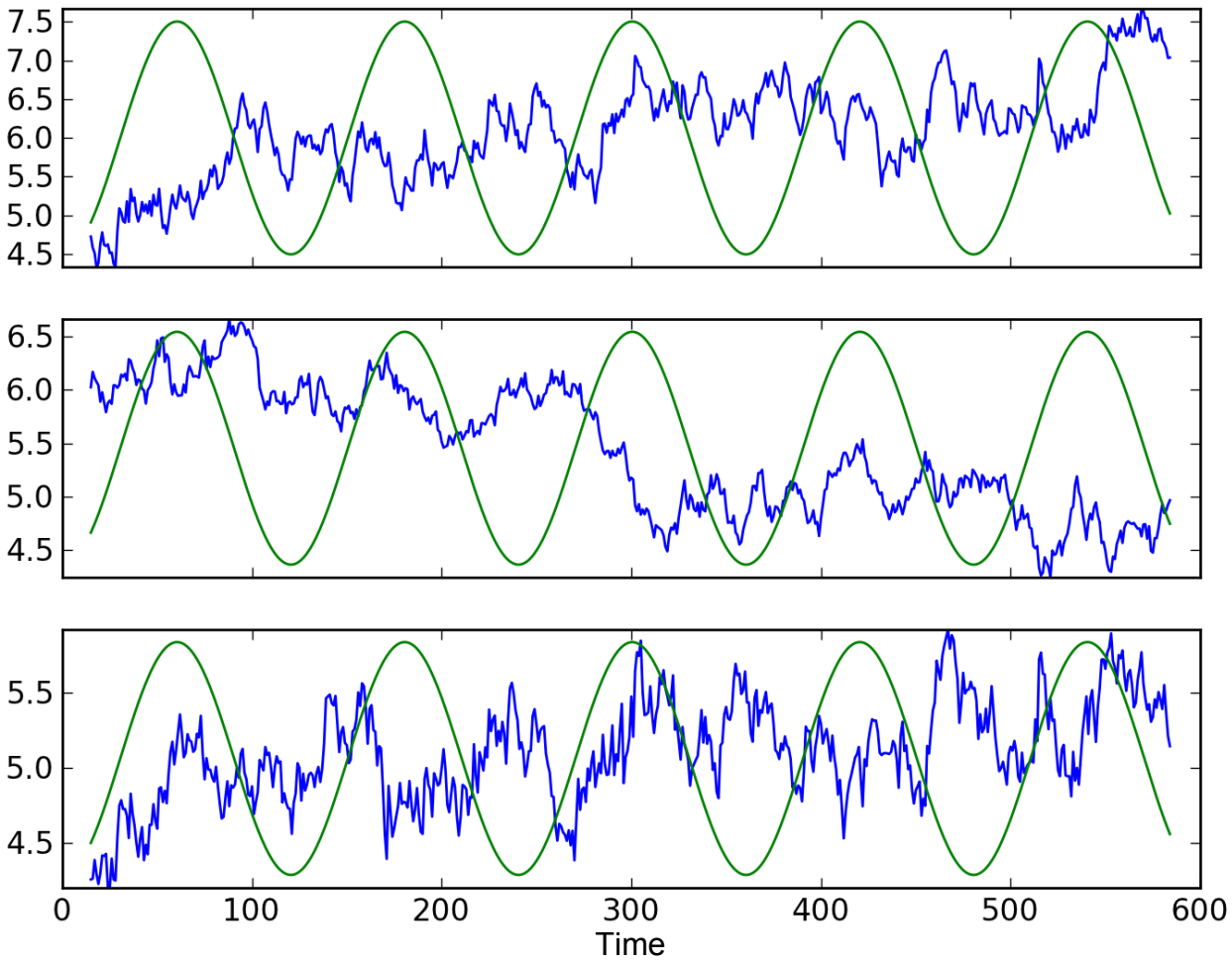
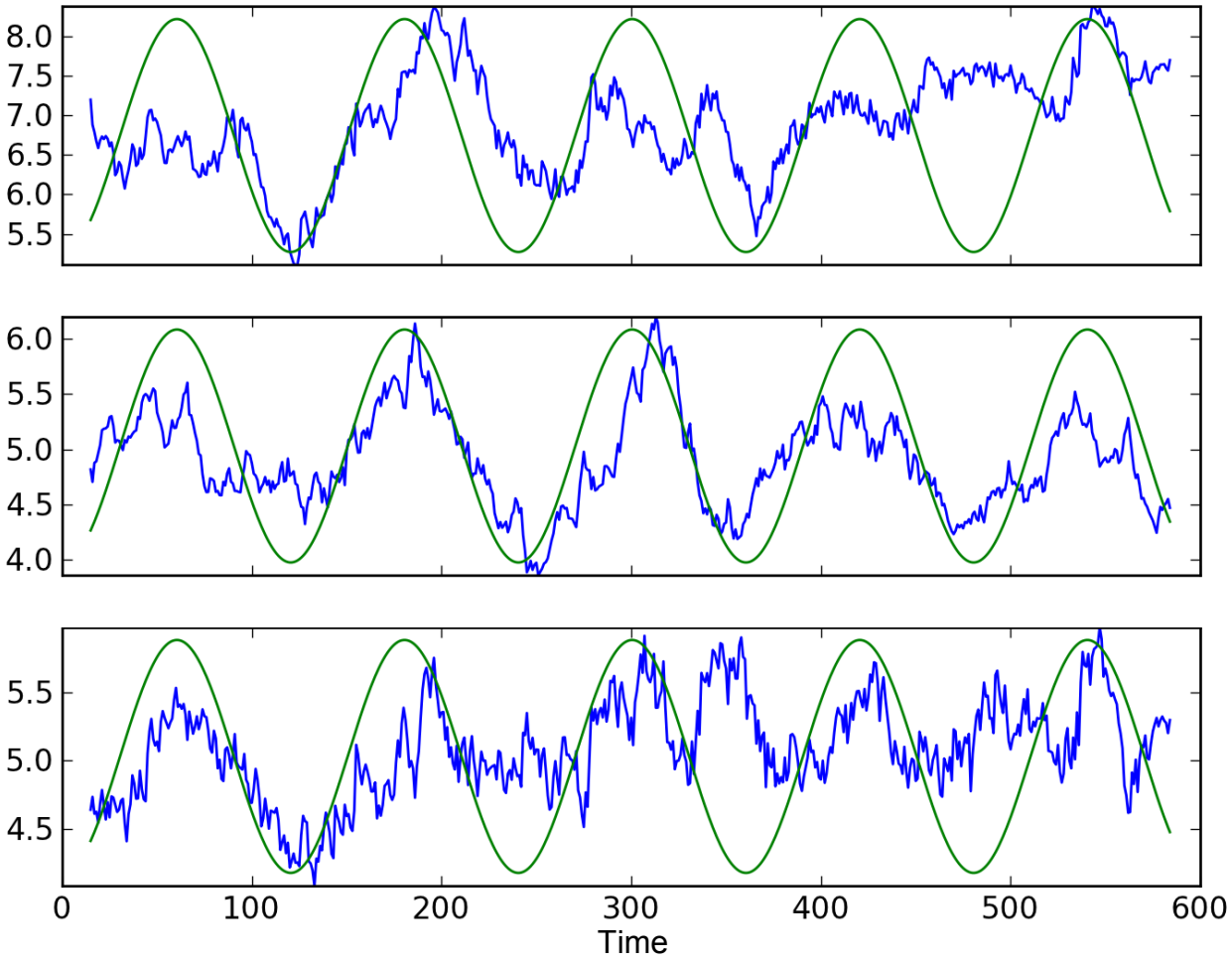


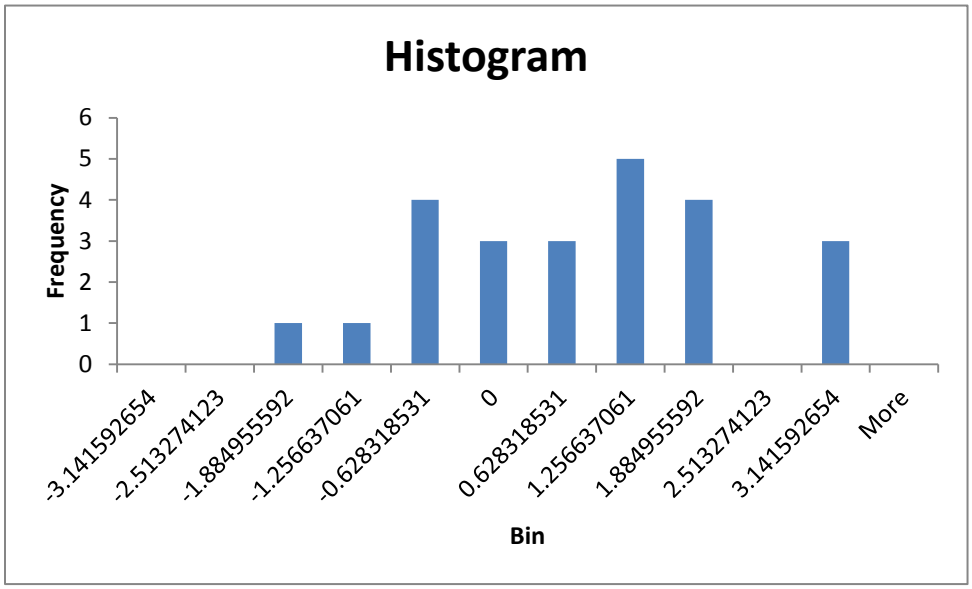
Figure 5.6.4 *RNG performance over time for Dual only. The panels depict RNG performance during the dual task trial. The top panel reflects cycling performance, the middle panel reflects seriation, and the bottom panel reflects repetition. The green trace depicts how difficulty changed during the CTT only and dual task only trials. Seriation was reliably influenced by the changing tracking difficulty*

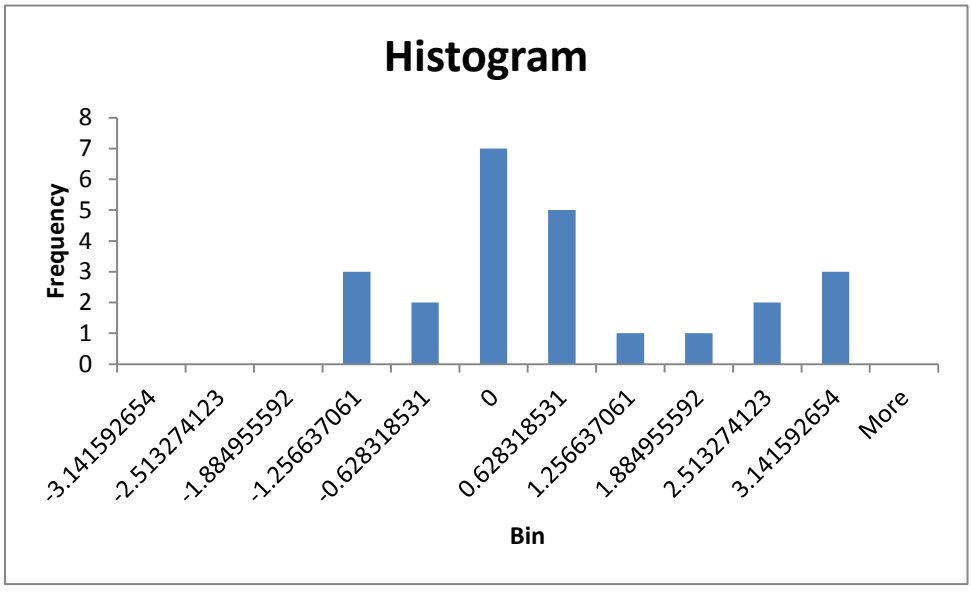


5.6.2.3 Effects of task on random number generation performance. To examine whether simultaneously performing the critical tracking task and random number generation task degraded the performance of random number generation the normalized power spectral density (NPSD) at 0.0083 Hz (period of 120 seconds) was obtained using discrete time Fourier transformation. If the factor scores are positively or negatively correlated with the difficulty signal power at this frequency should increase. Across participants Factor I (cycling) had a 20% increase in power at 0.0083 Hz but a one-tailed t-test failed to reject the null ($t(44) = 1.279$, $p = .104$). Factor II (seriation) had a 28% increase in power at 0.0083 and a one-tailed t-test found significance at .0498 ($t(44) = 1.682$). Factor III (repetition) scores increased by 4% and yielded a significance level of 0.386. When the Factor I and Factor II scores were combined using the Euclidean norm (treats components as orthogonal) a one-tailed t-test is significant at 0.0179 ($t(44) = 2.165$). The Euclidean norm treats the components as orthogonal and the result suggests that cycling and seriation are affected by the changing pendulum length. See Table 5.6.2.

5.6.2.4 Phase of random number generation factor scores relative to subjective difficulty. The NPSDs suggests that the changing difficulty influences random number generation but it does not tell us whether random number generation degradation is positively or negatively correlated to tracking difficulty. Examining the phase lag between the difficulty signal and the factor scores can provide this information. Because the random number generation reflects the randomness over a moving window of 30 s the factor scores should lag behind the difficulty signal by 15 s or pi/4 radians. These phase lags relative to the difficulty signal with a 30 second rectangular window applied are depicted in Figure 5.6.5. In this figure negative phases indicate lags, and positive phases indicate leads. For the most part the phase data suggests RNG degradations are positively correlated with tracking difficulty. The phase leads in the data for some participants are a bit hard to explain. The participants could be anticipating the difficult part after they realize they are passed the easiest part, then when the tracking because extremely difficult they might give up and wait for

Figure 5.6.5 *RNG factor phase analysis. Histograms of phases at 0.0083 Hz relative to the windowed difficulty signal for the Factor I (top) and Factor II (bottom) RNG measures for the 23 participants.*





it to become easy enough to control. The participants who are completely out of phase could indicate changes in strategy. For example, when the tracking becomes difficult a participant might reduce cycling but increase seriating, or vice versa.

5.6.3 ***Conclusions and Discussion.*** As expected, simultaneously performing the random number generation task and compensatory tracking task caused systematic deficits in tracking performance. Simultaneously performing both tasks also increased seriation and cycling behavior of the randomly generated digit sequences. Collectively, experiments 5.4, 5.5, and 5.6 provide empirical evidence that subjective workload can be systematically manipulated by varying pendulum length, and that the task difficulty manipulation affects central executive processes. And further, that gradual cyclical changes in task difficulty/workload occurring over 30 s moving window can be measured using a RNG secondary task. No previous studies have been found that continuous measured task difficulty or continuous monitored cognitive workload.

Table 5.6.1
Paired t-tests comparing single vs. dual task tracking performance at the five disturbance frequencies

	Task						*Cohen's*
Frequency	CTT	Dual	*t*	*p*	*df*	*r*	*d*
0.13 Hz	55.825 (4.284)	49.807 (5.563)	18.88	3.2e-14	20	.990	1.24
0.21 Hz	52.113 (3.866)	45.973 (4.674)	27.93	1.7e-17	20	.990	1.47
0.47 Hz	47.925 (3.328)	38.111 (4.871)	25.46	1e-16	20	.977	2.41
1.13 Hz	35.183 (2.997)	27.124 (5.769)	11.73	2e-10	20	.935	1.80
1.93 Hz	20.716 (3.990)	24.751 (4.361)	-10.78	8.8e-10	20	.919	-0.99

Dependent variable is the power spectral density of pendulum angle at the frequency of interest. Lower power suggests participants are better at compensating for the disturbance.

Table 5.6.2
Independent one-tail t-tests comparing RNG Factor score power at 0.0083 Hz

	Task					*Cohen's*
Frequency	RNG	Dual	*t*	*p*	*df*	*d*
Factor I	0.302 (0.179)	0.363 (0.142)	1.28	.104	44	0.39
Factor II	0.348 (0.141)	0.446 (0.239)	1.68	.050	44	0.51
Factor III	0.317 (0.157)	0.330 (0.150)	0.29	.386	44	0.09
L2(Factor I, Factor II)	0.485 (0.169)	0.604 (0.203)	2.17	.018	44	0.65

Appendix 5.5.A Consent Form

CONSENT FORM

Idaho Visual Performance Laboratory
Department of Psychology and Communication Studies
College of Liberal Arts and Social Sciences
University of Idaho
Control of speed during altitude changes

During this experiment you will be presented a display in a virtual environment. Various parameters of this display will be manipulated to examine stress and mental workload. In this experiment you will be asked to control movement in the virtual world using an input device such as a joystick.

The data you provide will be kept anonymous. There will be absolutely no link between your identity and your particular set of data.

Your participation will help increase knowledge of stress and mental workload. Subsequent to your participation the purpose and methods of the study will be described to you and questions about the study will be answered. It is our sincere hope that you will learn something interesting about your visual system from this debriefing.

The risks in this study are minimal, however displays simulating movement may on rare occasion cause motion sickness or eye fatigue in sensitive individuals. If at any time during the experiment you feel eye fatigue, dizziness, headache or nausea, please let the experimenter know immediately so that you can take a break before these symptoms become too intense. We endeavor to design our displays to minimize eye fatigue and motion sickness, and schedule periodic breaks to further reduce their occurrence. As a result, these phenomena have not been a common problem in previous similar studies.

Your participation will require **1** session of approximately **60** minutes. You may withdraw from this study at anytime without penalty. You will receive partial credit for your time spent. However, please be aware that your data is useful to us only if you complete the experiment in its entirety.
This research project has been approved by the University of Idaho Human Assurance Committee. As such, new information developed during the course of the research which may relate to your willingness to continue participation will be provided to you.

Thank you for your participation

Signature_____ Date _____

If you have further questions or encounter problems please contact:
Dr. Brian P. Dyre
(208) 885-6927
bdyre@uidaho.edu

Appendix 5.5.B Debriefing Form

Debriefing Form

Department of Psychology and Communication Studies

College of Letters, Arts, and Social Sciences

Physiological Workload Measures

Experiment 4b

Participant:____

Date:_____

1. How often do you play video games?
    a. What is your video game skill? (Bad, okay or good)
    b. Are you right or left handed?
2. Are you male or female?
3. Did you notice that some of the tracking trials were more difficult than others?
4. Did you use any particular strategy to generate random numbers?
5. Did you use any particular strategy to stabilize the dot?
6. When performing both tasks did you prioritize one more than the other?
7. Did you feel fatigued by the end of the experiment?
    a. If yes: Did you feel like fatigue influenced your performance?

8. Do you feel like performing both tasks was more difficult than performing the tasks independently?

9. Did you have any eye-strain, fatigue, blurred vision, problems focusing on the target, etc. ?

Any additional comments

This experiment examines how varying parameters of the internal model of the dynamic system influences how difficult it is to control. When the system is more difficult to control your verbal responses are hypothesized to become less random. These results are intended to help us manipulate task difficulty in future experiments.

Appendix 5.5.C Human Assurances Approval

University of Idaho

Office of Research Assurances
Institutional Review Board
PO Box 443010
Moscow ID 83844-3010

Phone: 208-885-6162
Fax: 208-885-5752
irb@uidaho.edu

To: Brian Dyre

From: Traci Craig, PhD
Chair, University of Idaho Institutional Review Board
University Research Office
Moscow, ID 83844-3010

IRB No.: IRB00000843

FWA: FWA00005639

Date: August 29, 2011

Title: 'Human Cognitive Workload and Perceptual Performance in Virtual Environments'

Project: 10-037
Approved: 09/28/11
Expires: 09/27/12

On behalf of the Institutional Review Board at the University of Idaho, I am pleased to inform you that the **first-year extension** of your proposal is approved as offering no significant risk to human subjects as no changes in protocol have been made on this project.

This extension of approval is valid until the date stated above at which time a second extension will need to be requested if you are still working on this project. If not, please advise the IRB committee when the project is completed.

Thank you for submitting your extension request.

Traci Craig

Traci Craig

5.7 Experiment 7: Compensatory Tracking with continuously Varied Difficulty and Physiological Data Collection

The previous experiment provided empirical evidence that when participants are engaged in a tracking task in which difficulty varies cyclically and continuously, cyclical degradations in tracking performance and secondary task performance occur. In Experiment 7 we will examine whether our suite of physiological measures and novel analytical techniques is also sensitive to cyclical and continuously-changing levels of task difficulty. Importantly, we will examine whether these measures show promise as leading indicators of mental workload that can help predict degradations in task performance. Using the same tracking task and manipulation of task difficulty as Experiment 6, this experiment also collected the physiological measures of pupil diameter, respiration, skin conductance, and heart rate variability. These measures were analyzed offline using a suite of novel learning algorithms to search for leading indicators of workload amount the physiological measures.

Several hypotheses were examined in this experiment. Previous experiments have used Daubechies four tap discrete mother wavelet kernel. Here we examine if using a continuous Morlet wavelet to obtain amplitude and phase information aids the resulting accuracy of models. Calculating amplitudes and phases from continuous transforms is mathematically simpler than with discrete transforms. With complex transforms one only has to take the norm to obtain the power. Discrete transforms require a moving quadratic mean calculation. Hence, it is hypothesized that precomputing amplitudes and phases from Morlet transforms will improve performance.

Secondly, a variety of non-linear machine learning algorithms are examined (GP, random forests, adaboost, SVM, and decision trees). Previously, I have shown that GP is superior LDA. LDA however lacks some of the sophistication of these other approaches. GP is expected to do well, but no other explicit hypotheses are made in this regard.

Thirdly, the role of multiple physiological measures will be examined. My previous experiments have only used pupil diameter and skin conductance. Here HR/HRV is also recorded.

Although, much of the information is redundant with PD and SC it is expected to improve the accuracy of the resulting models.

Fourthly, Models will attempt to use the physiological measures to classify the derivative task difficulty signal. Being able to identify derivative changes in cognitive workload is a necessary condition to developing predictive (leading) models of workload.

Lastly, task performance will be analyzed. Task performance is expected to increase monotonically with task difficulty (e.i. pendulum length). Task performance is also hypothesized to lag the task difficulty manipulation.

5.7.1 *Method*

5.7.1.1 Participants. Eight participants with normal or corrected to normal Snellen visual acuity of 20/30 participated in this study. All were naïve to the hypotheses of the experiment. From those eight participants only five had fully intact datasets (tracking, PD, SC, respiration, HR/HRV) due to the technical difficulties associated with the physiological measurement equipment. All participants were ethically treated in accordance with experimental protocols approved by the University of Idaho's Human Assurance Committee (see Appendices 5.7.A – 5.7.C).

5.7.1.2 Stimuli and Apparatus. As with the previous experiment participants performed a critical tracking task. Here the length of the pendulum was varied in a manner that should result in the subjective workload changing as a sinusoidal function with a period of 120 seconds.

5.7.1.3 Procedure. Because the previous experiment found significant individual differences in tracking ability participants completed three blocks of magnitude estimation trials at five discrete pendulum lengths ranging between 0.06 and 1.00m. The reported difficulty estimates were used to calibrate the second phase of the experiment. Table 5.7.2 provides a synopsis of the difficulty functions used across the eight participants. After participants completed the calibration trials they were given a short break of approximately five minutes before being setup with a ASL 5000 head mounted eye tracker, Pro Comp Infiniti skin conductance monitor, Datalab 2000

respiration band and 3 lead EKG electrodes. Participants then completed a ten minute trial of compensatory tracking. Difficulty was manipulated sinusoidally with a period of 120 seconds and an amplitude spanning a subjective range each participant would considered 1 t0 6.

5.7.2 *Results*

5.7.2.1 Discrete versus complex wavelet kernels. Previous wavelet approaches have used discrete kernels. These discrete kernels yield real-valued coefficients reflecting activity at orthogonal frequency bands. Complex wavelets yield complex-valued coefficients of activity. Assessing spectral power with complex wavelets is a matter of taking the L2-norm of the real and imaginary components. On the other hand, assessing the power of time-varying requires rectification and smoothing. For this reason it is hypothesized machine learning techniques may be aided by having the magnitude and phase angles from complex kernelled wavelets as opposed to discrete coefficients of discrete kernels. To test this hypothesis several supervised machine learning techniques were used to identify whether the sinusoidally varying workload signal was between 1 and 3.5 or between 3.5 and 6. In total five machine learning techniques were compared: decision tree, adaboost, support vector machine (SVM), random forest, and symbolic regression with ALPS.

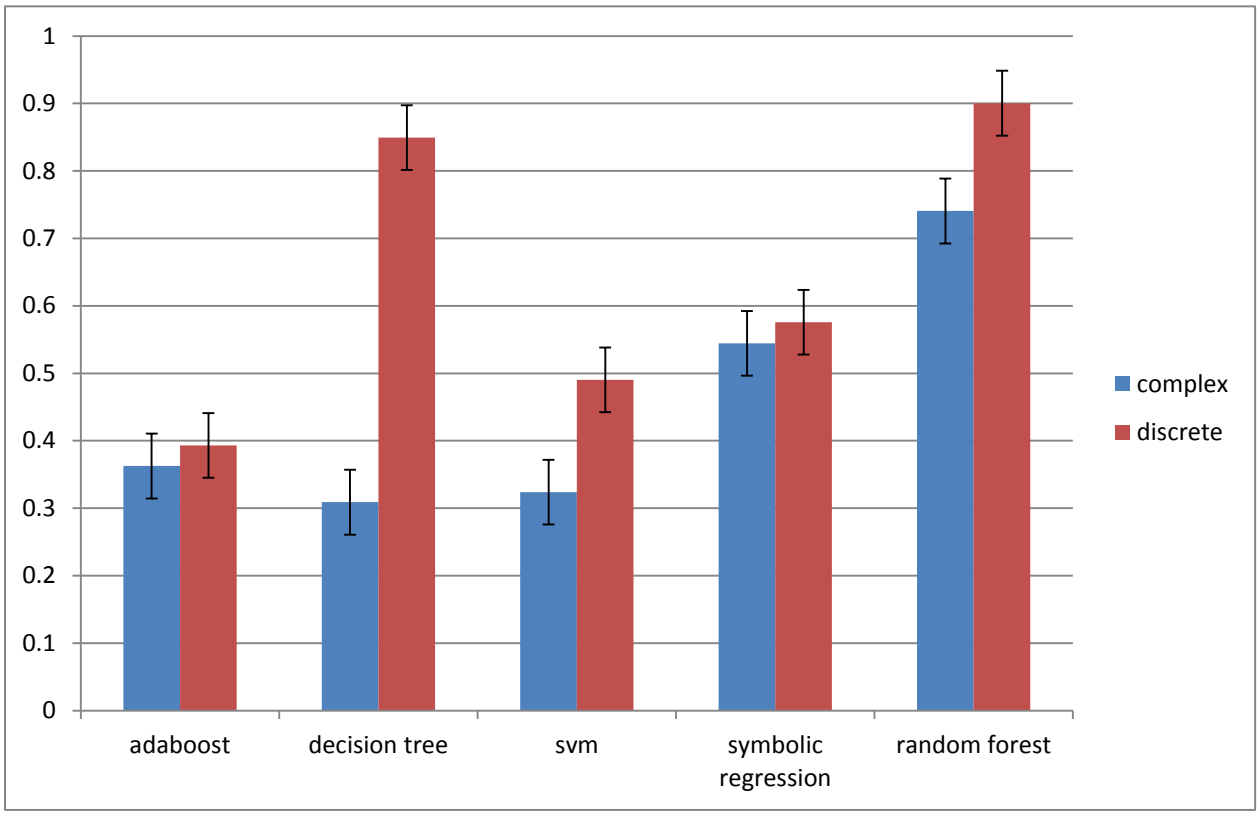
A 2 x 5 (wavelet, machine learning) repeated measures ANOVA (with Greenhouse-Geisser correction) on ten-fold cross validation accuracies disconfirmed our hypothesis that complex wavelets improve classification accuracy. A main effect of wavelet [$F(1, 4) = 41.051$, $p = .003$, $\eta^2_G = 2.458$, observed power = 1.000] indicates that in fact the converse, that discrete wavelets provide superior classification accuracy, is more likely. Even though calculating spectral power is simpler with complex wavelet, the complex wavelets have more spectral leakage that might hamper classification. The main effect of technique [$F(4, 16) = 74.279$, $p < .001$, $\varepsilon = 0.268$, $\eta^2_G = 6.988$, observed power = 0.999] and interaction between wavelet kernel and technique [$F(4, 16) = 41.758$, $p = .001$, $\varepsilon = 0.294$, $\eta^2_G = 2.458$, observed power = 0.833]were also found to be statistically reliable (see also Table 5.7.2). The two-way interaction is depicted in Figure 5.7.1 and Table 5.7.3. Best

performance was attained by random forests using discrete wavelet coefficients. Close behind was decision trees using discrete coefficients. A post-hoc Student-Newman-Keuls multiple comparisons analysis suggests this difference is not reliable. However the analysis did find that 35 of the 45 pairs are reliably different from one another. See Table 5.7.4

Some of the machine learning techniques cope with amplitude and phase estimates derived from complex kernels better than others. Most notably, decision tree performance decreased by over 60% (.850 to .309) when the complex kernel was used. Other techniques were less hindered by the type of wavelet used. Up to this point impressive symbolic regression had been at the forefront of this discussion. Here random forests outshine symbolic regressors. This is likely due to the fact that random forests are an ensemble learner. They have several models which more or less vote to form a final decision. This might also explain why performance is so bad for the decision trees using complex coefficients despite the fact that random forests are essentially a collection of decision trees. The distinction is that the trees that do not improve performance are given little weight to the final estimates.

5.7.2.2 Performance with multiple physiological measures. A second set of random tree classifiers were used to systematically examine whether including skin conductance wavelet coefficients and heart rate variability increased the accuracy of classifier predictions. For each participant heart rate variability was assessed by first identifying R peaks with the hybrid complex wavelet detect scheme described by Fard, Moradi, and Tajvidi (2008). Heart rate (HR, a.k.a. pulse rate) and heart rate variability (HRV) were then calculated using a moving window of 15 seconds. In total four classifiers were developed for each participant. Every classifier was given the discrete pupil diameter wavelet coefficients while the inclusion of skins conductance and (HR/HRV) was factorially manipulated. A 2 x 2 ANOVA was performed on the resulting cross-validation accuracies. Only a reliable main effect of skin conductance was found [$F(1,4) = 15.629$, $p = 0.017$, $\eta^2_G = 2.720$, observed power = 0.994].

Figure 5.7.1 *Wavelet Kernel x Machine Learning Technique Interaction on cross validation accuracy of binary estimates. Interaction suggests that some techniques are hindered more by the complex coefficients than by the discrete coefficients. The error bars represent 95% confidence intervals. Post-hoc analysis suggests that decision tree-discrete is not reliably different from random forest-discrete.*

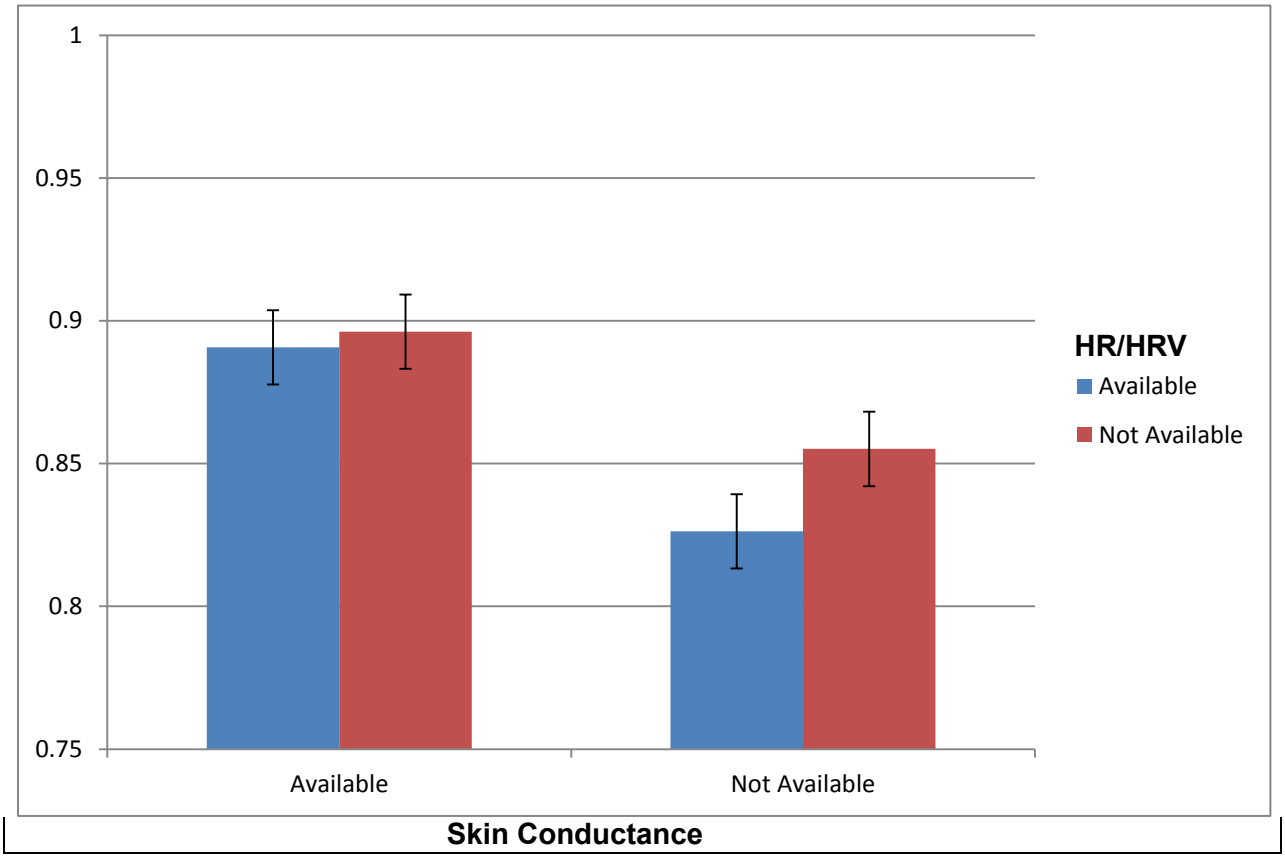


This main effect suggests that skin conductance does indeed improve classification accuracy (by 5% from 84.3 to 89.3%). The HR/HRV trend, while statistically inconclusive [$F(1,4) = 5.653$, $p = 0.076$, $\eta^2_G = 0.241$, observed power = 0.800], points towards HR/HRV actually having a detrimental effect on performance (see Table 5.7.5 for full ANOVA Summary). A Student-Newman-Keuls post-hoc analysis suggests that the worse performing classifier (PD with HR/HRV) is reliably different from the best classifier (PD with SC) and the second best classifier (PD with SC and HR/HRV; see Figure 5.7.2 and Table 5.7.6). No other differences were deemed reliable at a Type I error rate of 0.05.

The results presented above strongly suggest that skin conductance aids mental workload classifications, and that the effect of HR/HRV is inconclusive. In chapter 2 I discussed how some believe low frequency HRV (<0.2 Hz) reflects sympathetic activity and higher level activity reflects parasympathetic activity (Houle & Billman, 1999; Guger, et al. 2004; Wiederhold, Davis, & Wiederhold, 1998). To give HRV a second chance, low and high frequency activity was segregated with an 8th order Butterworth filter with a cutoff frequency of 0.2 Hz. This additional segregation did not make the results any more conclusive.

Theoretically, all of the measures used here should be affected by sympathetic nervous system activity. Thus the measures are conceptualized as containing redundant information. The fact that HR/HRV might actually hamper performance indicates that the measures contain a great deal of noise. To provide a better understanding of how pupil diameter, skin conductance, and heart rate variability are linearly related a correlational analysis was performed on the slowest two pupil diameter and skin conductance coefficients. A preliminary omnibus correlational analysis suggested that higher frequency correlations between different physiological measures were not meaningful. This also reduces the number of pairwise comparisons from 231 to 28. Due to the large number of cases (49920) resampling was used to randomly select 100 cases from each of the five participants.

Figure 5.7.2 *Cross Validation Accuracies based on availability of SC and HR/HRV.Making HR/HRV data available to machine learning does not improve the accuracy of the resulting models.*



The Larzelere and Mulaik step-down procedure was used to identify the reliability of the pairwise correlations. The analysis suggests that HRV is dominated by frequency content below 0.2 Hz (r = 0.859, p = 7e-147). Heart rate was found to correlate with the slowest GSR coefficient (r = .451, p = 1e-36). It also shows a reliable correlation of r = 0.451 (p = 2e-26) between the slowest SC coefficient and the slowest PD coefficient. Weak correlations (*r values* < 0.254) were also found between HRV and the slowest pupil diameter and skin conductance coefficients (see Table 5.7.7, and Figure 5.7.3).

5.7.2.3 Classification of Workload Derivative. Random Forests with PD and SC can classify the workload signal with close to 90% accuracy. If a person's mental workload and the rate at which it is changing are known then first order predictions of future workload can be obtained. To identify whether it is possible to classify the derivative of the workload signal another set of optimization runs were used with random forests. Results found that random forests were able to classify the derivative signal with at least the same cross-validation accuracy with 4 of the 5 participants (See Figure 5.7.4). This suggests that fairly accurate predictions of mental workload could be obtained from pupil diameter and skin conductance.

5.7.2.4 RMS Tracking Error Analysis. To understand how tracking error is related to the task difficulty signal magnitude estimation was used to predict tracking error from difficulty. This was accomplished by first calculating RMS tracking error using a 13 second moving window. Then the trial was segregated into 46 epochs of 13 seconds each (epoch needs to be aperiodic of the sinusoidal disturbance to increase the variability of the aggregated difficulty means). This analysis found reliable correlations (at α=0.05) for 6 of the 7 participants (See Table 5.7.8). For 6 of the 7 participants the regressions account for over 81% of the variability. The estimates suggest that tracking error is monotonically related to difficulty (see Figure 5.7.5). To be practical, a measure of workload should lead primary task performance. To examine whether the difficulty signal leads tracking error with this particular task discrete time Fourier transforms at 1/120 Hz were obtained

Figure 5.7.3 *Correlations between PD, SC, and HR/HRV. Scatterplots between physiological measures and coefficients. HRV and HRV were low-transformed to reduce positive skew.*

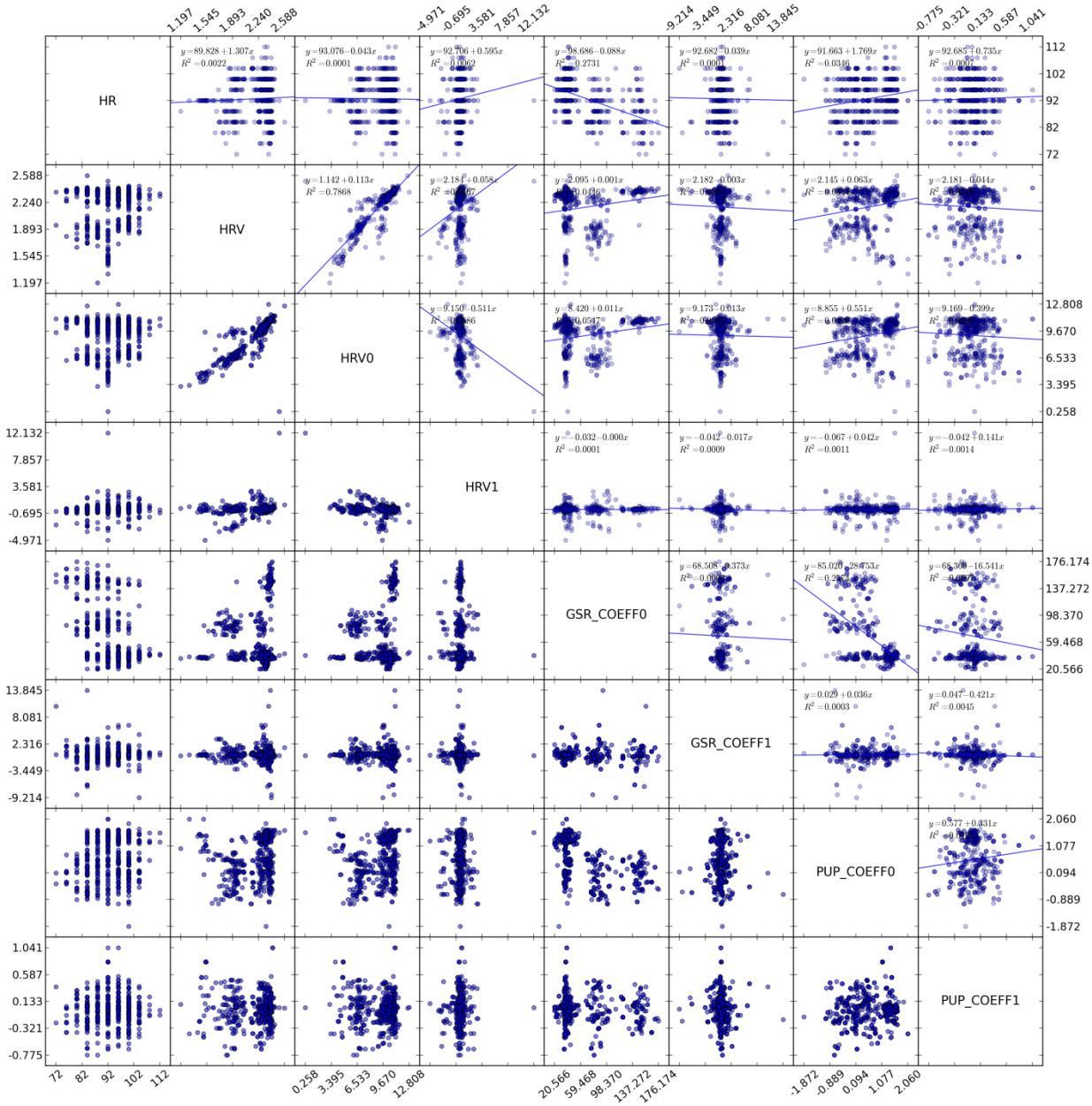


Figure 5.7.4 *Cross-validation accuracies of random forests by participant at predicting workload and the derivative of workload*

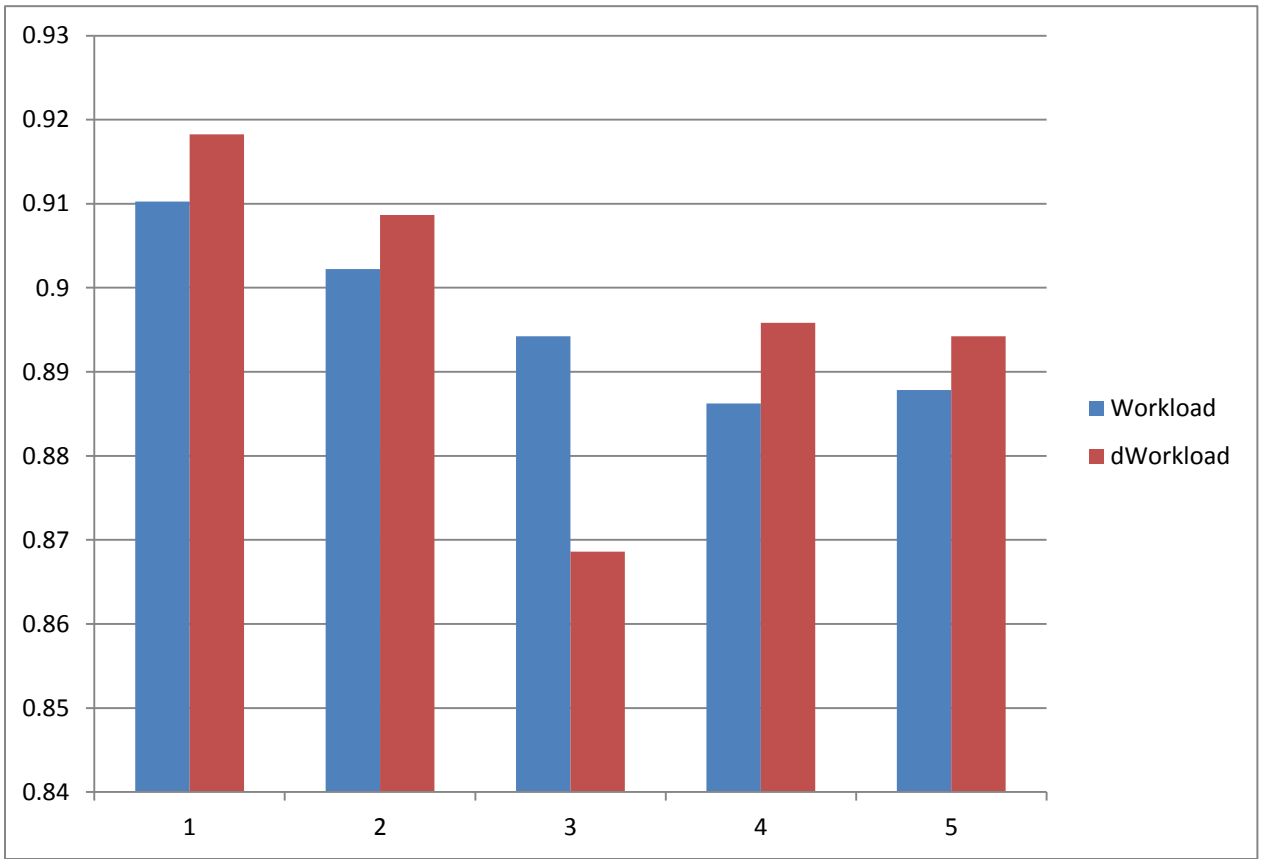
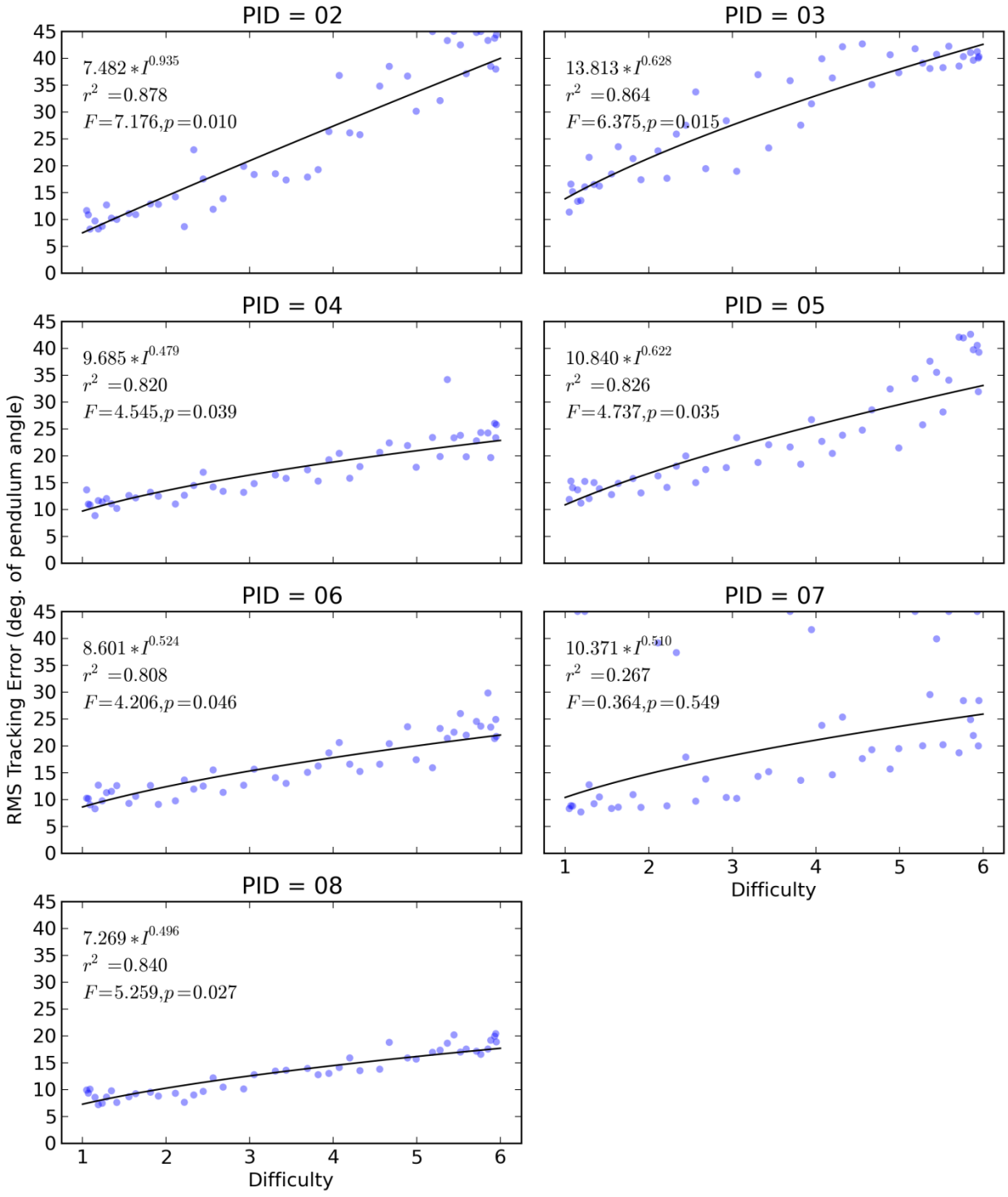


Figure 5.7.5 *RMS Tracking Error by Participant. Tracking error is monotonically related to difficulty.*



from the tracking error and difficulty signal. Phase lags and times could then be calculated between the two variables. The results of this analysis is in Table 5.7.9 and suggest that there is very little lag between task difficulty and primary task performance with this particular methodology.

5.7.3 *Conclusions and Discussion*

By continuously varying the length of the pendulum in the tracking task this experiment was able to manipulate task difficulty. These relatively subtle changes in workload were found to elicit changes in tracking performance and physiological measures. Physiological correlates of difficulty could be identified by using discrete wavelet transforms to decompose the physiological signals and applying machine learning.

Contrary to my hypothesis including additional physiological measures did not significantly improve workload classification by the learning algorithms. Much of the information in HR/HRV signals is redundant with PD and SC. Additional measures also complicate the data collection. Applied systems would need to validate the quality of physiological systems and switch classifiers based on the available metrics.

Most importantly, being able to quantify both workload magnitude and the derivative of workload suggests that predictive measures of workload can be obtained. These measures could potentially be used to augment human-machine interfaces to improve the overall safety and efficiency of human-machine systems.

Table 5.7.1
Magnitude estimation parameters derived from discrete tracking trials.

Participant	*a*	*k*	r^2	*F*	*p*
1	-0.341	3.029	0.387	0.632	0.441
2	-0.486	2.142	0.705	2.387	0.146
3	-0.363	2.400	0.652	1.874	0.181
4	-0.293	3.423	0.658	1.920	0.176
5	-0.690	1.669	0.831	4.904	0.035
6	-0.524	1.823	0.840	5.238	0.029
7	-0.483	1.874	0.499	0.995	0.326
8	-0.370	2.437	0.828	4.813	0.036

Table 5.7.2
Wavelet Kernel x Machine Learning Technique Summary ANOVA Table

Source	df	F	p	ε	η^2_G	observed power
Wavelet	1, 4	41.051	.003	-	2.428	1.000
Technique	4, 16	74.279	< .001	.268	2.428	0.999
Wavelet x Technique	4, 16	41.758	.001	.294	2.458	0.833

Table 5.7.3
Marginal Means for Wavelet Kernel by Machine Learning Technique

Technique	Wavelet	Mean	Std. Error	95% Confidence Interval
Adaboost	complex	0.363	0.027	0.309 - 0.416
Adaboost	discrete	0.393	0.031	0.333 - 0.453
Symbolic Regression	complex	0.544	0.025	0.495 - 0.594
Symbolic Regression	discrete	0.576	0.012	0.553 - 0.599
Decision tree	complex	0.309	0.025	0.260 - 0.358
Decision tree	discrete	0.849	0.005	0.839 - 0.860
Random Forest	complex	0.741	0.009	0.722 - 0.759
Random Forest	discrete	0.900	0.001	0.898 - 0.903
SVM	complex	0.324	0.015	0.294 - 0.353
SVM	discrete	0.490	0.074	0.346 - 0.635

Table 5.7.4
SNK: Step-down table of q-statistics

Pair	i	|diff|	q	range	df	p
decision tree_complex vs. random forest_discrete	1	0.591	19.843	36	14	< .001
random forest_discrete vs. svm_complex	2	0.577	19.346	35	14	< .001
decision tree_complex vs. decision tree_discrete	3	0.540	18.131	34	14	< .001
adaboost_complex vs. random forest_discrete	4	0.538	18.044	33	14	< .001
decision tree_discrete vs. svm_complex	5	0.526	17.635	32	14	< .001
adaboost_discrete vs. random forest_discrete	6	0.507	17.017	31	14	< .001
adaboost_complex vs. decision tree_discrete	7	0.487	16.333	30	14	< .001
adaboost_discrete vs. decision tree_discrete	8	0.456	15.306	29	14	< .001
decision tree_complex vs. random forest_complex	9	0.432	14.481	28	14	< .001
random forest_complex vs. svm_complex	10	0.417	13.984	27	14	< .001
random forest_discrete vs. svm_discrete	11	0.410	13.756	26	14	< .001
adaboost_complex vs. random forest_complex	12	0.378	12.683	25	14	< .001
decision tree_discrete vs. svm_discrete	13	0.359	12.045	24	14	< .001
alps_complex vs. random forest_discrete	14	0.356	11.944	23	14	< .001
adaboost_discrete vs. random forest_complex	15	0.347	11.656	22	14	< .001
alps_discrete vs. random forest_discrete	16	0.325	10.891	21	14	< .001
alps_complex vs. decision tree_discrete	17	0.305	10.233	20	14	< .001
alps_discrete vs. decision tree_discrete	18	0.274	9.180	19	14	0.001
alps_discrete vs. decision tree_complex	19	0.267	8.952	18	14	0.001
alps_discrete vs. svm_complex	20	0.252	8.455	17	14	0.002
random forest_complex vs. svm_discrete	21	0.250	8.395	16	14	0.002
alps_complex vs. decision tree_complex	22	0.235	7.898	15	14	0.004
alps_complex vs. svm_complex	23	0.221	7.402	14	14	0.006
adaboost_complex vs. alps_discrete	24	0.213	7.153	13	14	0.007
alps_complex vs. random forest_complex	25	0.196	6.583	12	14	0.013
adaboost_discrete vs. alps_discrete	26	0.183	6.127	11	14	0.019
adaboost_complex vs. alps_complex	27	0.182	6.100	10	14	0.017
decision tree_complex vs. svm_discrete	28	0.181	6.086	9	14	0.015
svm_complex vs. svm_discrete	29	0.167	5.590	8	14	0.023
alps_discrete vs. random forest_complex	30	0.165	5.529	7	14	0.020
random forest_complex vs. random forest_discrete	31	0.160	5.362	6	14	0.020
adaboost_discrete vs. alps_complex	32	0.151	5.073	5	14	0.021
adaboost_complex vs. svm_discrete	33	0.128	4.288	4	14	0.040
decision tree_discrete vs. random forest_complex	34	0.109	3.650	3	14	0.053
adaboost_discrete vs. svm_discrete	35	0.097	3.261	2	14	0.037
alps_discrete vs. svm_discrete	36	0.085	-	-	-	-
adaboost_discrete vs. decision tree_complex	37	0.084	-	-	-	-
adaboost_discrete vs. svm_complex	38	0.069	-	-	-	-
alps_complex vs. svm_discrete	39	0.054	-	-	-	-
adaboost_complex vs. decision tree_complex	40	0.054	-	-	-	-
decision tree_discrete vs. random forest_discrete	41	0.051	-	-	-	-
adaboost_complex vs. svm_complex	42	0.039	-	-	-	-
alps_complex vs. alps_discrete	43	0.031	-	-	-	-
adaboost_complex vs. adaboost_discrete	44	0.031	-	-	-	-
decision tree_complex vs. svm_complex	45	0.015	-	-	-	-

Table 5.7.5
Additional Physiological Measures Summary ANOVA Table

Source	df	F	p	η^2_G	observed power
SC	1, 4	15.629	.017	2.720	0.994
HR/HRV	1, 4	5.653	.076	0.241	0.800
SC x HR/HRV	1, 4	4.630	.098	0.098	0.450

Table 5.7.6
SNK: Step-down table of q-statistics

Pair	i	|diff|	q	range	df	p
PD and HR/HRV vs. PD and SC	1	0.066	8.559	3	17	.001
PD and HR/HRV vs. PD, HR/HRV and SC	2	0.060	7.849	2	17	.001
pd vs. PD and SC	3	0.041	-	-	-	-
pd vs. PD, HR/HRV and SC	4	0.036	-	-	-	-
pd vs. PD and HR/HRV	5	0.025	-	-	-	-
PD, HR/HRV and SC vs. PD and SC	6	0.005	-	-	-	-

Table 5.7.7
Larzelere and Mulaik Significance Testing

Pair	i	Correlation	p	alpha/(k-i+1)
HRV vs. HRV0	1	0.859	7e-147	0.002
GSR_COEFF0 vs. HR	2	0.525	1e-36	0.002
GSR_COEFF0 vs. PUP_COEFF0	3	0.451	2e-26	0.002
HRV0 vs. HRV1	4	0.316	5e-13	0.002
HRV0 vs. PUP_COEFF0	5	0.254	9e-09	0.002
HRV vs. PUP_COEFF0	6	0.240	5e-08	0.002
GSR_COEFF0 vs. HRV0	7	0.228	2e-07	0.002
GSR_COEFF0 vs. HRV	8	0.207	3e-06	0.002
HRV vs. HRV1	9	0.192	2e-05	0.003
HR vs. PUP_COEFF0	10	0.168	2e-04	0.003
PUP_COEFF0 vs. PUP_COEFF1	11	0.100	0.025	0.003
GSR_COEFF0 vs. PUP_COEFF1	12	0.090	0.045	0.003
HR vs. HRV1	13	0.059	0.185	0.003
HRV0 vs. PUP_COEFF1	14	0.054	0.230	0.003
HRV1 vs. PUP_COEFF1	15	0.052	0.242	0.004
HR vs. HRV	16	0.046	0.308	0.004
HR vs. PUP_COEFF1	17	0.044	0.328	0.004
GSR_COEFF1 vs. PUP_COEFF0	18	0.041	0.358	0.005
GSR_COEFF0 vs. GSR_COEFF1	19	0.036	0.425	0.005
GSR_COEFF1 vs. PUP_COEFF1	20	0.031	0.483	0.006
HRV1 vs. PUP_COEFF0	21	0.020	0.657	0.006
GSR_COEFF1 vs. HRV1	22	0.020	0.659	0.007
GSR_COEFF0 vs. HRV1	23	0.017	0.697	0.008
HRV vs. PUP_COEFF1	24	0.016	0.716	0.010
GSR_COEFF1 vs. HRV	25	0.013	0.772	0.013
GSR_COEFF1 vs. HR	26	0.012	0.789	0.017
HR vs. HRV0	27	0.007	0.867	0.025
GSR_COEFF1 vs. HRV0	28	0.002	0.956	0.050

Table 5.7.8
Tracking Error by Difficulty Magnitude Estimate Results by Participant

Participant	df	F	p	r^2	MSE
2	1, 45	7.176	.010	0.878	0.008
3	1, 45	6.375	.015	0.864	0.004
4	1, 45	4.545	.039	0.820	0.003
5	1, 45	4.737	.035	0.826	0.005
6	1, 45	4.206	.046	0.808	0.004
7	1, 45	0.364	.549	0.267	0.048
8	1, 45	5.259	.027	0.840	0.003

Table 5.7.9
Primary Task Performance Lag by Participant

Participant	gain (dB)*	lag (s)
2	19.21	1.677
3	16.96	4.258
4	10.98	0.675
5	16.12	0.156
6	11.29	0.669
7	23.01	11.836
8	8.92	0.731

* Gain calculated as $20 * \log 10 \left(\frac{RMS\ Tracking\ Power}{Difficulty\ Power}\right)$

ɬ Participant 7's RMS tracking error did not reliable correlate with difficulty. Their calculated lag may be spurious.

Appendix 5.7.A Consent Form

CONSENT FORM

Idaho Visual Performance Laboratory
Department of Psychology and Communication Studies
College of Liberal Arts and Social Sciences
University of Idaho
Control of speed during altitude changes

During this experiment you will be presented a display in a virtual environment. Various parameters of this display will be manipulated to examine stress and mental workload. In this experiment you will be asked to control movement in the virtual world using an input device such as a joystick.

The data you provide will be kept anonymous. There will be absolutely no link between your identity and your particular set of data.

Your participation will help increase knowledge of stress and mental workload. Subsequent to your participation the purpose and methods of the study will be described to you and questions about the study will be answered. It is our sincere hope that you will learn something interesting about your visual system from this debriefing.

The risks in this study are minimal, however displays simulating movement may on rare occasion cause motion sickness or eye fatigue in sensitive individuals. If at any time during the experiment you feel eye fatigue, dizziness, headache or nausea, please let the experimenter know immediately so that you can take a break before these symptoms become too intense. We endeavor to design our displays to minimize eye fatigue and motion sickness, and schedule periodic breaks to further reduce their occurrence. As a result, these phenomena have not been a common problem in previous similar studies.

Your participation will require **1** session of approximately **30** minutes. You may withdraw from this study at anytime without penalty. You will receive partial credit for your time spent. However, please be aware that your data is useful to us only if you complete the experiment in its entirety.
This research project has been approved by the University of Idaho Human Assurance Committee. As such, new information developed during the course of the research which may relate to your willingness to continue participation will be provided to you.

Thank you for your participation

Signature_____ Date _____

If you have further questions or encounter problems please contact:
Dr. Brian P. Dyre
(208) 885-6927
bdyre@uidaho.edu

Appendix 5.7.B Debriefing Form

Debriefing Form

Department of Psychology and Communication Studies

College of Letters, Arts, and Social Sciences

INL Physiological Predictors of Workload

Experiment 7

Participant:____

Date:_____

1. Did you move your left hand during the course of the trial while the GSR was still hooked up?

2. How often do you play video games?

    a. What is your video game skill? (Bad, okay or good)

    b. Are you right or left handed?

3. Do you identify yourself as male or female?

4. During the second part did you notice that difficulty changed throughout the trial?

5. How uncomfortable was the eye-tracker when you first put it on? (1-10)

6. How uncomfortable was the eye-tracker when you finished? (1-10)

7. Did you find the eye-tracker distracting from the task at hand?

8. Do you think that fatigue played a role in your performance?

a. How about fatigue from the eye-tracker?

9. Did you have any eye-strain, fatigue, blurred vision, problems focusing on the target, etc. ?

Any additional comments

Appendix 5.7.C Human Assurances Approval

University of Idaho

Office of Research Assurances
Institutional Review Board
PO Box 443010
Moscow ID 83844-3010

Phone: 208-885-6162
Fax: 208-885-5752
irb@uidaho.edu

To: Brian Dyre

From: Traci Craig, PhD
Chair, University of Idaho Institutional Review Board
University Research Office
Moscow, ID 83844-3010

IRB No.: IRB00000843

FWA: FWA00005639

Date: August 29, 2011

Title: 'Human Cognitive Workload and Perceptual Performance in Virtual Environments'

Project: 10-037
Approved: 09/28/11
Expires: 09/27/12

On behalf of the Institutional Review Board at the University of Idaho, I am pleased to inform you that the **first-year extension** of your proposal is approved as offering no significant risk to human subjects as no changes in protocol have been made on this project.

This extension of approval is valid until the date stated above at which time a second extension will need to be requested if you are still working on this project. If not, please advise the IRB committee when the project is completed.

Thank you for submitting your extension request.

Traci Craig

Traci Craig

Chapter 6: Conclusions

The subject central to this dissertation pertains to measuring cognitive workload measures from physiological indicators. In time critical operations, such as process control, operators may not have time to report their subjective workload, or may not be explicitly aware of their mental states (see Chapter 2). Nuclear power plant operators often joke that operating a nuclear power plant is 99% sheer boredom interspersed with 1% sheer terror. When humans are under stress the autonomic nervous system initiates a fight or flight response. During the 99% of sheer boredom mild or even moderate stressors may increase arousal and aid decision making, but during the 1% of sheer terror too much stress may produce cognitive interference that may prevent operators from grasping the larger picture and selecting the appropriate actions even in situations where operators may be highly trained and normally competent (see Chapter 2). Despite these limitations, human operators play an essential and critical role to the overall reliability and functioning of plant operations.

Being able to identify sensitive and diagnostic measures of mental workload is a complex challenge. Physiological measures are influenced by autonomic and environmental factors as well as cognitive workload. Physiological measures can also be influenced by a person's overall health. The information we are actually interested in ends up buried and noise and obtaining usable measures requires a good deal of filtering and processing. In the beginning, the approaches used to decipher workload were rather naïve. The first experiment examined how pupil diameter and skin conductance magnitudes were affected by a series of faulted components while participants engaged in a simplistic process control task (DURESS). This analysis was refined by applying short time Fourier transforms to the physiological signals and applying linear discriminate analysis. While the DURESS task offered external validity, the faulted components often went unnoticed. Unnoticed faults are unlikely to elicit changes in cognitive workload. This was remedied by using a pursuit control task. Rotating the control mappings was found to produce an immediate and salient

increase in tracking difficulty across participants. Evolutionary algorithms were able to use wavelet decomposed pupil diameter and skin conductance to classify the mapping state better than linear discriminate analysis.

To systematically manipulate task difficulty a compensatory tracking task was developed. Experiment 5.4 validated that discrete changes in task difficulty caused systematic and reliable changes in subjective difficulty. Experiment 5.5 required participants to simultaneously performing a verbal random number generation task and compensatory. Previous research suggests that random number generation requires central executive processes to inhibit automated serialized responses (see Section 5.6); thus secondary task performance can be used as a measure of cognitive workload. As the compensatory became more unstable and consequently more difficult to track the generated sequence of digits became less random. Participants showed increased cycling and seriating behavior. This suggests that the compensatory tracking task loads central executive processes (critical for planning and decision making). Experiment 5.6 established that dual task performance decrements generalized to continuous changes in task difficulty.

In the final experiment, evidence was found that both the magnitude and derivative of cognitive workload could be estimated with > 90% accuracy. This suggests that leading indicators of workload could be formed from relatively simply, potentially unobtrusive physiological measures. This holds significant practical relevance, as being able to identify cognitive workload when an operator is at a high risk of making a poor decision could offer magnitude orders of safety in operations were the human component is the primary source of risk.

6.1 Empirical Contributions and Limitations

Previous studies have demonstrated that secondary tasks can be used to measure residual cognitive ressources (Kahneman, 1973; Navon & Gopher, 1979; Moray, 1988). Dual task experiments using random number generation have shown that random number generation interfers with central executive processing (Zelanznik, Spencer, & Ivry, 2002; Noordzij, van der

Lubbe, Neggers, & Postma, 2004; Koike, Marumo, Kinou, Kawakubo, Rogers, & Kasai, 2011; Spatt, 1996). No studies to date have examined the RNG in conjunction with critical tracking task. Here we demonstrated RNG randomness degraded as the difficult of the CTT was increased. This work was also able to use a moving window to produce a continuous measure of RNG randomness. This continuous measure was sensitive enough to detect slowing wavering changes in task difficulty.

While the corpus of literature for subjective and task performance based measures of workload is large, the number of studies examining physiological based measures of workload is small. This work not only expands this body of knowledge, but substantiates it by linking it to both subjective ratings, and secondary task performance.

Here physiological measures of workload were calculated offline. A logical next step would be to predict cognitive workload in real-time and use this measure to dynamically maximize primary task performance. The critical tracking task could be devised with a scoring system that rewarded small deviations and shorter pendulum lengths. An adaptive system would need to constantly monitor workload to keep the pendulum length at a point where the user can maintain control and maximize the rate the scoring accumulation rate.

A second shortcoming is that primary performance does not significantly lag task difficulty with the CTT. This makes it difficult to assess whether physiological measures of workload can predict task performance. This could be due to the relatively quick time dynamics of the tracking task; with the CTT the system can become unstable in a fraction of a second and be can be stabilized in a matter of seconds. With process control tasks small deviations may take several minutes or even hours to accumulate. Returning to process control may yield leading indicators if the system dynamics are slower to evolve and a means of subtly manipulating task difficulty can be developed.

On the empirical front, future work should focus on the diagnosticity of these measures. These measures may be influenced by not only cognitive workload but also arousal. These constructs are likely interdependent. The Yerkes-Dobson law describes how moderate levels of

arousal are necessary for optimal levels of performance (Yerkes & Dodson, 1908). When arousal is too high or too low cognitive resources cannot be efficiently allocated and performance suffers (Berkun, 1964; Capretta & Berkun, 1962; Lundberg, 1993; Porcelli & Delgado, 2009) . While high levels of arousal may led to high workload, high arousal is not necessarily bad, and misclassifying arousal may lead to unintended consequences. To use a sports analogy we can think about the "in the zone" phenomena. While in the zone athletes are performing extremely well even though there arousal may be high. They are have the cognitive resources to keep up with task demands so their workload is at a nominal level. Mistaking arousal for high workload would be analogous to benching Michael Jordan while he is on a hot streak. From this illustration it should be clear why we should try to segregate arousal from workload. Because, the machine learning algorithms essentially fit the data to pre-specified examples it is possible that machine learning could be used to predict several psychological constructs.

6.2 Analytical Contributions and Limitations

The success here suggests that the application of machine learning is underutilized by cognitive science and perhaps essential to cutting edge science. It also demonstrates that less expensive/obtrusive physiological measures can be used to form measures of workload.

While previous work has examined how pupil diameter is linked to cognitive workload (Nakayama & Katsukura, 2007; Nakayama & Shimizu, 2002) and has applied wavelet decomposition to pupil diameter (Marshall, 2000; 2002; 2007) the application of genetic programming and other machine learning techniques is novel. This work has found that machine learning is an effective tool to develop robust classifiers of cognitive workload. In particular, Brieman's (2001) random forests and symbolic regression were particularly effective.

The use of multiple measures is also novel. Current research tends to focus on a particular type of physiological measure (e.i. EGG, transcranial doppler, fNIR). Experiment 5.7 demonstrated that multiple measures can improve cross validation accuracies, but measures that are highly

redundant may actually hamper the resulting models.

Random forests classifiers were found to best performance of the machine learning techniques tested. This may be attributed to random forests being ensemble learners. They have a "team" of models which "vote" to classify the input data. Other types of machine learning algorithms can be implemented as ensemble learners and may offer performance on par or superior to random forests. Readers should keep in mind that machine learning in this context should be considered a reverse engineering tool that allow us to decipher what information is carried by physiological signals and how it is represented. In the long-term, *specific* descriptions and models may be superior to these *general* problem solvers.

Despite evidence that the critical tracking task loads central executive resources, these results may lack external validity to real world applications such as process control. In such settings operators may have to multitask while integrating information across multiple modalities while ignoring visual and auditory noise (visual clutter, alarms, communication amongst other operators, machines, etc.). The results here should be seen as establishing some of the first principles that may led towards integrated solutions.

References

Amit, Y., & Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Computation, 9*(7), 1545-1588.

Backs, R. W. (1995). Going beyond heart rate: Autonomic space and cardiovascular assessment of mental workload. *The International Journal of Aviation Psychology, 5*(1), 25-48.

Baddeley, A. D. (1996). Exploring the central executive. *The quarterly Journal of Experimental Psychology, 49*(1), 5-28.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. Bower, *The psychology of learning and motivation* (pp. 47 – 89). New York: Academic Press.

Baddely, A. D., & Wilson, B. (1988). Frontal amnesia and the dysexecutive syndrome. *Brian and Cognition, 7*, 212-230.

Barr, L., Howarth, H., Popkin, S., & Carroll, R. J. (2005). *Proceedings from the International Conference on Fatigue Management in Transportation.* Seattle, WA.

Barr, L., Popkin, S., & Howarth, H. (2009). *An evaluation of emerging driver fatigue detection measures and technologies.* U. S. Department of Transportation Federal Motor Carrier Satey Administration.

Beech, J. R. (1984). The effects of visual and spatial interference on spatial working memory. *Journal of General Psychology, 110*(2), 141-149.

Berkun, M. M. (1964). Performance decrement under psychological stress. *Human Factors, 6*(1), 21-30.

Besner, D. (1987). Phonology, lexical access in reading, and articulartory suppression: A critical review. *Quarterly Journal of Experimental Psychology, 39A*, 467-478.

Billman, G. E. (2011). Heart rate variability – A historical perspective. *Frontiers in Physiology, 2*(86), 1-13.

Birbaumer, N. (2006). Breaking the silence: Brain–computer interfaces (BCI) for communication and motor control. *43*, 517-532.

Blackledge, J. M. (2003). *Digital Signal Processing: Mathematical and Computational Methods, Software Development and Applications.* Woodhead Publishing.

Boring, R. L., & Kelly, D. (2008). *P-203: Human Reliability Analysis (HRA) Training Course.* US Nuclear Regulatory Commission.

Brieman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32.

Broadbent, D. (1958). *Perception and Communication.* London: Pergamon Press.

Cantor, G. (1887). Über die verschiedenen Ansichten in Bezug auf die actualunendlichen Zahlen." Bihang till Kongl. Svenska Vetenskaps-Akademien Handlingar. *Bd.*

Capretta, P. J., & Berkun, M. M. (1962). Validity and reliability of certain measures of psychological stress. *Psychological Reports, 10*, 875-878.

Casey, S. M. (1998). *Set Phasers on Stun: And other true tales of design, technology, and human error.* Aegean Publishing.

Chomsky, N. (1959). Review of Skinner's Verbal Behavior. *Language, 35*, 26–58.

Cosentino, B., & Ross, A. (1999). *Design and Implementation of a Java Version of the DURESS II Simulator.* Toronto, Canada: University of Toronto, Cognitive Engineering Laboratory.

Cosmides, L., & Tooby, J. (1997). The multimodular nature of human intelligence. In A. Schiebel, & W. Schopf (Eds.), *Origin and evolution of intelligence* (pp. 71-101). Los Angeles, CA: Center for the Study of the Evolution and Origin of Life, UCLA.

Crawford, M. (2009). *Shop class as soulcraft: An inquiry into the value of work.* Penguin Press HC.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inf. Theory, 36*(5), 961-1005.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics.

Daubechies, I., Grossmann, A., & Y., M. (1986). Painless non-orthogonal expansions. *J. Math. Phys., 27*, 1271-1283.

Davies, D. R., & Parasuraman, R. (1982). *The Psychology of Vigilance.* London: Academic Press.

Degani, A. (2004). *Taming HAL: Designing interfaces beyond 2001.* New York, NY: St. Martin's Press/Palgrave-Macmillan.

Dekker, S. W. (2006). *The field guide to understanding human error.* Aldershot, UK: Ashgate Publishing Co.

Drilling, N. C. (2011). *Deep Water: The Gulf Oil Disaster and the Future of Offshore Drilling.*

Duffin, R. J., & Schaefer, A. C. (1952). A class of nonharmonic Fourier series. *72*, 341-366.

Dykes, A. R. (1946). Chairman's Address to the Scottish Branch of the Institution of Structural Engineers.

Dyre, B. P., Grimes, J., & Lew, R. (2009). *ViEWER: Virtual Environment Workbench for Education and Research.* Moscow, ID: University of Idaho, Department of Psychology and Communication Studies.

Eiben, A. E., & Smith, J. E. (2003). *Introduction to Evolutionary Computing.* Springer-Verlag.

Ellingwood, B. R., & Wen, Y. K. (2005). Risk benefit-based design decisions for low probabbility/high-consequence earthquate events in Mid-Ameria. *Progress in Structural Engineering and Materials, 7*, 56-70.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge Handbook of Expertise and Expert Performance.* New York, NY: Cambridge University Press.

Evans, F. J. (1978). Monitoring attention deployment by random number generation: An index to measure subjective randomness. *Bulletin of the Psychonomic Society, 12*, 35-38.

Fard, Moradi, & Tajvidi. (2008). A novel approach in R peak detection using Hybrid Complex Wavelet (HCW). *International Journal of Cardiology, 124*, 250-253.

Figner, B., & Murphy, R. O. (2010). Using skin conductance in judgment and decision making research. In M. Schulte-Mecklenbeck, A. Kuehberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research.* New York, NY: Psychology Press.

Folland, G., & Sitaram, A. (1997). The Uncertainty Principle: A mathematical survey. *Journal of Fourier Analysis and Applications, 3*(3), 207-238.

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. *International Journal of Psychophysiology, 31*, 129-145.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation, 121*(2), 256-285.

Freund, Y., & Schapire, R. E. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence, 14*(5), 771-780.

Galton. (1907). Vox Populi. *Nature*.

Gavon, P. (2010, 2 23). Sleep Cycle: iPhone app that helps tired parents. *WIRED.co.uk*.

Gazzaniga, M., Ivry, R., & Mangun, G. (2002). *Cognitive Neuroscience : The Biology of the Mind* (2nd ed.). W.W. Norton.

Geisser, S., & Greenhouse, S. (1958). An extention of Box's result on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics, 29*, 885-891.

Gertman, D. I., & Blackman, H. S. (1994). *Human Reliability & Safety Analysis Data Handbook.* Wiley-Interscience.

Ginsburg, N., & Karpiuk, P. (1994). Random generation: Analysis of the responses. *Perceptual and Motor Skills, 79*, 1059-1067.

Ginsburg, N., & Wiegersma, S. (1991). Response bias and the generation of random sequences. *Perceptual and Motor Skills*, 1332-1334.

Good, I. J. (1958). The interaction algorithm and practical Fourier analysis. *Journal of the Royal Statistical Society, Series B20*(2), 361-372.

Gopher, D., & Donchin, E. (1986). Workload – An examination of the concept. In L. Kaufman, & T. J. P., *Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance* (pp. 41-49). John Wiley and Sons, Inc.

Greenhouse, S., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Greenwood, R. E. (1955). Coupon collector's test for random digits. *Mathematical Tables and Computation*, 1-4.

Grimnes, S. (1982). Psychogalvanic reflex and changes in electrical parameters of dry skin. *Medical and Biological Engineering and Computing, 20*, 734-740.

Guger, C. E., Leeb, R., Pfurtcheller, G., Antley, A., Garau, M., & al., e. (2004). Heart-rate variability and event-related ECG in virtual envrinoments. *Presence 2004: The 7th Annual International Workshop on Presence*, (pp. 240-245).

Harding, G., & Punzo, F. (1971). Response uncertainty and skin conductance. *Journal of Experimental Psychology, 68*(2), 265-272.

Harris, F. J. (1978). On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proceedings of the IEEE*, *66*, pp. 51-83.

Harrison, D., Boyce, S., Loughnan, P., Dargaville, P., Storm, H., & Johnston, L. (2006). Skin conductance as a measure of pain and stress in hospitalized infants. *Early Human Development, 82*, 603-608.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In H. P., & M. N., *Human mental workload* (pp. 139-183). Amsterdam: North Holland.

Haynal, S., & Haynal, H. (2011). Generating and searching families of FFT algorithms. *Journal on Satisfiability, Boolean Modeling and Computation, 145-187*, 145-187.

Ho, T. K. (1995). Random Decision Forest. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, (pp. 278-282). Montreal, QC.

Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844.

Hogwei, W. (2009). Evaluation of Various Window Function using Multi-Instrument.

Hornak, J. P. (2010). *The Basics of MRI*. Retrieved January 24, 2010, from Interactive Learning Software: http://www.cis.rit.edu/htbooks/mri/index.html

Hornby, G. S. (2009). A steady-state version of the age-layered population structure EA. In R. Riolo, U.-M. O'Reilly, & T. McConaghy (Eds.), *Genetic Programming Theory and Practice VII.* New York, NY: Springer-Verlag.

Houle, M. S., & Billman, G. E. (1999). Low-frequency component of the heart rate variability spectrum: a poor marker of sympathetic activity. *Am. J. Physiol. Heart Circ. Physiol., 267*, H215–H223.

Huang, R.-S., T.-P., J., & Makeig, S. (2007). Event-related brain dynamics in continuous sustained attention. In D. D. Schmorrow (Ed.), *Foundations of Augmented Cognition: Third international conference* (pp. 65-73). Springer Publishing.

Hughes, J. (2004). *Citizen Cyborg: Why Democratic Societies Must Respond to the Redesigned Human of the Future.* Westview Press.

Jacobs, S. C., Friedman, R., Parker, J. D., Tofler, G. H., Jimenez, A. H., Muller, J. E., et al. (2001). Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research. *Am Heart J., 128*(6), 1170-1177.

Jahanshahi, M., Profice, P., Brown, R. G., Ridding, M. C., Dirnberger, G., & Rothwell, J. C. (1998). The effects of transcranial magnetic stimulation over the dorsolateral prefrontal cortex on suppression of habitual counting during random number generation. *Brain, 121*, 1533-1544.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology, 47*, 310-339.

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. *Theorertical Issues in Ergonomic Science, 4*(1-2), 56-88.

Just, M. A., Keller, T. A., & Cynkar, J. A. (2008). A decrease in brain activation associated with driving when listening to someone speak. *Brain Research, 1205*, 70-80.

Kahneman, D. (1973). *Attention and Effort.* Englewood Cliffs, NJ: Prentice Hall.

Kantowitz, H. B., & Knight, J. L. (1976). Testing tapping timesharing: I. Auditory secondary task. *Acta Psychologica, 40*, 343-362.

Keizer, M. (2004). Scaled Symbolic Regression. *Genetic Programming and Evolvable Machines, 5*, 259-269.

Kendall, M. G., & Smith, B. B. (1938). Randomness and random sampling numbers. *Journal of the Royal Statistical Society, 101*, 147-166.

Kitcher, P. S. (2003). *Science, Truth, and Democracy.* USA: Oxford University Press.

Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C. E. (1993). *Decision Making in Action: Models and Methods.* Norwood, NJ: Ablex Publishing Company.

Knuth, D. (1981). *The Art of Computer Programming* (Vol. 2). Reading, MA: Addison-Wesley.

Koike, S. T., Marumo, K., Kinou, M., Kawakubo, Y., Rogers, M. A., & Kasai, K. (2011). Association between severe dorsolateral prefrontal dysfunction during random number generation and earlier onset in schizophrenia. *Clinical Neurophysiology, 122*, 1533-1540.

Komiyama, H., & Kraines, S. (2008). *Vision 2050: Roadmap for a Sustainable Earth.* Springer: Japan.

Kotchetkov, I. S., Hwang, B. Y., Appelboom, G., Kellner, C. P., & Connolly, E. S. (2010). Brain-computer interfaces:military, neurosurgical, and ethical perspective. *Neurosurgical Focus, 28*(5), 1-6.

Kugelmass, S., Lieblich, I., Ben-ishai, A., Opatowski, A., & Kaplan, M. (1968). Experimental evaluation of galvanic skin response and blood pressure change indices during criminal interrogation. *Criminology and Police Science, 59*(4), 632-636.

Lew, R., Dyre, B. P., Soule, T., Ragsdale, S. A., & Werner, S. (2010). Assessing mental workload from skin conductance and pupillometry using wavelets and genetic programming. *In Proceedings of the 54th Annual Meeting of the Human Factors and Ergonomics Society.*

Lew, R., Dyre, B. P., Werner, S., Wotring, B., & Tran, T. (2008). Exploring the potential of short-time Fourier transforms for analyzing skin conductance and pupillometry in real-time applications. *In Proceedings of the 52th Annual Meeting of the Human Factors and Ergonomics Society*, (pp. 1536-1540).

Liaw, A. (2013). *Package 'randomForest'.* CRAN.

Linderman, M. D., Santhanam, G., Kemere, C., Gilja, V., O'Driscoll, S., Yu, B., et al. (2008). Signal processing challenges for neural prostheses. *IEEE Signal Processing Magazine, special issue on brain-computer interfaces, 25*, 18-28.

Logie, R. H. (1995). *Visual-spatial working memory.* Hove,UK: Lawrence Erlbaum Associates, Inc.

Lotto, F., Congedo, M., Lecuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering, 4*(R1).

Lundberg, U. (1993). On the psychobiology of stress and health. In O. Svenson, & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 41-53). New York, NY: Plenum Press.

Lyons, R. (2008). *Quadrature signals: Complex but not complicated*. Retrieved from DSP Guru: http://www.dspguru.com/sites/dspguru/files/QuadSignals.pdf

Marshall, S. (2000). *Patent No. 6,090,051.*

Marshall, S. (2002). The index of cognitive activity: Measuring cognitive workload. *IEEE 7th Human Factors Meeting.* IEEE.

Marshall, S. (2007). Identifying cognitive state from eye metrics. *Aviat Space Environ Med, 78*(5), B165-175.

McCarley, J. S., Vais, M. J., Pringle, H., Kamer, A. F., Irwin, D. E., & Strayer, D. L. (2004). Conversation distupts change detection in complex traffic scenes. *Human Factors, 46*, 424-436.

McDonnell, J., & Jex, J. (1967). *A "critical" tracking task for man-machine research related to the operator's effective elay time. Part II: Experimental effects os a system input spectra, control of stick stiffness, and controlled element order, CR-674.* NASA.

McLarty, C. (1997). Poincaré: Mathematics & logic & intuition. *Philosophica Mathematica, 5*(2), 97-115.

McRuer, D., & Graham, D. (1965). *Human pilot dynamics in compensatory systems, AFFDL-TR-65-15.* USAF.

Milner, D. G. (2006). *The Visual Brain in Action* (2nd ed.). USA: Oxford University Pres.

Miot, S. (2012, 8 27). Smartphone Adoption Rate Fastest in Tech History. *PCMag*.

Moray, N. (1988). Mental workload since 1979. In D. J. Oborne (Ed.), *International Reviews of Ergonomics: Current trends in human factors research and practice.*

Moslehian, M. S., Rowland, T., & Weisstein, E. W. (2011). *Banach Space*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/BanachSpace.html

Motavalli, J. (2010, February 4). The dozens of computers that make modern cars go (and stop). *The New York Times*.

Munson, A. (2010). Bourbaki at Seventy-Five: Its Influence in France and Beyond. *Journal of Mathematics Education at Teachers College, 1*, 18-21.

Murai, K., Hayashi, Y., Nagata, N., & Inokuchi, S. (2003). Mental workload of ship's navigator: A few comments on heart rate variability during navigational watch keeping. *7th International Conference, KES 2003.* Oxford UK.

Murata, A., & Iwase, H. (2000). Evaluation of mental workload by variability of pupil area. *IEICE Trans. Inf. & Syst., E83-D*(5), 1187-1190.

Nagai, Y., Critchley, H. D., Featherstone, E., Trimble, M. R., & Dolan, R. J. (2004). Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: a physiological account of a "default mode" of brain function. *NeuroImage, 22*(1), 243-251.

Nakayama, M., & Katsukura, M. (2007). Feasibility of assessing usability with pupillary responses. *Proc. of AUIC 2007*, (pp. 15-22).

Nakayama, M., & Shimizu, Y. (2002). An estimation model of pupil size for 'Blink Artifact' and it's applications. *Proceedings of 10th European Symposium on Artificial Neural Networks*, (pp. 251-256).

Nakayama, M., & Shimizu, Y. (2004). Frequency analysis of task evoked pupillary response and eye-movement. *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, (pp. 71-76).

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review, 86*, 214–255.

Newsome, M. (2007, 1 24). He Took On the Whole Power-Tool Industry. *Inc.*

NHTSA. (1997). *Drowsy Driving and Automobile Crashes.*

Noordzij, M. L., van der Lubbe, R. H., Neggers, S. F., & Postma, A. (2004). Spatial Tapping Interferes With the Processing of Linguistic Spatial Relations. *Canadian Journal of Experimental Psychology, 58*(4), 259-271.

Norman, D. A. (2011). *Living with Complexity.* Cambridge, MA: MIT Press.

Norman, D. A., & Shallice, T. (1980). *Attention to action. Willed and automatic control of behavior.* University of California San Diego CHIP Report 99.

O'Neill, K. (2006). Modular software for an augmented cognition system. In D. Schmorrow (Ed.), *Foundations of Augmented Cognition* (pp. 309-330). Mahwah, NJ; London, UK: Lawrence Erlbaum Associates Publishers.

Orme-Johnson, D. W. (1979). Autonomic stability and transcendental meditation. *Psychosomatic Medicine, 35*, 341-349.

Parasuraman, R., & Rizzo, M. (Eds.). (2008). *Neuroergonomics: The Brain at Work.* Oxford University Press.

Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., & Barnes, M. (1992). *Theory and design of adaptive automation in aviation systems (Technical Report No. NAWCADWAR-92033-60).* Naval Air Warfare Center, Aircraft Division.

Pavel, M., Wang, G., & Li, K. (2002). Augmented cognition: Allocation of attention. *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS '03).*

Peplow, M. (2011). Chernobyl's Legacy. *Nature*, 562-565.

Petzoldt, T., Bär, N., & Krems, J. F. (2009). Gender effects on lane change test (LCT) performance. *PROCEEDINGS of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design.*

Pogue, D. (2011, 11 30). A Thermostat That's Clever, Not Clunky. *The New York Times*.

Poincaré, H. (1913/2012). *Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method.* (J. M. Cattell, Ed., & G. B. Halsted, Trans.) New York and Garrison, NY: The Science Press.

Porcelli, A. J., & Delgado, M. R. (2009). Acute stress modulates risk taking in financial decision making. *Psychological Science, 20*(3), 278-283.

Quintopia. (2007). *Favorite math jokes*. Retrieved from echochamber: http://forums.xkcd.com/viewtopic.php?f=17&start=120&t=5683

Rabinowitz, F. M. (1970). Characteristic sequential dependencies in multiple-choice situations. *Psychological Bulletin, 74*, 141-148.

Rayleigh, J. W. (1889). On the Character of Complete Radiation at a Given Temperature. *Philos. Mag. 27. Reprinted in Scientific Papers*.

Robbins, T. W., Anderson, E. J., Barker, D. R., Bradley, A. C., C., F., Henson, R., et al. (1996). Working memory in chess. *Memory & Cognition, 24*(1), 83-93.

Rowland, T. (2011). *Hermitian Inner Product*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/HermitianInnerProduct.html

Rowland, T. (2011). *Lp-space*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/Lp-Space.html

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Sabatinelli, D., Lang, P. J., Bradley, M. M., Costa, V. D., & Keil, A. (2009). The timing of emotional discrimination in human amygdala and ventral visual cortex. *Journal of Neuroscience, 29*(47), 14864-14868.

Sajda, P., Muller, K.-R., & Shenoy, K. V. (2008). Brain-computer interfaces. *Signal Processing Magazine, IEEE, 25*(1), 16-17.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal, 3*.

Santhanam, G., Ryu, S. I., Yu, B., Afshar, A., & Shenoy, K. (2006). A high performance brian-computer interface. *Nature, 442*, 195-198.

Scerbo, M. W. (1996). Theroretical perspectives on adaptive automation. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmorrow, D. D., & McBride, D. (2004). Introduction. *International Journal of Human-Computer Interaction, 17*(2), 127-130.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transaction of the Royal Society London, B, 29*(8), 199-209.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge, UK: Cambridge University Press.

Shannon, C. E. (1949). Communication in the presence of noise. *Proc. IEEE, 86*(2).

Sharikadze, M., Cong, D. K., Staude, G., Deubel, H., & W., W. (2009). Dual-tasking: Is manual tapping independent of concurrently executed saccades? *Brain Research, 1283*, 41-49.

Shaw, T. H., Guagliardo, L., de Visser, E., & Parasuraman, R. (2010). Using transcranial Doppler sonography to measure cognitive load in a command and control task. *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, 249-253.

Simoni, D. A. (2008). Reverse engineering the visual system via genetic programs. In L. M. Reeves, & D. D. Schmorrow (Eds.), *Foundations of Augmented Cognition.* Springer.

Smith, J. (2007). *Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications* (2nd ed.).

Spatt, J. (1996). Evaluation of a simulation of human performance on random-digit generation: Measures of concept and redundancy. *Perceptual and Motor Skills, 83*, 319-322.

St. John, M., Kobus, D., Morrison, J., & Schmorrow, D. D. (2004). *Overview of the DARPA Augmented Cognition Technical Integration Experiement.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, S. S. (1953). On the brightness of lights and the loudness of sounds. *Science, 118*, 576.

Stevens, S. S. (1970). Neural events and the psychophysical law. *Science, 170*, 1043–1050.

Storm, H., Myre, K., Rostrup, M., Stokland, O., Lien, M. D., & Ræder, J. C. (2002). Skin conductance correlates with preoperative stress. *Acta Anaesthesiologica Scandinavica, 46*, 887-895.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular phone. *Psychological Science, 12*, 462-466.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied, 9*, 23.

Sussman, H. M. (1982). Contrastive patterns of intrahemispheric interference to verbal and spatial concurrent tasks in right-handed, left-handed and stuttering populations. *Neuropsychologia, 20*(6), 675-684.

Svenson, O., & Maule, J. (Eds.). (1993). *Time pressure and stress in human judgment and decision making.* New York, NY: Plenum.

Terlecki, M., Brown, J., Harner-Steciw, L., Irvin-Hannum, J., Marchetto-Ryan, N., Ruhl, L., et al. (2010). Sex differences and similarities in video game experience, preferences, and self-efficacy: Implications for the gaming industry. *Current Psychology, 3*(1), 22-33.

Tesler, L., & Saffer., D. (2007). Larry Tesler interview: The laws of interaction design. In D. Saffer, *In Designing for Interaction: Creating Smart Applications and Clever Devices.* Berkeley, California: New Riders. Published in association with AIGA Design Press.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw, & J. Bransford (Eds.), *In Perceiving, acting and knowing* (pp. 211-265). Hillsdale, N.J.: Erlbaum Associates.

Vincente, K. J., & Pawlak, W. S. (1994). *Cognitive work analysis for the DURESS II system, (CEL 94–03).* Toronto, Canada: University of Toronto, Cognitive Engineering Laboratory.

Weis, M., Romer, F., Haardt, M., Jannek, D., & Husar, P. (2009). Multi-dimensional space-time-frequency component analysis of event related EEG data using closed-form PARAFAC. *Acoustics, Speech and Signal Processing*, 19-24.

Weisstein, E. W. (2011). *Autocorrelation*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/Autocorrelation.html

Weisstein, E. W. (2011). *Fourier Series*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/FourierSeries.html

Weisstein, E. W. (2011). *Function Space*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/FunctionSpace.html

Weisstein, E. W. (2011). *Hilbert Space*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/HilbertSpace.html

Weisstein, E. W. (2011). *Inner Product Space*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/InnerProductSpace.html

Weisstein, E. W. (2011). *Metric Space*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/MetricSpace.html

Weisstein, E. W. (2011). *Normed Space*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/NormedSpace.html

Weisstein, E. W. (2011). *Parseval's Theorem*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/ParsevalsTheorem.html

Weisstein, E. W. (2011). *Plancherel's Theorem*. Retrieved from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/PlancherelsTheorem.html

Wells, R. (2007). *Introduction to Biological Signal Processing and Computational Neuroscience.* Moscow, ID.

Wells, R. (2012). Personal Communication. (R. Lew, Interviewer)

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Wiederhold, B. K., Davis, R., & Wiederhold, M. D. (1998). Correlation of physiological arousal and immersion levels in virtual worlds. In G. &. Riva (Ed.), *Virtual Environments in Clinical Psychology and Neuroscience: Methods and Techniques in Advanced Patient-Therapist Interaction* (pp. 52-60). IOS Press.

Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of finger-tapping task variations: An ALE meta-analysis. *NeuroImage, 42*, 343-356.

Wittgenstein, L. (1953/2001). *Philosophical Investigations.* Blackwell Publishing.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*, 459-482.

Young, P. M., Clegg, B. A., & Smith, C. A. (2004). Dynamic models of augmented cognition. *International Journal Of Human-Computer Interation, 17*(2), 259-273.

Zelanznik, H. N., Spencer, R. M., & Ivry, R. B. (2002). Dissociation of explicit and implicit timing in repetitive tapping and drawing movements. *Journal of Experiment Psychology: Human Perception & Performance, 28*, 575-588.

Zelanznik, H. N., Spencer, R. M., & Ivry, R. B. (n.d.). Dissociation of explicit and implicit timing in repetitive tapping and drawing movements. *Journal of Experiment Psychology: Human Perception & Performance, 28*, 575-588.

Zhang, Y., van Drongelen, W., Kohrman, M., & He, B. (2008). Three-dimensional brain current source reconstruction from intra-cranial ECoG recordings. *Neuroimage, 42*, 683-695.