Cognition and Psychological Scaling:

Model, Method, and Application

# of Constrained Scaling[*]

by

Ronald Laurids Boring, B.A. (Hons.), M.A.

A thesis submitted to

the Faculty of Graduate Studies and Research

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Institute of Cognitive Science

Carleton University

Ottawa, Ontario, Canada

September 24, 2004

[*]Carleton University Cognitive Science Technical Report 2004-10
http://www.carleton.ca/ics/techReports.html

The undersigned recommend to the

Faculty of Graduate Studies and Research

acceptance of the thesis

“Cognition and Psychological Scaling:

Model, Method, and Application of Constrained Scaling”

submitted by

Ronald Laurids Boring, B.A. (Hons.), M.A.,

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Chair of Examining Board

Department Chair, Institute of Cognitive Science

Thesis Supervisor

External Examiner

Carleton University

September 7, 2004

To Wendy and Elliott—what matters over mind.

In Memoriam

Professor Richard Dillon

Professor Lew Stelmach

ABSTRACT

This dissertation builds on previous research on constrained scaling, a technique for training individuals to translate mental magnitudes to numeric scales. Constrained scaling has been found to reduce significantly the variability in scale use both within and between individuals. A series of 15 brief experiments related to constrained scaling is presented. Specific findings include: (1) loudness constrained scaling experiments can be implemented on a conventional personal computer without the need for specialized hardware; (2) loudness scaling experiments can be successfully conducted without the need for a sound attenuating chamber; (3) cross-modal constrained scaling exhibits scale carryover from the training to the testing stimuli; (4) cross-modal constrained scaling is also susceptible to stimulus range effects; (5) brightness stimuli should be flashed in order to minimize the possibility of participant light adaptation; (6) conventional computer monitors are effective for displaying color brightness scaling stimuli; (7) constrained scaling of color brightness results in significantly reduced variability compared to magnitude estimation; (8) interval stimuli are more effective than ordinal stimuli for scale training; (9) random noise should be added to feedback values when using ordinal training stimuli; (10) the optimal ratio of training to testing trials is 1:1; (11) the ratio of training to testing scaling exponents is constant across scaling modalities; (12) there is considerable individual difference in scaling the subjective utility of money; (13) constrained scaling increases sensitivity to individual differences in scaling; (14) constrained scaling is more sensitive than magnitude estimation for rating the subjective visual appeal of Web pages; (15) constrained scaling can be applied successfully to aid software users in making parameter selections for streaming media. These 15 experiments demonstrate that constrained scaling is easy to implement without costly or specialized psychophysical laboratory equipment. Further, the experiments highlight the current breadth of constrained scaling research, including traditional psychophysical domains such as loudness and brightness scaling to novel psychometric domains such as rating the subjective utility of money or the visual appeal of Web pages. Finally, they show that constrained scaling offers unmatched reliability in introspective elicitation, making it a powerful tool for cognitive research.

CLASSIFIER KEYWORDS

magnitude estimation; constrained scaling; mental magnitude; measurement theory; psychological calibration; methodological triangulation; variance; reliability; cognition; psychophysics; psychometrics; cross-modal matching; ordinal scale; interval scale; loudness; brightness; color; happiness; visual appeal; quality of service

ACKNOWLEDGEMENTS

Foremost, I wish to thank my family for their help in seeing this dissertation to completion. I cannot express strongly enough my gratitude for the crucial role that my wife, Wendy, played. Her unwavering support as well as her persistent push made this document a reality. In the early stages of writing this dissertation, we had our son, Elliott, who reminded me that there were more exciting things in life than working on a dissertation. I thank you, Elliott, for giving me such a good reason to finish this document sooner rather than later. I also wish to thank my parents, Don and Grit Boring, and my sister, Mona Mondava, for their constant encouragement.

I owe special gratitude to my dissertation committee. Professor Robert West served as my thesis supervisor and academic advisor at Carleton University. He sold me on the topic of constrained scaling by allowing me to explore its methods and concepts at my own pace and in my own manner. Much of what is contained in this dissertation is the product of the many fruitful methodological and conceptual discussions I had with Professor West. I also wish to thank Professors Andrew Brook and John Logan, my other dissertation committee members. Both served as instructors and mentors to me throughout my PhD, and I was fortunate to continue working with them on my dissertation. Professor Brook was extremely helpful in refining the philosophical discussions contained in this dissertation; Professor Logan proved invaluable in upholding a sound methodological foundation as well as high writing standards throughout this dissertation. I thank my entire committee for their patience and diligence in working with me.

I also wish to thank my external examiners, Professor Robert Langlois from Carleton University's Department of Mechanical and Aerospace Engineering, and Professor Willy Wong from the University of Toronto's Department of Electrical and Computer Engineering. Both provided thought-provoking questions during my dissertation defense and provided useful direction for my final revisions to this document.

Several students at Carleton University aided the completion of this dissertation. Stephen Moore and Kayoko Tanabe served as research assistants, and both helped immensely in the data collection effort. As well, my writing benefited from countless insightful and productive discussions with Gary Fernandes of the Psychology Department and with Sanjay Chandrasekharan, Terry Stewart, and Matthew Rutledge-Taylor of the Institute of Cognitive Science.

Finally, I acknowledge the sources of financial support that allowed me to complete this research. The Faculty of Graduate Studies and Research at Carleton University provided a variety of stipends during my doctoral studies. Parts of this research were funded by grants to Professor Richard Dillon through the Council of Information Technology Ontario (CITO) as well as by grants to Professor Robert West through the Natural Sciences and Engineering Research Council (NSERC) and the Canadian Foundation for Innovation (CFI). The Idaho National Engineering and Environmental Laboratory (INEEL) assisted with tuition at the conclusion of this dissertation. I thank Drs. Harold Blackman, Walt Newcomb, Jerry Harbour, Julie Marble, and David Gertman of the INEEL for their roles in securing this final source of funding.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION[1]

The Problem Space

Meister (2001, 2003) points out that many psychological researchers take measurement for granted. While the statistical tools to evaluate psychological measures are carefully considered, the actual nature and use of these measures are not. In the following dissertation, I explore this shortcoming in terms of scaling, one of the most frequently employed forms of human measurement. Without a proper understanding of scaling, there is the risk that scaling becomes an unreliable measure at best. With a proper understanding of scaling, it is possible to increase the reliability of scaling in human psychological research.

Human psychological scaling involves the process of translating subjective experience into quantitative values. Because it is difficult to match overt behaviors or physiological responses to subjective experience, scaling serves as a primary method available for researchers to assess mental states. Although scaling is an introspective technique, the findings from scaling research are empirically robust. In psychophysical research, for example, the relationship between the intensity of a physical stimulus and an individual's perception of magnitude follows a lawlike relationship (Stevens, 1975). In psychometric research, the Likert scale and its many derivatives provide exceptional insight into human mental states (Dumas, 2001), even when those mental states are not precipitated by readily identifiable cues in the environment. Psychometric scaling

[1] Portions of this introduction first appeared in West, Boring, and Moore (2002) and Boring (2003).

techniques allow researchers to document dynamic cognitive states as well as enduring personality traits and attitudes.

While psychophysical and psychometric scaling techniques are well-established and powerful research methods, they are not without shortcomings. Foremost among these shortcomings is the problem of variability between individuals (Algom & Marks, 1990). With classical scaling methods, there is no guarantee that two users with the same subjective experience will use a scale in the same way. The failure of different individuals to scale in the same way results in considerable score variability. This score variability decreases the effectiveness of scaling in research and necessitates the collection of scaling data from many individuals in order to arrive at conclusive research findings.

A problem arises when two users have identical perceptions, but they scale their perceptions differently. As Poulton (1989) notes, people exercise different biases when scaling. One user might be biased to assign scale values close to the midpoint of the scale, regardless of his or her subjective experience. Another user might be biased to use extreme low and high scale values, avoiding midpoint values on the scale even when they represent the most appropriate match to his or her subjective experience.

The scaling biases that Poulton (1989) documents can lead to a high level of interparticipant scaling variability. When two or more scalers have identical underlying perceptions but use a scale differently, there is a divergent range of scale values assigned for each level of mental magnitude. In such a case, the mapping between the scale and the underlying mental phenomenon is unreliable across scalers.

An additional form of scaling variability occurs within individual scalers. Intraparticipant scaling variability results when an individual scaler fails to use a scale consistently. The scaler may, for example, switch scaling strategies during a scaling session, causing an inconsistent mapping between the scale and the individual's sense of mental magnitude.

The Solution Set

Any measurement apparatus that lacks consistency and calibration is ineffective. The same thing can be said about humans as scalers. Humans are generally not in the regular practice of quantifying subjective experiences. When asked to do so, we often do so in an inconsistent manner and without calibration to the scale we are asked to use. We simply lack sufficient exposure to our target scale to be effective scalers.

The goal of any study of scaling is to determine methodologies that allow people to communicate accurately the magnitudes of specific dimensions of conscious experience. The goal of perceptual (a.k.a., psychophysical) scaling is to find the mathematical functions that map the magnitudes of external stimulus dimensions to the conscious perception of magnitude. In turn, the goal of subjective (a.k.a., psychometric) scaling is to unravel the relationship between conscious experience and external manifestations of that experience, which is presumed to follow a functional mapping that allows researchers to deduce the underlying cognitive state of magnitude elicitations.

Numerous scaling techniques exist. Within this dissertation, my focus is on constrained scaling, a technique in which participants are trained to use a naturalistic measurement scale. Through a modest amount of training, the participant's internal perceptions are calibrated to an external numeric scale. Once the participant establishes the relationship between his or her mental magnitudes and the scale, the participant is able to translate this scale to represent a number of mental modalities. Compared to other scaling methods, constrained scaling has been demonstrated to reduce significantly both intra- and inter-participant variability (Boring, West, & Moore, 2002; West & Ward, 1994; West, Ward, & Khosla, 2000).

Fields that rely on human data have a tenuous relationship with psychological scaling methods. Although scaling affords the opportunity to gain insight into users' perceptions, it does so at a cost. The traditional cost of scaling is low adherence to a common scale and subsequent low scaling reliability. Constrained scaling offers a simple yet effective augmentation to classical scaling methods. Using constrained scaling, individuals learn to scale according to a common scale, thereby decreasing idiosyncratic variability between scalers.

Getting There

Over the course of this dissertation, I explore various facets of constrained scaling. The first chapter presents the central theoretical thesis of this document, namely, that the mind is a magnitude processor. Essentially this chapter sets itself the goals of extending the existing information processing framework in cognitive science and creating a novel model of mind. In this chapter, I attempt to highlight a major shortcoming in the conventional approach of cognitive science.

The second chapter is a brief history of scaling, which leads into the next chapter on constrained scaling. If the mind is a magnitude processor, then cognitive science had better have some good ways to measure mental magnitudes. In this chapter, I suggest constrained scaling is the method of choice and justify this claim with an exposition of general issues and the solutions that constrained scaling offers.

The next chapters discuss at length the 15 experiments on constrained scaling that I conducted for this dissertation. The experiments bridge many domains, from loudness and color brightness psychophysical scaling, to the psychometric scaling of the subjective happiness afforded by different amounts of money, to human factors scaling involving visual appeal and streaming media settings. These fifteen experiments serve to elucidate the model, method, and application of constrained scaling through a series of experiments that replicate, refine, extend, and apply constrained scaling.

The final chapter brings together the disparate findings from the 15 experiments and articulates the core findings, theoretical bridges, experimental shortcomings, and proposed future research. Finally, a series of appendices complements the content of the dissertation by providing technical details about the implementation of the experimental control software and about the calibration of the psychophysical stimuli.

Being There

The results presented in this dissertation provide compelling evidence that cognitively augmenting participants with a learned scale can substantially increase the reliability of psychophysical scaling and increase the sensitivity to individual differences in psychometric scaling. It is my hope that the experiments presented in this dissertation will strengthen the still emerging cognitive model of psychological scaling and lead to further implementation of constrained scaling as a technique to strengthen the role of psychological measurement to have the same reliability and validity as physical measurement techniques.

It has been said that a mind is a terrible thing to waste.[2] To this I would add that the measurement of a mind is an equally terrible thing to waste. Many existing psychological measurements fail to capture mental magnitudes effectively. By calibrating individual scale use, constrained scaling conduces better psychological measurement and thereby a better understanding of human mental processes.

[2] Attributed to the United Negro College Fund.

MIND AS MAGNITUDE

Introduction

The mind is a magnitude processing system. In this chapter I discuss magnitude processing as an important but overlooked component of cognition. I build upon the well-established information processing model of cognition by positing that magnitude is at the heart of information processing. It is my aim that this chapter will clarify magnitude as a starting point for cognitive research. Further, I aim to elucidate the fundamental connection between mental magnitude and scaling.

Cognition as Information Processing

Cognitive science has long embraced the information processing model of the mind (Dawson, 1998; Fancher, 1996; Gardner, 1985; Johnson-Laird, 1993), in which the mind is seen as a type of thoroughfare of mental information analogous to the flow of binary data within a computing system. Historically, the information processing view of cognition allowed psychological researchers to maintain many of the findings from the stimulus-response framework of the previously dominant behavioristic paradigm, while simultaneously shifting emphasis to the actual processing of that information. Whereas behaviorism's accounts of stimuli and responses served as the informational components of psychology, the emerging paradigm of cognition addressed the processing of that information. The transition from behaviorism to cognitive science is often painted in terms of a dramatic paradigm shift (Gardner, 1985). This transition may also be viewed less dramatically not as the wholesale abandonment of the behaviorist tradition but rather as the reframing of behaviorism's stimulus-response model in terms of information. In recasting behaviorism as information, the way was prepared for information processing to emerge.

The information processing model of cognition has not been without criticism (Dawson, 1998). The most succinct and poignant critique comes from Dreyfus (1992), who notes that artificial intelligence research based on the information processing model of cognition has developed techniques that succeed where humans fail and fail where humans succeed. In other words, there is a significant disconnect between human cognition and the type of cognition that has been modeled on an information processing system. The information processing approach lends itself well to solving logical problems, but it fails to account for the inherent illogic with which the human mind approaches its environment. Similarly, Gardner (1985, p. 385) noted,

> …(A)s one moves to more complex and belief-tainted processes such as classification of ontological domains or judgments concerning rival courses of action, the computational model becomes less adequate. Human beings apparently do not approach these tasks in a manner that can be characterized as logical or rational or that entails step-by-step symbolic processing. Rather, they employ heuristics, strategies, biases, images, and other vague and approximate approaches. The kinds of symbol-manipulation models invoked by Newell, Simon, and others in the first generation of cognitivists do not seem optimal for describing such central human capacities.

Cognitive science was born out of a tradition of logical, mathematical information processing. But, cognitive science has subsequently struggled to characterize the entirety of human cognition according to this model.

Despite criticisms of the information processing model of the mind, it has provided a measure of theoretical cohesiveness to the disparate subdisciplines of cognitive science (Boring, 2002). It has also proved itself a robust model of the mind in that new approaches to cognition have been assimilated into the mold of information processing. For example, the advent of connectionist frameworks has seriously challenged the notion that mental information is processed in the serial, central fashion akin to early computing systems (Dawkins, 1998; Quinlan, 1991). Connectionism under its many guises (e.g., parallel distributed processing or neural networks) has demonstrated that cognition can occur in a distributed and parallel fashion, not just in a serial fashion through a single central processing unit. Connectionist accounts of cognition are also more biologically credible than classical information processing accounts (Zipser, 1986). Importantly, even though connectionism has fundamentally altered the way researchers perceive cognitive processing, the basic notion of the mind as an information processing system endures. The specific details of how mental information is processed have been rearticulated through connectionism, but the basic model of cognition as information processing remains unaltered.

Even the so-called *hard problems* of cognition such as consciousness have not spelled the demise of the information processing model of cognition. For example, Chalmers (1996) suggests that the empirical study of consciousness may have thus far proved elusive because researchers have failed to identify that consciousness is a form of information that coexists with other mental events. Chalmers' model of consciousness frames consciousness as information processing. The challenge for cognitive science according to Chalmers is not in creating a cognitive model capable of accounting for consciousness but rather in developing tools and techniques that can measure the informational component of consciousness. Chalmers' approach can be construed as a type of information dualism, in which traditional forms of mental information are readily apparent to the researcher but in which the conscious form of information remains elusive. While Chalmers' approach is problematic, it is nonetheless intriguing in that it extends the information processing approach into heretofore elusive domains of cognition.

It is to this robust foundation of cognition as information processing that I append the notion of magnitude. Figure 1 illustrates the emergence of information as a psychological concept, from origins in stimulus-response behaviorism, to the nascence of information processing in the cognitive era, to my proposed incorporation of magnitude as a form of information. Before I proceed to detail the role of magnitude in cognition, I will first review the importance of magnitude in general terms to science.

Magnitude in Science

Magnitude is the measurable, countable, or comparative quality of something. Magnitude reflects a continuous quantum rather than discrete, categorical membership. It also serves as the basis of the empirical physical sciences, which have developed sophisticated methods to quantify physical magnitudes through measurement (Ellis, 1968). Galileo set the early stage for the importance of measurement by declaring that the goal of science was to "measure what is measurable and to try to render measurable what is not yet so" (cited in Berka, 1992, p. 181).

Berka (1992) accords the following characteristics to the materialistic measurement used in the physical sciences (p. 182-3):

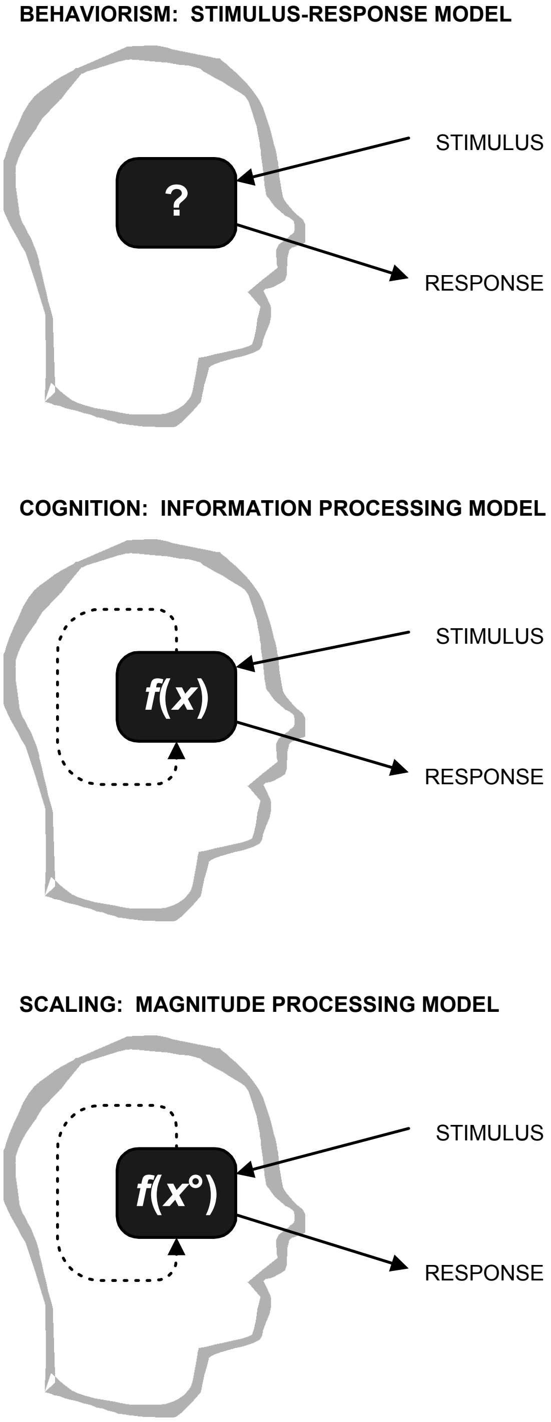

BEHAVIORISM: STIMULUS-RESPONSE MODEL
?
STIMULUS
RESPONSE
COGNITION: INFORMATION PROCESSING MODEL
f(x)
STIMULUS
RESPONSE
SCALING: MAGNITUDE PROCESSING MODEL
f(x°)
STIMULUS
RESPONSE

Figure 1. A comparison of three psychological models.

> 1. Measurement is ontologically committed (i.e., rooted in and, hence, grounded by objective reality.
> 2. Magnitudes are historically and theoretically determined reflections of quantitative aspects of objectively existing entities and not merely outcome of metricization or measuring procedures.
> 3. The object of measurement exists prior to metricization or measuring procedures.
> 4. In agreement with the historical determination of every phenomenon, a transfer of methods from one universe of discourse into another one is adequate only on the objective condition that certain structural similarities hold between the domains in question.

These points are derived from Berka's attempt to define measurement as currently used in the physical sciences, not from an a priori historical formalism that has guided measurement in practice. These four axioms roughly translate to mean that measurement is based on physical magnitude dimensions and that measurements should be a type of natural reflection of physical magnitudes, not a contrived formulation only made possible by complex measurement instruments.[3]

Historically, the *conservative conception of measurement* (Savage & Ehrlich, 1992) first espoused by Helmholtz (1887), conceives of measurement as the one-to-one correspondence of a physical property to a real number. For example, a metal rod of a given length might be used as one unit of length measurement. This rod would correspond to a numeric value of 1 in terms of measurement. A second rod of equal length placed adjoining the first rod would equal two units of that measurement. When presented with a novel object, the metal rods may be used to measure the length of the

[3] These statements represent strong positions that may not reflect measurement in practice, especially in the cognitive sciences. Those considerations will be discussed later in this paper.

new object. In the conservative conception of measurement, there is always a direct relationship between the magnitude dimension of an object and a real object that represents a numerical quantity.

The conservative conception of measurement encompasses physical magnitudes such as length, weight, angle, and the like (Savage & Ehrlich, 1992). It, however, fails to account for certain magnitude dimensions that cannot be directly linked to a physical object. For example, temperature remains an elusive magnitude to measure directly. Instead, temperature must be measured indirectly. Since it is known that objects expand and contract relative to temperature, it is possible to use this expansion and contraction in a lawlike manner to measure the effect of temperature on an object. It is not possible to measure temperature directly. A conventional thermometer actually measures the height of a temperature-sensitive fluid like mercury contained in a thin, long, translucent tube. As such, conventional measures of temperature are simply measures of the length of a fluid. The measurement relationship between temperature and underlying physical units remains indirect.

In order to account for the necessity of indirect measurements of physical magnitudes, more recent formulations of measurement theory use a *liberal conception of measurement*[4] (Savage & Ehrlich, 1992). In this formulation, numbers follow a specified functional relationship to magnitudes. The liberal conception of measurement affords a

[4] Savage (1970) also refers to the conservative and liberal conceptions of measurement respectively as the narrow and broad views of measurement.

more flexible view that accommodates the necessity to measure certain magnitude dimensions according to other magnitude dimensions.

An important concept in measurement in the physical sciences centers on the multidimensionality of measurement for any given object. As Kyburg (1984, p. 17) notes:

> Measurement is often characterized as the assignment of numbers to objects (or processes). Thus we may assign one number to a steel rod to reflect its length, another to indicate its mass, yet another to correspond to its electrical resistance, and so on. It is thus natural to view a quantity as a function whose domain is the set of things that quantity may characterize, and whose range is included in the set of real numbers.

Any given object has a multitude of magnitude dimensions in which it may be measured. While in many cases these magnitude dimensions may be orthogonal, they are often interrelated. For example, the 1889 definition of the magnitude of a meter was defined by the International Bureau of Weights and Measures to be equivalent to a graduated platinum-iridium cross section at 0º C (Penzes, 2002). Note that the fidelity of the measurement depended on temperature, another magnitude dimension. More recently, the 1983 definition of a meter is “the length of the path traveled by light in vacuum during a time interval of 1/299,792,458 of a second,” where the speed of light is 299,792,458 m/s and the light is defined as a helium-neon laser with a wavelength equal to 632.99139822 nm (cited in Penzes, 2002). The current definition of the length of a meter is thus measured in terms of precisely defined magnitude measurements of time and light wavelength.

As any middle-school Physics pupil discovers with glee, the gravitational pull of the earth on objects accelerates as an inverse function of time (i.e., $g = 9.8 \text{ m/s}^2$). The physical sciences similarly exercise an increasing enthusiasm for measuring physical magnitudes according to interrelated dimensions. The intention of the increasing multidimensionality of standardized measurements is not to obfuscate or to walk a precariously close line to recursion. Rather, these multidimensional measurements serve to minimize the variability in measurement. Whereas a physical object such as a rod made out of platinum-iridium might be subject to fluctuations beyond those accounted for by temperature, a wavelength of a burst of light measured in time brings a higher constancy to the measurement standard. Increasing the constancy of the standard ensures that measures made on physical magnitudes accurately reflect the characteristics of those magnitudes. The precision of empirical laws is necessarily limited by the noisiness of magnitude measurements. Hence, the goal of science is to achieve the highest measurement constancy and fidelity that are possible.

Magnitude in Cognitive Science

There exists a methodological rift between the physical and psychological sciences (Michell, 1997). While the empiricism of the physical sciences requires magnitude measurement, psychological empiricism has eschewed a strict reliance on magnitude measurement. While pre-cognitive methodological texts in psychology emphasized the primary importance of the measurement of mental magnitudes (Stevens, 1951), recent texts on cognitive methods make no reference to the measurement of mental magnitudes (Bower & Clapper, 1989).

This is not to say, of course, that psychology is not actively engaged in measuring mental processes. Measurement is, in fact, an essential part of most psychological experimentation. The key difference between psychological measurement and that employed in the physical sciences is that which is purported to be measured. With infrequent exception, cognitive psychology does not attempt to measure underlying mental magnitudes. The measurements of interest to cognitive psychologists are those measurements that are central to explaining what processing occurs, not what degree of processing occurs.

The research literature in cognition shifts much of the empirical focus from mental magnitudes to the process of mental categorization (Estes, 1994). For example, researchers in the field of visual object recognition investigate how shapes are grouped into holistic objects (Biederman, 1995), but they often ignore the sensory shape intensities that are necessary for shape recognition. Cognitive linguists focus on categorizing words into grammatical units while ignoring the fact that special grammatical categories—adjectives and adverbs—serve almost exclusively the role of magnitude placeholders (Simpson, 1944). Likewise, philosophers of mind may discuss *qualia*—what is like to experience a categorical state—while omitting a thorough discussion of the mental magnitudes that shape and define that experience. While I do not wish to contest the fact that the mind actively categorizes the information with which it is presented, the cognitive literature largely ignores the equally important fact that the mind requires magnitudes in order to categorize. Cognitive categorization is a means of delimiting magnitude information. As such, magnitude is a crucial component—if not the very essence—of cognition.

Human cognition is dominated by magnitude processing. For example, humans perceive stimulus intensities, whereby those intensities are magnitude representations. Humans make subjective judgments, in which the subjectivity culls a variety of magnitude information about which the judgment is made. Humans have emotions of various intensities. (In fact, magnitude is so central to human emotion that many forms of mental illness are defined in terms of insufficient or excessive affect.) Humans perform mental arithmetic, a process that involves mental representations of quantity. Humans also make relative comparisons, which entail weighing the magnitude dimensions of two or more things. While cognitive science rarely discusses mental magnitudes, magnitude is so inherent in human mental operations that the greater challenge may, in fact, be to identify mental processes that do *not* involve magnitude.

The most common techniques used in cognitive psychological research do not produce measures that are sensitive to magnitude. The staple measures of cognitive psychology are reaction time and error rate[5] (Bower & Clapper, 1989), which are overt manifestations of behavioral performance that can be linked to cognition. While reaction time does, in fact, record temporal magnitude, it may be argued that the time course of cognition when measured as reaction time is not an index of the use of magnitude as

[5] *Human error rates* are an inverse measure of *learning*. When learning occurs, error rates decrease (Reason, 1990). For this discussion, measures of learning are treated under the general umbrella of human error.

information. Unless the topic of investigation is specifically about time perception, reaction time serves as an indirect measure of cognition akin to the use of fluid heights to measure temperature. While reaction time certainly yields valuable insights into cognitive processes, it typically employed as a measure of cognitive performance, but not explicitly as a measure of cognitive magnitudes. Likewise, error rates are measures of cognitive performance that do not directly assess magnitude processing in the mind.

In order to clarify mental magnitudes, I put forth three levels of magnitude processing in the mind. I call the first level *sensation*. This first level involves the transformation of stimulus information into neural activity. Such processing occurs sensorially by inputting information about the environment to the mind. Sensory processes are well documented (Coren, Ward, & Enns, 1999; Goldstein, 2001), and the use of magnitude information at this stage of processing is widely accepted.[6] Unlike more malleable cognitive processes, sensation is largely hardwired into the organism. Every organism has a particular set of sensory stimuli to which it is receptive, and the biological mechanisms of receiving sensory information are largely invariant across organisms of the same species. While some adaptation does occur at the sensory level,[7] sensation is automatic and largely beyond the control of the organism.

[6] Magnitude is often relegated to sensory processing. I argue that magnitude permeates all cognitive processes.

[7] For example, the human eye adapts to the level of light it encounters. Light and dark adaptation are not under the control of the individual; the sensory organ—the eye—adjusts its light sensitivity automatically to meet the environmental context.

The second level of magnitude processing is *perception*. This level involves the interpretation and use of sensory information by the mind. It is at this level that sensory magnitudes are turned into meaningful information. It is also at this level that categorization occurs, in which sensory information is grouped into meaningful representations of the environment.[8] While perceptual categorization involves the parsing of some magnitude information, perception is not solely about categorization. Much magnitude information is retained during perception, as evidenced by the high degree of salient magnitude information that is available when someone tries to access this information.[9] While the most apparent focus of the person's attention is initially on categorizing information at the perceptual phase, the person retains full access to magnitude information.

The final level of magnitude processing is *cognition*. Cognition involves the use of internal information by the mind. To avoid confusion, what I mean by cognition is this

[8] Marr (1982) might argue for a stage of information processing between sensation and perception. Particularly in the case of stereoscopic vision, Marr's 2½-D sketch is a case in which magnitude information such as shading, lightness, and color is used in to create a magnitude-rich representation of the external world. This representation is further simplified and categorized in the final perceptual stage of visual information processing.

[9] For example, a person who walks through a deciduous forest during the autumn may generally perceive that the leaves of the trees are changing color. This is a form of cognitive categorization: the trees are categorized as a forest and the leaves are categorized as fall foliage. If, however, that person is asked to describe the fall foliage, he or she is able to give a reasonably detailed account of the myriad colors that comprise the fall foliage. The magnitude information used to categorize is still consciously available.

sense is not thinking in general but rather thinking as a self-guided process. Cognition, narrowly defined is this sense, draws heavily on memory, drives, and affective states as information input. Whereas all mental information processing encompasses cognition in the broad sense, this narrow definition of cognition only entails self-originating information processing. The information originates inside the head rather than through external sensory and perceptual input. Clearly, it is not possible to have a cognitive system that is completely isolated from external factors. In this case, the external factors are construed to have only indirect effects on the mental information involved. For example, external factors may trigger a particular affective state in the person. That affective state in itself promulgates memories and thoughts that are not necessarily directly relevant to the external event that triggered it.

Figure 2 illustrates the relationship of the three levels of magnitude processing in terms of information flow. While sensation and perception are clearly involved in magnitude input, cognition brings internal magnitude states into consciousness. These three levels of magnitude processing straddle the border between unconscious and conscious processing. Sensation, as already discussed, is a largely automatic process that serves to take in as much environmental information as possible. Perception serves as a filtering mechanism by which irrelevant information is eliminated. The point at which environmental information becomes perceptual information marks the threshold of consciousness. Cognition, narrowly defined, serves to bring unconscious information such as memories, drives, and affective states into consciousness.

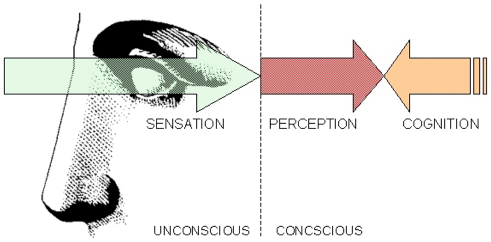
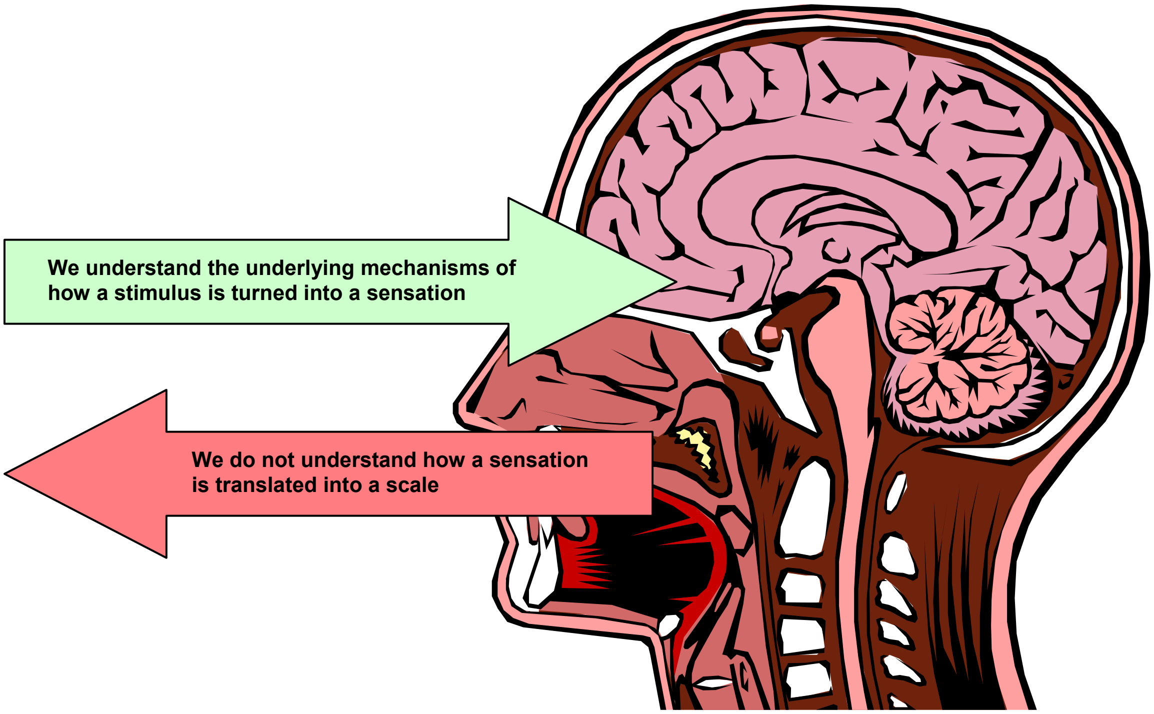


Figure 2. A depiction of the three levels of magnitude processing.



We understand the underlying mechanisms of how a stimulus is turned into a sensation

We do not understand how a sensation is translated into a scale

Figure 3. The perceptual process of inputting stimulus sensation is mostly understood; the generative process of outputting a magnitude on a scale is not.

The Measurement of Mental Magnitude

The psychological analog to measurement in the physical sciences is scaling. While cognition largely overlooks the measurement of mental magnitudes, the psychological subfields of psychophysics and psychometrics are built around measuring mental magnitudes. Both subfields use scaling—the process of introspectively accessing magnitude information and linking it to a numerical or other scale (see Figure 3). Psychophysics is concerned with scaling perceptual magnitudes (Stevens, 1975), while psychometrics is concerned with scaling subjective magnitudes (Kline, 1998).

Both psychophysics and psychometrics make four tacit assumptions about mental magnitudes:

1. The mind is a representational system.
2. Mental representations have magnitudes.
3. It is possible to access the level of mental magnitude through conscious introspection.
4. Mental magnitude maps to a numerical scale.

Point 1 is a standard assumption in cognitive science, namely that the mind mirrors internally what it perceives externally in the world. The specifics of how the mind actually represents this information are still the subject of considerable debate and will not be discussed here. Regarding Point 2, scaling researchers require that the internal representations include a magnitude component. A psychophysicist, for example, assumes that the mental representation retains the magnitude qualities of frequency and amplitude that are inherent in an actual tone played in the physical world. Similarly, a psychometrician assumes that an internal representation such as an opinion or subjective affect carries with it magnitude information not unlike those representations that reflect perceptions of the physical world. Both psychophysicists and psychometricians make the further assumptions that mental representations elicit magnitude information that is consciously accessible and that this information can be matched in the mind to a representation of numerical magnitude. Because this information is consciously accessible, the numerical magnitude representation can be relayed, thus translating the external magnitude into an external measurement.

Why Magnitude Matters

I have thus far discussed categorization as if it generally fails to retain magnitude information. Cognitive science treats cognition as the process of taking raw information and translating it into high level symbolic representations. For example, lines on paper are categorized into geonic objects (Biederman, 1995) and raw vocal utterances are categorized into phonemes (Liberman, Harris, Hoffman, & Griffith, 1957). In this view, meaning is derived from categorization, and meaning is conveyed in terms of concept membership. A categorizational view of cognition does not need magnitude information, because magnitude information is at best something to be filtered out in the process of forming clear concepts.

It is unclear why the categorizational view of the mind is so singularly attractive to cognitive science. Perhaps it is a carryover from the propositional logic of the Fregean tradition or a byproduct of a too literal embrace of the parallels between the mind and the computer. Regardless of its origins, it has fueled a long line of cognitive inquiry that has omitted mental magnitude processing.

I have not yet addressed why magnitude matters to cognition. Clearly, magnitude supports categorization, the most well documented aspect of cognitive information processing. But, why does magnitude need to exist beyond categorization? Why isn't categorization sufficient to account for cognition?

The simple answer to these questions is that categorization is a form of representational data reduction. Categorization impoverishes the information that is processed in the mind. While categorization ultimately eases the processing load on the mind, it strips the mental representation of its richness. Without magnitude in cognition, it would be impossible to gauge the uniqueness of a situation, the importance of novel information, or the vividness of experience. Without magnitude there would be no subjective experience, no memorable life events, and no rewards. In some ways, the behaviorists got it right, for they understood that a stimulus was not just a stimulus. A stimulus was a laden piece of information that affected the organism depending on the magnitude of its importance to that organism. A piece of chocolate is inherently rewarding to most humans, not because it is a categorizable object but because it is an object that triggers a variety of physiological and psychological responses that culminate in a sense of reward. This sense of reward results from the interplay of magnitude information in the mind, not from the categorization of objects. Magnitude matters, because without it organisms would be information processing automata for whom information was of little interest.

Categorization is, of course, an essential part of cognition. My intention is not to disregard the vital role of categorization. Instead, I wish to consider for a moment the synergistic role between mental categorization and mental magnitude processing. Magnitude processing feeds into categorization, while categorization serves as a framework for understanding magnitudes. Both are necessary facets of everyday mental functions. This dissertation redresses a shortcoming in the research literature by exploring the role of magnitude processing in cognition. The specific domain of exploration is scaling.

Summary

I have made five key claims in this chapter. These are:

1. The mind processes information as magnitudes.
2. Magnitudes are used to categorize mental information into meaningful composites.
3. Cognitive categorization does not preclude the possibility that magnitude information remains intact to co-exist with other mental representations.
4. Cognitive science has focused almost exclusively on categorization, ignoring the equally interesting cognitive domain of magnitude processing.
5. Scaling is a means of accessing mental magnitudes.

In the next chapter, I will explain more about the interaction of scaling and magnitude. I will also introduce a novel method of scaling, which, unlike existing approaches to scaling, incorporates cognition as a central component of the scaling process.

HISTORICAL OVERVIEW OF SCALING

Fechnerian Origins

It has taken considerable time for psychology to arrive at the five premises identified in the last chapter. The origin of psychophysics, and indeed the formal study of psychology, was with Gustav Fechner (1860a, 1860b), who suggested that the physical world was mirrored in the mental world in a lawlike, measurable fashion. While the notion of mental representation followed millennia-old epistemological and ontological traditions, the mathematical formulation that linked the physical world to the mental world was novel. Fechner called these two domains *outer* and *inner psychophysics* (*die Außenpsychophysik* and *die Innenpsychophysik*), respectively, signifying the relationship of the physical to the psychological world. Those representational states that fell within the realm of the mental world were inner psychophysics. Objective reality, which fell outside the realm of the mental world, was outer psychophysics.

Fechner's lineage as a student of and later collaborator with Helmholtz are foundational to the birth of psychology as a discipline (Fancher, 1996). Helmholtz is often considered the founder of measurement theory for his careful formulation of the best uses of measurement as a calibrated tool for objective empirical measurement. Just as Helmholtz developed measures to account for the physical world, Fechner developed measures to account for the mental world.

Fechner (1860b) considered the relationship between outer psychophysics and innerpsychophysics as the path between the stimulus (*der Reiz*) and the sensation (*die Empfindung*). This path was mediated by a lawlike nexus of psychophysical activity

(*Organ der psychophysischen Tätigkeit*), which was influenced by both memory and observation. Outer psychophysics involved translating stimulus information into neural information, a process that was unconscious. Inner psychophysics involved the interpretation of these neural messages into a mental representation. Since the relationship between stimulus and sensation was governed by lawlike processes, Fechner believed it was possible to measure this relationship with a mathematical equation. Weber had earlier discovered that the amount of change necessary to perceive a change in stimulus intensity was a constant function:

$$\frac{\Delta\phi}{\phi} = k, \tag{1}$$

where ϕ is a stimulus and *k* is a constant (Gescheider, 1997). This formula revealed that as the stimulus intensity increased, proportionately so did the amount of stimulus change necessary to perceive a difference. Weber's discovery, now appropriately called Weber's Law, specified only the relationship between stimuli. It did not yet connect physical stimuli with the mental world.

Fechner (1860a) discovered that stimulus units following Weber's Law could be chained together to create a measurement scale of sensation. The formal explication of the relationship between a stimulus (ϕ) and a sensation (ψ)—between outer psychophysics and inner psychophysics—was thereby established:

$$\psi = k \log \phi. \tag{2}$$

Fechner’s scale started at the absolute threshold of a stimulus and was measured in terms of just noticeable differences. The logarithmic function accounted for the curvilinear relationship between the stimulus and the sensation.

The just noticeable difference scale that Fechner proposed became known as Fechner’s Law (Gescheider, 1997). The lawlike relationship between the physical and the mental world was an exciting starting point for psychology (Boring, 1950), and it was quickly embraced by early proponents of the new discipline of psychology. One such proponent, Wilhelm Wundt, was so intrigued by the fact that it was possible to bridge the physical and mental worlds that he developed a series of techniques to allow researchers to access the mental world. The mode of access Wundt developed was introspection, a technique fraught with enough controversy so as eventually to render a rift between mainstream psychology and the study of mental magnitudes. In this vein, William James influentially concluded that Fechner’s work offered precisely nothing to the field of psychology (1890), affirming the split between psychology and psychophysics. This split has endured even to the present day’s division between the magnitudinalists in psychophysics and the categorists in cognitive psychology.

Steven’s Power Law

Fechner’s Law is a type of indirect measure of mental magnitude. Fechner’s scale measures the levels of stimulus necessary to perceive differences, but it does not actually yield the direct magnitude perception triggered by the stimulus. S.S. Stevens (1975) surmised that a direct measure of magnitude was possible. His approach was simply to pair magnitude perceptions to a numeric scale as uttered in the form of a number. This approach, called *magnitude estimation*, yields a different relationship between stimulus and sensation than Fechner's scale, requiring a power transformation:

$$\psi = k\phi^{\beta}, \tag{3}$$

where ψ is sensation, k is a constant based on the unit of measure, ϕ is the stimulus, and β is a sensory-specific exponent value. Equation 3 is now known as Stevens' Power Law to denote the lawlike relationship between stimulus and sensation for direct scaling measures.

Stevens and Galanter (1957) found that the Power Law approximated Fechner's Law for categorical scaling but not for continuous scaling. Ward (1974), among others, found that the Power Law holds for categorical judgments but produces an exponent that is about half the size of the exponent produced for continuous scale judgments. Krueger (1989) has suggested that to reconcile Fechner's Law with Steven's Power Law, it is necessary for Fechnerian psychophysics to forfeit its strict reliance on Weber's Law and for Stevensian psychophysics to reconsider the notion that magnitude estimates are a direct measure of underlying sensations. A mathematical treatment that requires less compromise is offered by Norwich (1987, 1993; Norwich & Wong, 1997), which states:[10]

$$\psi = \frac{k\ln(1+\gamma\phi^{\beta})}{2}, \tag{4}$$

where γ' is a new constant. For large values of $\gamma\phi^{\beta}$, the equation takes the form:

[10] Norwich's equations have been aligned with Steven's notation.

$$\psi = \frac{k\beta \ln \phi}{2} + \frac{k \ln \gamma}{2}, \qquad (5)$$

which is a form of Fechner's Law. For small values of $\gamma\phi^{\beta}$, the equation takes the form:

$$\psi = \frac{k\gamma\phi^{\beta}}{2}, \qquad (6)$$

which is a form of Stevens' Power Law. Norwich's so-called Informational Law successfully reconciles Fechnerian and Stevensian psychophysics under one parent law.

Psychometrics

Measurement of mental magnitudes in psychometrics is similar to psychophysics. Psychometrics takes the ideas postulated by Fechner one degree further. Whereas Fechner outlined the measurement course for psychophysics as the relationship between the physical world and the mental world, psychometrics omits the physical world. Thought is not always a response to stimulus promptings from the physical world. Thus, psychometrics measures the level of internal magnitudes for which there is no clear physical stimulus. The measurements used in psychometrics are necessarily unidimensional, because there is no clear secondary dimension to which mental magnitudes can be related (Kline, 1998). Consequently, psychometrics does not benefit from the lawlike multidimensional functions characteristic of psychophysics. Nonetheless, there exists a rich tradition of psychometric research, starting with early attitude scales by Thurstone (1919) and Likert (1932) and extending to a rich variety of multivariate techniques currently in practice.

Magnitude and Categorization

Psychophysics and psychometrics offer scaling techniques to measure mental magnitude, but cognitive science seldom counts these among its methods. While it is easy to point at cognitive science for this shortcoming, this methodological oversight is actually bi-directional. While cognitive science has failed to embrace scaling methods, psychophysics and psychometrics have likewise failed to embrace insights from cognitive science. In the next chapter I outline how cognition is absolutely necessary to improve the scaling fidelity of psychophysical and psychometric methods. First, however, it is important to clarify the co-existence of magnitude and categorization in the mind, for these are the domains of psychophysics/psychometrics and cognitive science, respectively.

Holistic and Analytic Information Processing

I have already discussed the co-occurrence of categorization and magnitude in terms of mental information processing. Before I proceed with further points about mental magnitude and scaling, it is important to consider the theoretical basis of the distinction between categorization and magnitude processing. The essence of this distinction is with holistic and analytic views of cognition. Holistic and analytic distinctions have been the source of considerable research and discussion in psychology, both prior to and concurrent with their emergence in information processing.

Already early in the history of experimental psychology, there was debate over the deconstructability of mental processes. Edward B. Titchener, chief proponent of the structuralist movement in psychology, suggested that consciousness should be investigated in terms of reduced component processes, rather than as a singular, irreducible process (Titchener, 1911).[11] Titchener's atomistic view of mental processes contrasted strongly with the views of his German mentor, Wilhelm Wundt, who believed that the introspective methodology then at psychology's disposal could not adequately analyze the components of higher level mental processes. Wundt's critique, however, was generally overshadowed by the prominence achieved by Titchener in the English-speaking world (Schulz & Schulz, 1999).

It was the Gestalt psychologists, most notably Max Wertheimer, Wolfgang Köhler, and Kurt Koffka, who developed the strongest argument against structuralist analysis of thought (Fancher, 1996). Through a series of now famous perceptual illustrations, the Gestalt psychologists demonstrated how the mind operated according to grouping principles. No matter how hard a person might try to separate the components of visual objects that he or she saw, these objects were unfailingly perceived in groups. For example, when looking at a tree in an empty field, the tree would appear as a single object in the mind of the perceiver. Of course, the viewer might also focus on an individual aspect of the tree such as its bark, its branches, or its leaves. However, this conscious decomposition could never diminish the wholeness of the tree nor the

[11] What Titchener and other introspectionists meant by consciousness is essentially the conscious part of cognition. Because introspectionism was the chief empirical method available to structuralist psychologists and because introspectionism relied on conscious access to cognition, it can be argued that Titchener's arguments were not specifically meant to preclude unconscious cognitive processes. Here, I assume that Titchener would have meant for his arguments to apply equally well to cognition in general, not simply to conscious cognition.

subconscious perception that the bark, branches, and leaves were part of a single object.

The key to the Gestalt argument was that this mental categorization occurred automatically, as a subconscious process. In modern parlance, the process of grouping is said to be cognitively impenetrable (Pylyshyn, 1984), meaning the perceiver does not have conscious access to the mental processes that caused the perception of a single object from the bark, branches, and leaves. Moreover, because the process is cognitively impenetrable, the perceiver is not able to control the categorization processes that are at work. Objects such as bark, branches, and leaves, when arranged in the fashion accorded by nature, always comprise a tree. Normal human object perception can never disconnect those components that group together to create a single object. This is not to say that Gestalt psychology believes perceivers can never be aware of the individual items that are grouped to make an object. Gestalt psychologists would never mean to suggest that stimuli are indivisible into constituent properties; rather they would suggest that the perception of stimuli occurs as the synthesis of these parts rather than as separate processes (Smith, 1988). Using the earlier example, the perceiver is, of course, aware of the bark, branches, and leaves. What the perceiver is not aware of is how the mind joins these items to form tree objectness. It is the process of mental union of these items—categorization—that is cognitively impenetrable.

At the heart of the structuralist/Gestaltist debate is holistic and analytic theory. Titchener and other structuralists espoused an analytic view of the mind, which is to say that a mental process is the sum of its constituent parts. Conversely, the Gestalt psychologists argued for a holistic view of mind in which a mental process is viewed as more than the sum of its parts. At issue is also whether or not thoughts are cognitively penetrable. An analytic model of the mind holds that mental processes are comprised of discrete mental steps, which may be accessed on a conscious level. A holistic model of mind holds that mental processes are comprised of automatic processes that are not consciously accessible.

Neither the structuralists nor the Gestaltists expressly discussed mental magnitudes. Nonetheless, these early foundations of the analytic and holistic camps implicitly inform the discussion of mental magnitudes. Analytic theorists, intent on finding the atoms of human thought, would tend to espouse a view that incorporated magnitude as an essential and cognitively penetrable part of cognition. Holistic theorists, in contrast, would tend toward a view that emphasized categorization as the most elemental level at which cognition should be discussed.

The debate on the holistic-analytic distinction of mental processes has continued, but recent research has offered explanations that point to the coexistence of holistic and analytic processes. For example, Smith (1988) suggests that holistic perception is linked to automatic, subconscious cognitive processes, whereas analytic perception is a result of deliberate, conscious cognitive processes. Smith points to common two-stage models of perception, in which there is typically an initial, fast, and automatic process that functions holistically. There is also a second, slow, and effortful process that works analytically. Both processes must occur in order for perception to take place.

The coexistence of holistic and analytic processes is also found in dual-route models of cognition, such as Coltheart's (1978) dual-route model of written word recognition. According to these models, when people read aloud, they read words that occur frequently in a language automatically, according to a lexical route of processing. In this case, the word itself forms a unit of perception that corresponds with a series of phonemes to be uttered. The lexical route operates holistically, since the word itself forms the smallest unit of perception in the mental lexicon. For words that occur less frequently in language, people may make use of an effortful naming process according to a phonological route of processing. In phonological processing, a word is sounded out in a process that transcribes particular orthographic combinations into phonological utterances.[12] This process is often considered an analytic process, since word naming occurs as a series of discrete, cognitively penetrable mental processes. Coltheart's model suggests that either holistic or analytic processes are used depending on the context. Paap and Noel (1991) demonstrate that the two processes may occur simultaneously, in a type of race to see which route names the word the fastest. Their model extends the holistic/analytic framework to suggest that mental processes, particularly in word recognition, occur simultaneously in holistic and analytic fashion. Paap and Noel's findings have not yet been fully explored in terms of their implications for holistic and analytic theories of mental processing. Nonetheless, their model holds important implications for establishing a balanced view of holistic and analytic processes.

[12] Of course, some level of holistic processing occurs even when reading an individual letter of text. Different lines are grouped together to create a holistic representation of that letter. It is important to note that I am describing a process by which holistic *Gestalts* are successively broken into smaller and smaller units. The decoding of whole *Gestalts* eventually reaches a level at which the shapes become sensory magnitudes.

A unified holistic-analytic approach to cognition holds important implications for magnitude processing. It suggests that there can be a symbiosis between categorization and magnitude processing. While most automatic and conscious mental processing results in categorization, underlying this categorization is magnitude processing. This magnitude processing, while generally overshadowed by categorization, exists and may be consciously accessed when a person wishes to do so. Moreover, in cases where categorization is not automatic, magnitude information may become conscious as the person attempts to make sense of the information at hand.

A variation on analytic and holistic theory is Fodor's theory of modularity (1983). Fodor holds that the mind necessarily processes information discretely in specialized mental organs known as modules. Fodor does not, however, suggest that these modules must be cognitively penetrable. In fact, Fodor holds that conscious awareness only occurs holistically. The complete mental picture of mental processing only occurs once the discrete modules have assembled their information. Fodor's view can be considered a composite model. The modules operate very much according to an analytic framework, with the exception that modules are encapsulated and not cognitively penetrable. Once the modules assemble all stages of processing, Fodor's model is akin to a holistic model, since he emphasizes that it is only in the composite of modular processes that processing becomes meaningful.

Fodor's theory of modularity is often contested (Karmiloff-Smith, 1992, 1999; Lyons, 2001), but it illustrates some of the complexity that is prevalent when attempting to merge holistic and analytic theory. Fodor's theory of modularity is also the key to understanding the underlying importance of magnitude to cognition. The basic question concerning modularity is to what extent the modularity model accommodates mental magnitudes. I have suggested that categorization is essentially a holistic process, whereas magnitude processing is essentially an analytic process. I have also suggested that Fodor reconciles holistic and analytic approaches through his modularity model. The question remains: Does modularity reconcile categorization and magnitude processing in cognitive science?

Fodor clearly did not have magnitude in mind when writing *The Modularity of the Mind* (1983). Fodor's modularity largely continues the well established tradition in cognitive science of equating categorization and cognition.[13] Modules are simply means to categorize the flow of information in the mind. To reconcile modularity and magnitude, there are three possibilities:

1. Discount modularity and suggest that it is not a viable model of cognition.
2. Suggest that magnitude coexists with the categorization that occurs in the modules.
3. Suggest that there is a magnitude module (or series of modules).

There is good evidence to suggest that modularity is a viable model of cognition (Gazzaniga, 1989). It would therefore not be a productive avenue to seek to discount

[13] Fodor (1998) has recently offered that concept formation—a form of categorization—is at the heart of cognitive science. While he suggests that this concept fixation has misguided cognitive science, he does not go so far as to offer a magnitude based alternative account of cognition. His point is based on the vacuousness of concept representations, a point which certainly holds true if information is represented devoid of all magnitude information. It is unclear, however, to what extent his argument holds if magnitude is an intrinsic part of cognition and of concept formation.

modularity in favor of magnitude. Clearly, both need to exist in order to realize a sufficient model of cognition. There is also good evidence to suggest that sensory information is processed in multiple modules (Livingstone & Hubel, 1988), which implies that at the very least magnitude information serves as input for modules. Experiments in perceptual scaling (West et al., 2000) also hint that a common scaling process applies to different sensory modalities, suggesting the possibility that there may be a magnitude module that is responsible for taking magnitude information and translating it into meaningful, consciously accessible quantity information. There is no conclusive account for the role of modularity in a magnitude-based cognition. I suggest that the only feasible account at this point is that magnitude information is either retained through modular categorization or coexists to join post-module information in central consciousness. Because magnitude information is consciously accessible and cognitively penetrable, it is clear that magnitude information is retained through any modular processing.

CONSTRAINED SCALING

Introduction

In the previous chapter, I explained the centrality of magnitude to cognition. I demonstrated that it is plausible to think of magnitude as occupying three stages of information processing, incorporating sensation, perception, and cognition. I also explained how magnitude coexists and supports other cognitive processing such as categorization. In this chapter, I discuss the measurement of mental magnitude in more detail.

The mind processes magnitude information, and psychophysical and psychometric scaling are the ways to access these magnitudes.[14] What is missing from this tidy account is an explanation of how scaling occurs. Figure 4 illustrates the processes involved in scaling the magnitude of a physical stimulus. Magnitude information is first extracted through the process of sensation. Then, perceptual processes create a mental representation of the magnitude. Finally, cognition translates the magnitude representation to a scale. Commonly, magnitude representations are translated into numerical scales, which require the enlistment of mathematical cognitive skills to produce a numerical representation to match the magnitude representation in the mind. Note that Figure 4 does not fully incorporate the process of generating a scaling

[14] Another class of magnitude measurements consists of physiological measures. Using measures such as galvanic skin response, electrocardiograms, and heart rate, it is possible to determine a person's physiological response to a stimulus. These measures circumvent introspective response in a way that does not necessarily gauge conscious response. These are indirect measures, since they do not necessarily measure conscious experience directly but rather the effect of conscious experience on the body.

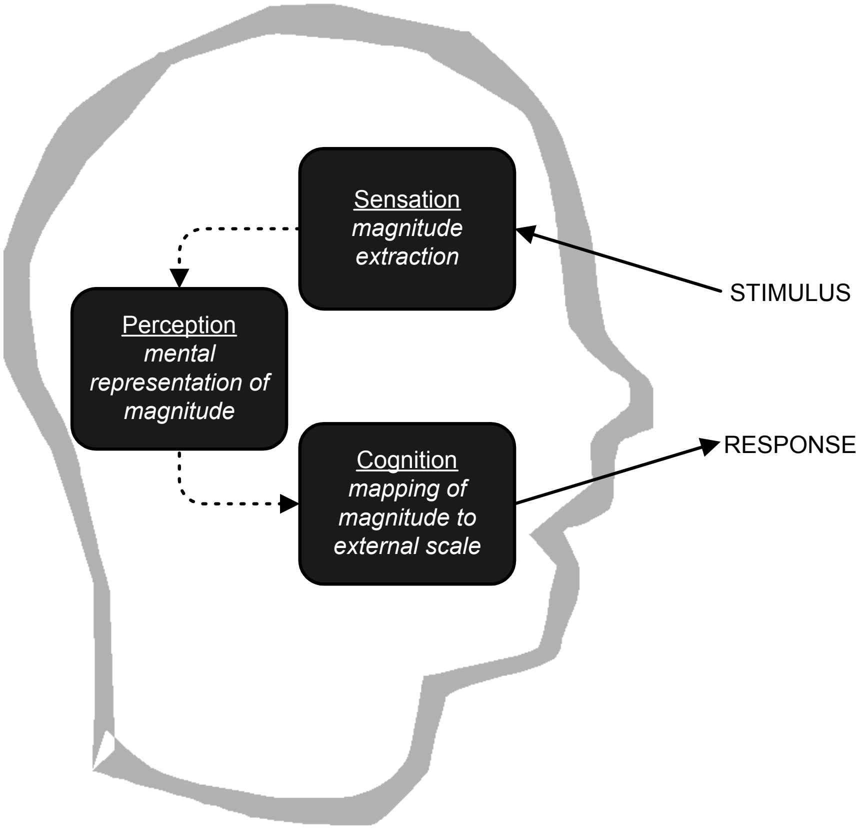

Sensation
magnitude extraction
STIMULUS
Perception
mental representation of magnitude
Cognition
mapping of magnitude to external scale
RESPONSE

Figure 4. Processes involved in scaling a magnitude perception.

response, since this may entail different proficiencies depending on the mode of response. To generate a verbal response of the magnitude quantity, spoken language processes would be required; to generate a written response of the magnitude quantity, written language processes would be required; to generate a response on a line scale, a variety of processes including psychomotor processes would be required. Figure 4 simply illustrates the basic processes required in any magnitude scaling.

Calibration and Scale Units

As I discussed previously, measurement is the key to physical science. An equally important hallmark of physical science is the need to calibrate the measurement apparatus. Calibration refers to the notion that one apparatus will obtain the same measurements as another apparatus. Measurement in physical science is not simply about affixing numbers to observations. Measurement is about affixing numbers to observations in a consistent manner across both observations and observers. Calibration allows replication, in that it is possible for one scientist to reproduce the measurements of another scientist. Calibration also facilitates generalization, as the scientist may be sure of the connection of one measurement to another measurement. For example, the equality of 1 liter of water and 1 kilogram of mass at 0° C at sea level can be relied upon when the measurement instrumentation is calibrated to the metric standard.

An important distinction needs to be drawn between calibrating a scale and setting its unit of measure. It is possible to use any variety of units for a scale. The classic examples of units of scale are the metric and Imperial measures. While the metric system boasts a cleaner lineage and a wider implementation, there is no measurement advantage of using one system over another. A metric unit of measure has an *exact* equivalent Imperial unit of measure. The two measurement systems use different scale units to measure the same thing, but a unit on one scale always maps precisely to a unit on the other scale.

Calibration, on the other hand, prescribes the fidelity of that measurement unit. Calibration is the extent to which a measurement apparatus matches a predefined standard of measurement. A calibrated metric apparatus will report unit values in accordance with metric standards, whereas a calibrated Imperial apparatus will report unit values in accordance with Imperial standards. The degree of fidelity to the predefined standard marks the level of calibration, whereas the mapping of a physical entity to a specific scale constitutes the process of setting the unit of measure.

A clear example of the importance of scale calibration and scale units comes from the historic chain of events that led to the modern thermometer (Middleton, 1966). It was the Florentine *Accademia del Cimento* between 1657 and 1667 that first tried to match different thermometers to a common scale. The Florentine Academy produced three sizes of spirit-in-glass thermometers with three corresponding temperature ranges of 50, 100, and 300 degrees. The smallest of the thermometers, the 50-degree thermometers, were observed to have the greatest temperature agreement between different thermometers in the series. Early accounts of this finding suggested that the glassblowers had an easier time forming consistent glass tubes for the small 50-degree thermometers, whereas the large 100- and 300-degeree thermometers exhibited a high level of variability in the manufacture of the glass tubes. Thus, the earliest account of thermometer calibration links the glassblowers' consistency in creating thermometers of exactly the same size with measurement consistency.

A different approach to calibration was developed by Robert Hooke in England around 1665 (Middleton, 1966). Instead of trying to make thermometers the same size, he calibrated thermometers against each other by changing the temperature of both thermometers and marking the corresponding points on the thermometer over time. Starting near freezing, he marked the fluid level in each thermometer and then gradually increased the temperature, marking the fluid levels on both thermometers as the temperature rose.

Once this method of calibrating thermometers was developed, it was possible for subsequent experimenters to devise scales against which thermometers could be calibrated (Middleton, 1966). In the seventeenth century, Daniel Gabriel Fahrenheit calibrated his thermometer at two points, the temperature of melting ice (0˚) and the temperature of the human body (100˚). In the eighteenth century, Anders Celsius also calibrated his thermometer at two points, which corresponded to the temperature at which water freezes (100˚) and the temperature at which water boils (0˚). Note that the polarity of the Celsius scale was later reversed, although his calibration points remained the same.

Calibration and the unit of measurement are important scientific complements of each other. Combined, they allow scientists to talk about the same observable phenomenon in the same manner. While they are separate concepts, calibration and the unit of measurement interact to give consistency to measurement. Without calibration and a common set of measurement units, science is fragmented by its inability to communicate its results. A similar process is at play in scaling of mental pheneomena. Without calibration and a common set of measurement units, psychological scaling cannot adequately explain mental phenomena in a consistent or replicable manner.

The Mind as a Measurement Apparatus

In the context of scaling, the mind operates much like a measurement apparatus. In psychophysical scaling, it senses a physical stimulus and produces a measurement to reflect the magnitude of that stimulus. In psychometric scaling, it produces a measurement to reflect mental states that do not necessarily have a clear correlate in the external, nonmental world.

Using the earlier example of thermometers, the analog between a physical measurement apparatus and human scaling is clear. Humans are sensitive to temperature and can provide approximations of temperature according to the following scale:

$$W = k(T - 305.7)^{1.6}, \tag{7}$$

where *W* is the perceived warmth, *k* is a constant, and *T* is the temperature in degrees Kelvin (Stevens & Stevens, 1960). *W* is an estimate of the temperature of a piece of heated aluminum touching the skin. The exponent value changes to 1.0 when a cooled piece of aluminum is applied to the skin.

Given that humans do not respond consistently to cold vs. warm temperatures, it would seem that humans are poor thermometers. One could imagine that human temperature sensing abilities are even more confounded outside the constraints of the laboratory. Physical thermometers provide accurate measures of temperature in degrees Celsius or Fahrenheit, yet humans as thermometers often feel the temperature to be a little colder or a little hotter than it actually is. Humans are especially sensitive to wind and humidity, making our subjective perception of temperature sometimes quite different than reality. Meteorologists have developed wind chill factors and humidexes to account for our proclivity to stray from objective mercuric temperatures. While we primarily use wind chill factors and humidexes to gauge how much or how little clothing we need to wear in order to stay comfortable on a given day, these scale adjustments are actually transformations that calibrate human temperature perception to an objective scale. Given a known level of humidity or a known wind speed, it is possible to calibrate subjective temperature perception to an objective temperature scale.

Contrary to first appearances, humans are not poor thermometers. We are simply poorly calibrated thermometers. Applying a calibration correction for wind speed and humidity makes humans into relatively good thermometers.

What about the scale units that the mind uses to report perceived temperature? In contrast to most mental magnitudes, temperature presents a case in which the units of measurement are learned by humans at an early age.[15] An American child growing up in northern Montana learns to associate 90˚ F with a warm summer day. A Canadian child growing up 50 miles (or 80 km) to the North in southern Alberta learns to associate 32˚ C

[15] Other commonly learned units of measure include distance, weight, fluid mass, and money. Interestingly, in places like Canada and Britain that have officially gone to the metric system from the Imperial system, it is these units that are the slowest to be converted in the minds of that country's inhabitants. This phenomenon is especially pronounced with the measures that are intimately connected to a person. While a Canadian may talk of a 500 km trip, his or her personal height is still measured in feet and inches. While a Briton may talk about a 1200 kg car, his or her personal weight is still measured in stones and pounds.

with the same warm summer day. These two children grow up with an identical climate—and, thus, identical temperature calibration points—but they learn two very different temperature scales. Should that American later become an expatriate living in Canada, he or she will have great difficulty relearning the Celsius units of measure that are ubiquitous to Canada. Likewise, should the Canadian later become an American immigrant, he or she would likely struggle to match Fahrenheit units of measurement to the well learned Celsius units of measurement. These people would eventually become adept at converting temperatures according to the following equations (or some quickly calculated approximation of these equations):

$$^{\circ}F = \frac{9}{5}(^{\circ}C) + 32 \quad \text{or} \quad ^{\circ}C = \frac{5}{9}(^{\circ}F - 32), \tag{8}$$

but the actual relearning of the new temperature scale would be slow at best.

This example of humans as thermometers illustrates several important concepts in psychophysical and psychometric scaling.

1. In many cases, it is possible through corrections and transformations to map human subjective scaling to objective physical measurement. This is the case with psychophysical scaling.
2. Human scaling is calibrated according to exposure to the range of the phenomenon. For example, a person raised in a cold climate would be calibrated to a colder range of temperatures than another person raised in a tropical climate.
3. Both calibration and the unit of measurement are learned in human scaling.

4. It is possible for scaling to occur that is neither calibrated nor attuned to a particular unit of measurement.

It is this latter point that concerns the remainder of this paper. There are phenomena to which we have learned to respond with at least a modicum of accuracy according to a physical measurement scale. Temperature is a prime example of such a phenomenon. But, there is a possibly infinite number of mental magnitudes for which there are no calibration points and no learned scale units. What does this lack of mental calibration or measurement units mean for psychological scaling? Just as in physical science, calibration and measurement units should be the hallmarks of psychological measurement. A failure to control for these factors would bring into serious question the validity of scaling results for mental phenomena.

Problems with the Mind as a Measurement Instrument

Given a short, pure tone of average frequency and amplitude (e.g., 1000 Hz at 70 dB), what scale value would someone assign to represent the loudness of that tone? The answer depends on a number of factors. First, the scale value depends on the range of the scale. A scale from 1 to 10 would produce quite different results than a scale from 1 to 100. Still, even when the range is specified, the reliability of the scale value across different people could be quite low. Assume for the moment that two people have equal hearing abilities and the same mental magnitude perception of the loudness of the tone. Using a simple ten-point scale, one person's loudness rating of "5" might mean the same thing as another person's rating of "7." Setting the endpoints of a scale does not calibrate the mind. In calibrating a physical measurement apparatus, it is sufficient to calibrate a few points on the scale, especially if the properties of the measurement apparatus are known. For example, when calibrating standard spirit-in-glass thermometers, it is sufficient to match a few points for temperature equivalence, because the thermal properties of the fluid in the thermometer are understood. There is no luxury of mental equivalence in humans. Even when humans are given the endpoints on a scale, this does not necessarily affect the way in which they scale between those endpoints.

Poulton (1989) provides a comprehensive list of biases in scaling. By *bias* is meant the fact that people do not treat a scale in a consistent, linear fashion.

1. There is, for example, a series of contraction biases, in which people do not use the full range of the scale available to them. A common form of this occurs when people scale most magnitudes too closely to a central scale value.
2. Other biases occur when people use inappropriate units of magnitude. Such would be the case, for example, when judging temperature on a 100-point scale according to the already familiar Celsius scale. The familiar Celsius scale could hamper the ability of a person to scale using a different temperature scale.
3. When people do not use a scale consistently across the scale range, they often exhibit a logarithmic response bias. This may happen when a person does not know how to map a magnitude to a scale in a particular range. For example, if a person scales a stimulus in one-step increments in a low range and then uses ten-step increments in a high range, this would result in a logarithmic shaped curve.[16]

[16] Note that the logarithmic curve is expected in psychophysical scaling and is, in fact, the basis of Fechner's Law and Stevens' Power Law.

4. In some cases, people scale using the entire range of a scale, even when the stimulus does not warrant a full range of responses. In such cases, people exhibit range equalizing biases. If, for example, a subset of stimuli is presented at around the midpoint of a scale yet the person provides responses spanning the full range, then that person is equalizing the range of his or her scale response.
5. When people use all scale responses above and below the midpoint of a scale equally often, there are centering biases. For example, when a person is presented with a range of loud stimuli, that person might assign loudness values localized around the midpoint of the scale. When presented with a range of quiet stimuli, the person might assign roughly the same range of responses. Although the stimuli are clearly different, the response bias of the individual tends to center the scale responses, thereby minimizing the scale differences between the two stimuli. Helson's adaptation-level theory (1964) accounts for this phenomenon, in which people adapt their perceptual response according to the intensity of the stimuli with which they are presented.

Although most of Poulton's scaling biases are most clearly illustrated through psychophysical scaling examples, these basic biases are endemic to psychometric scaling as well. These biases most commonly manifest themselves in the form of a failure to use the full range of the psychometric scale (i.e., a contraction bias) or a tendency inappropriately to use the extreme ends of a scale (i.e., a range equalizing bias).

With these many biases coming into play, psychological researchers need to exercise caution before treating their measures of mental magnitude with the kind of confidence that a physical scientist might exercise. Too often, psychological researchers correct for biased scaling results without remedying the cause of the biases. Large numbers of participants are enlisted for a study and/or a large number of repeated trials are executed to accommodate the need for statistical effect sizes. Buried in statistically significant averages are often individual differences in the use of scales. The statistical tests that are the mainstay of psychological research carry the inherent ability to occlude these differences and biases in scale usage. Without calibrating mental scale usage, psychological research risks being an artifact of chance scaling overlap rather than a true science of measurement.

Laming (1997) expounds on the scaling biases identified by Poulton. He suggests that the results found in psychophysical studies are typically an artifact of the scaling method more than a reflection of internal sensory states. Laming argues that magnitude estimation, for example, produces results that are a function of the relative judgment of the range of stimuli rather than a reflection of sensory percepts. In his view, without the presentation of a range of stimuli, the magnitude estimate does not exist, because the estimate cannot occur independent of the context of neighboring stimuli. As evidence, Laming presents the considerable variance found in magnitude estimation results, whereby two-thirds of score variance is inherited from the preceding score. Laming is skeptical of the ability of any measurement tool to capture the sensation[17] of mental

[17] Note that Laming's use of "sensation" excludes a full account of the cognitive representational system that I have classified as "perception." Perception is the basis for psychological scaling in the present document. Laming's blended view of sensation and perception seems at odds with a cognitivist view of the mind as a representational system.

percepts. He suggests, “Judgments of sensory magnitudes are known to be much subject to bias…. A theory of the measurement procedure would tell us how to avoid those biases or, at least, how to correct for them” (p. 25). As shall be seen in the remainder of this chapter, constrained scaling offers exactly such a method to calibrate the individual and thereby reduce score variance. Constrained scaling is both a theory of mental measurement as well as a procedure for improving the measurement of perception.

Calibrating the Mind

Ward (1992) introduced an important consideration into the psychological scaling literature. He suggested that much of psychophysics aims to eliminate biases in order to reveal the *true* psychological scale. Ward argued that the notion of a true scale was based on a static model of the mind. As an alternative, he offered an account of the mind as dynamic and distributed. Ward believed that the biases that surfaced in scaling were clues to the status of the mind at the time it was being measured. He further suggested that when the object of study was not the mind in flux but rather the underlying processes that control mental phenomena, it was necessary to control the situation as much as possible (p. 220):

> If we want to make precise and repeatable statements about phenomena that are mediated by a dynamic distributed mind we must control the situation so that the mind we are engaging is both known in detail and as invariant as we can make it. This implies that scales should be defined so as to be useful (render laws simple and elegant) and then subjects should be taught how to use the scales in situations that have been studied and analyzed so as to engage a known and consistent subset of the agents of mind.

With this statement, a new approach to scaling was initiated, one that subsequently came to be known as *constrained scaling* (West & Ward, 1994).

Constrained scaling works by both calibrating the individual to a mental scale (West & Ward, 1994) and by providing a natural set of scaling units (West et al., 2000). An individual is calibrated to a scale by receiving a set of training stimuli. Each stimulus is presented, after which an experimental participant estimates the magnitude of the presented stimulus. Finally, the actual scale value is presented to the participant. Over a series of training trials, the participant learns to match his or her magnitude perception to the scale. In order that a scale may be readily learned, it must represent a natural scale such that it can be fit to Stevens' Power Law in Equation 3.[18] Once the participant has learned to match his or her perceptual magnitudes to a scale, the participant receives a novel set of stimuli to scale according to the learned scale. The participant receives no feedback for the novel stimuli, but scale learning is supported and enhanced with reminder trials in which stimuli from the original scale are presented again with feedback.

Whether answering the need for a dynamic model of scaling (Ward, 1992, 2002) or addressing a way to remove bias in order to arrive at true psychological measures

[18] A natural scale refers to a scale that may be readily learned (Ward, 1992), because it follows the characteristics of how a stimulus range is mapped to mental magnitudes. For the purposes of this discussion, I assume that power laws using Stevens' (1975) exponent values, derived through magnitude estimation, are natural scales. For some of these it has actually been shown that they are easy to learn (Marks, Galanter, & Baird, 1995; West et al., 2000). It is interesting to note that these same studies indicate that exponent values above Stevens' values are harder to learn. Teghtsoonian (1971) suggested that the function of the exponent value is to compress different stimulus ranges onto a single, fixed magnitude range. Under this interpretation, these results could be interpreted to mean that people can use less than the full internal range but not more.

(Poulton, 1989), constrained scaling offers a method that calibrates the human mind to a particular scale. Initially, the central issue of constrained scaling was to what extent a scale could be learned (West & Ward, 1994). Early results revealed that with a modest amount of training to a loudness scale of 1000 Hz, participants were subsequently able to apply that scale to different tones. Note, however, that the object of constrained scaling was not that participants could specifically learn one loudness scale to measure a novel set of loudness stimuli, but rather that it would be possible to teach participants a general scale that could be applied to any domain. Later research confirmed the generalizability of the learned scale, which was applied successfully to evaluate the subjective utility of money (West & Ward, 1998) and the subjective brightness of light (West et al., 2000).

Recent research (West et al., 2000) has found that constrained scaling reduces the interparticipant variability in psychophysical response. To determine the average level of response variability, West et al. reviewed the results from 14 previous psychophysical studies that had used magnitude estimation or cross-modality matching methods (see Table 1). West et al. used the coefficient of variation, a measure of the percentage of variability relative to the mean, expressed in terms of the standard deviation divided by the mean (*SD*/*M*). The coefficient of variation provided a standardized measure of variability that was suitable for direct comparison across studies that had used different scaling ranges. The average coefficient of variation across the previous studies was 0.333. West et al. also used the ratio of highest-to-lowest (*H*:*L*) slope values from the scale plots as another gauge of variability. Within each study, the highest slope value

Table 1. Variability in a convenient sample of magnitude estimation and cross-modality matching experiments.

Study	***SD*/*M***	***H*:*L***	**Method**	***N***	**Stimulus**	**Study**
1	0.286	2.750	ME	11	loudness	Stevens & Guirao (1964)
2	0.436	n.a.	ME	32	loudness	Teghtsoonian & Teghtsoonian (1983)
3	0.388	n.a.	ME	35	loudness	Teghtsoonian & Teghtsoonian (1983)
4	0.290	3.951	ME	8	loudness	Algom & Marks (1990)
5	0.293	2.296	ME	11	loudness	Algom & Marks (1990)
6	0.444	3.320	ME	8	Loudness	Ward (1982)
7	0.446	6.000	ME	8	brightness	Ward (1982)
8	0.186	1.600	ME	10	loudness	Hellman & Meiselman (1988)
9	0.274	2.267	ME	6	heaviness	Luce & Mo (1965)
10	0.231	1.746	ME	6	loudness	Luce & Mo (1965)
11	0.328	3.368	CMM	20	duration to loudness	Lilienthal & Dawson (1976)
12	0.347	3.808	CMM	20	loudness to duration	Lilienthal & Dawson (1976)
13	0.392	2.435	CMM	5	loudness to line length	Zwislocki (1983)
14	0.326	2.4	CMM	10	duration to loudness	Ward (1975)

Definitions: *SD*/*M* = standard deviation of exponent values divided by the mean of the exponent values; *H*:*L* = the ratio of the highest to lowest individual exponent; *N* = number of participants; ME = magnitude estimation; CMM = cross-modality matching; n.a. = not applicable.

Note: Table from West, Ward, and Khosla (2000) used by permission of the first author.

was compared to the lowest slope value. Again, this relative proportion allowed a direct comparison between the different studies. The average of the highest-to-lowest values for the 12 studies for which it could be calculated was 2.995:1. In contrast, using constrained scaling, West et al. found a coefficient of variation equal to 0.086 and a highest-to-lowest ratio equal to 1.28:1 across trials for the first reported experiment. The use of constrained scaling reduced the interparticipant variability in psychophysical scaling nearly fourfold in terms of the coefficient of variation and by nearly two-and-a-half times in terms of the highest-to-lowest ratios. Similar results were found for the remainder of the constrained scaling experiments in the paper.

The results by West et al. (2000) suggested that constrained scaling significantly reduced the variability in scale usage compared with more conventional scaling methods. An analysis of the individual participant data further revealed that no participants exhibited the types of response biases described by Poulton (1989). Moreover, constrained scaling proved to be the method that produced the most consistent scaling responses by participants of any psychophysical method. As a measurement apparatus for human magnitude perception, constrained scaling exhibited a level of calibration that was not found in other scaling methods.

Constrained Scaling and Cognition

Constrained scaling is a cognitive model of scaling. Canonical approaches to psychophysical scaling suggest a stimulus-response approach to scaling (Marks, 1991), where:

$$M(S) = R, \tag{9}$$

in which *S* is the stimulus amplitude, *R* is the magnitude perception, and *M* is the psychophysical function that relates the two. Constrained scaling suggests that psychophysical scaling, *P*, is mediated by cognitive factors, *C*:

$$M(S) = C[P(S)] = R. \tag{10}$$

While the canonical approaches to psychophysics attempt to minimize the effects of *C*, West et al. (2000) sought to control for it and to calibrate it.

While *C* is a hypothetical construct, it represents the observable effects of cognition on scaling. West et al. (2000) made four key assumptions about the role of cognition in scaling:

1. *C* is cognitively penetrable in the sense that it can be controlled under the right conditions. The fundamental assumption of constrained scaling corresponds to this point, since constrained scaling assumes that through training, it is possible to influence an individual's cognitive matching of magnitude to a scale.
2. *C* is influenced by decisions made early during scaling. An individual considers factors like the stimulus intensity range and the appropriateness of a particular scale primarily during early exposure to scaling trials. It is therefore important that the learning trials that are part of constrained scaling be exercised early before response biases set in.
3. *C* makes heavy demands on the individual. Since the process of scaling makes heavy demands on attention and memory, the nature of the scale can actually result in an unstable scale. In practical terms, this assumption means that a constrained scale

needs to be a natural scale that can readily be matched to magnitude perceptions. A scale that does not map naturally to mental magnitudes would be too cognitively demanding to be effectively learned by the individual.

4. *C* is independent of perceptual modality. This assumption is important because it prescribes the possibility of a type of learned scale that can be used universally for all scaling needs, not simply for the stimulus modality in which the individual was first trained.

Again, it must be stressed that *C* is not a literal construct but rather a symbol of a variety of cognitive processes involved in translating a mental magnitude to a scale. Constrained scaling makes these four key assumptions about cognitive processes in scaling not as an exhaustive itemization of cognitive factors but as a starting point for the model.[19]

Constrained scaling is a cognitive model of scaling, and it continues to evolve. The experiments outlined in the next chapter aim to refine the constrained scaling model in order to make it a more comprehensive cognitive model. Among other things, I incorporate psychometrics into the constrained scaling model. Constrained scaling has

[19] These concepts are not unique to scaling. Any learned skill shows similar characteristics. For example, athletic skill requires (1) awareness of one's proprioceptive status and physical capabilities, (2) considerable training to achieve mastery, and (3) ongoing training to maintain mastery. Once mastery is achieved, (4) the athletic competence is often readily transferable to other domains involving the body. The process of psychological scaling is not as natural to most humans as is athletic skill, but athletic skill is attained over years if not decades of training, traceable to proprioceptive first steps in infanthood and childhood. Constrained scaling, like athleticism, capitalizes on humans' innate adaptability as the cornerstone of mastery. The domain of mastery in constrained scaling is the translation of mental magnitude to an external scale. Just as an athlete has highly honed physical skills, the constrained scaler exercises a condensed regimen of training to achieve calibration to a scale.

thus far been a model primarily of utility to psychophysical scaling. In the next chapter, I explain how I investigated the relevance of constrained scaling to subjective scaling that does not have a clear triggering stimulus in the non-mental world. Before I proceed, it is appropriate to review an example of constrained scaling in practice.

Reprise: Minds as Measurement Apparatuses

The importance of constrained scaling of magnitude is well illustrated in the rockumentary movie, *This is Spinal Tap* (Murphy & Reiner, 1984). In the film, Marty, a film producer, interviews Nigel, the lead guitarist for the heavy metal band, Spinal Tap, about his unusual electric guitar amplifier.

> NIGEL: ...This is the top to a…you know…what we use on stage. But, it's very…very special because if you can see…
>
> MARTY: Yeah....
>
> NIGEL: ...the numbers all go to eleven. Look...right across the board.
>
> MARTY: Ahh...oh, I see....
>
> NIGEL: Eleven...eleven...eleven....
>
> MARTY: ...and most of these amps go up to ten....
>
> NIGEL: Exactly!
>
> MARTY: Does that mean it's...louder? Is it any louder?
>
> NIGEL: Well, it's one louder, isn't it? It's not ten. You see, most...most blokes, you know, will be playing at ten. You're on ten here...all the way up...all the way up....
>
> MARTY: Yeah....

> NIGEL: ...all the way up. You're on ten on your guitar...where can you go from there? Where?
>
> MARTY: I don't know....
>
> NIGEL: Nowhere. Exactly. What we do is if we need that extra...push over the cliff...you know what we do?
>
> MARTY: Put it up to eleven.
>
> NIGEL: Eleven. Exactly. One louder.
>
> MARTY: Why don't you just make ten louder and make ten be the top... number...and make that a little louder?
>
> NIGEL: [long pause] ...these go to eleven.

Clearly, a volume setting of eleven is louder than ten. Or, is it? This classic comedic scene demonstrates the most important principle of constrained scaling: the endpoints of a scale have no direct relationship to magnitude unless they are calibrated. Constrained scaling is not about scaling physical stimulus intensities to a scale; rather, it is about scaling a person's subjective experience to a scale. One person's subjective scaling of loudness, for example, will vary considerably from another person's scaling of the same sound. However, if subjective experience is calibrated to a scale, what one person means when they say a volume of ten is subjectively the same as what another person means. Subjective experience of magnitude is simply calibrated to a scale in constrained scaling. Constrained scaling presents a scenario in which eleven really is louder than ten, at least if all guitar players are calibrated to the same scale. Nigel has impeccable logic in the world of constrained scaling!

OVERVIEW OF THE EXPERIMENTS

Replication, Refinement, Extension, and Application

As I discussed in the previous chapter, constrained scaling is both a tool for scaling and a cognitive model of how internal magnitude perceptions are translated into numeric estimations. In order better to understand the distinct tool and model components of constrained scaling, the present research addresses both facets through four types of experiments: *replication*, *refinement*, *extension*, and *application*.[20] Table 2 provides a rubric of how each proposed experiment falls within this classification. I briefly outline the 15 experiments in this section before providing a more detailed individual treatment of the experiments and results in subsequent chapters.

Experiment 1 is a *replication* experiment. Replication is simply a way to determine that the present experimental apparatus and design are compatible with existing research (Hubbard & Ryan, 2000). In this case, replication is crucial in determining that the current implementation of constrained scaling adheres to the tool outlined in earlier research. The ultimate goal of this replication is to ensure the robustness of the results from previous research to the present experimental framework. Experiment 1 is simply a replication of the method outlined in Experiment 1a and refined in Experiment 1b from West et al. (2000).   As shown in Table 2, both the training and testing stimuli involved loudness, as measured by the amplitude of tones. The training stimuli in Experiment 1 were 1000 Hz tones, while the testing stimuli were 65 Hz tones.

[20] Although I have classified each experiment according to a single type, these four classifiers need not be and, in fact, are not mutually exclusive.

Table 2. List of the experiments classified according to the contribution type of the experiment.

Experiment Number	Description	Training Stimuli	Testing Stimuli	Contribution
1	Basic loudness experiment replicating West et al. (2000)	Loudness (1000 Hz)	Loudness (65 Hz)	Replication
2	Basic loudness experiment as in Experiment 1, but without sound attenuating chamber	Loudness (1000 Hz)	Loudness (65 Hz)	Refinement
3	Cross-modal matching experiment: Loudness → Brightness	Loudness (1000 Hz)	Brightness (Grayscale)	Refinement
4	Cross-modal matching experiment: Brightness → Loudness	Brightness (Grayscale)	Loudness (1000 Hz)	Extension
5	Cross-modal matching experiment: Short-interval brightness → Loudness	Brightness (Grayscale)	Loudness (1000 Hz)	Extension
6	Color brightness magnitude estimation	—	Brightness (Grayscale + Color)	Extension
7	Color brightness experiment	Brightness (Grayscale)	Brightness (Color)	Extension
8	Brightness with categorical learning	Brightness (Grayscale)	Brightness (Color)	Refinement
9	Brightness with categorical learning plus random feedback noise	Brightness (Grayscale)	Brightness (Red)	Refinement
10	Brightness with decreased feedback ratio	Brightness (Grayscale)	Brightness (Red)	Refinement
11	Perceptual scale triangulation: Same stimuli trained to different brightness scales	1: Brightness 2: Brightness	1: Brightness 2: Brightness	Extension
12	Subjective triangulation magnitude estimation: Money → Happiness	—	Money	Extension
13	Subjective triangulation: Same stimuli trained to different brightness scales	1: Brightness 2: Brightness	1: Money 2: Money	Extension
14	Scaling of Web page visual appeal	Brightness (Grayscale)	Web pages	Application
15	Scaling of video quality of service	Video (Type 1)	Video (Types 2 + 3)	Application

Experiment 2 is a *refinement* experiment designed to assess constrained scaling as a tool. Refinement experiments aim to clarify methodology and reveal possible limits of constrained scaling. Experiment 2 repeats the stimuli and experimental apparatus used in Experiment 1. However, Experiment 2 was carried out in a normal room instead of a sound attenuating chamber. This refinement provides insight into the generalizability of loudness constrained scaling beyond laboratory facilities with precise acoustic abatement.

Likewise, Experiment 3 is a refinement experiment. In Experiment 3, the constrained brightness scaling experiment (Experiment 4) in West et al. (2000) was replicated and refined. Participants first learned a loudness scale and then applied it to scale brightness. The presentation of the brightness stimuli was refined and presented on a cathode ray tube (CRT) instead of displayed from a single light source light-emitting diode (LED) as had been done previously in West et al.

Experiments 8, 9, and 10 are also refinement experiments.[21] Experiment 8 investigates how well constrained scaling functions when the training stimuli are categorical (i.e., nominal or ordinal) instead of continuous (i.e., interval or ratio). Experiment 9 builds on Experiment 8, by adding random noise to the feedback values

[21] Experiments 8, 9, and 10 use a novel implementation of brightness stimuli instead of the loudness stimuli more common in previously published constrained scaling experiments. It is assumed that the methods and results would apply equally to a different stimulus modality such as loudness. The rationale for classifying these experiments as *refinement* experiments is because the outlined approach provides valuable insights into how to use constrained scaling as a tool. In other words, these experiments refine the constrained scaling method. They do not replicate previous research or extend constrained scaling theory, nor do they represent a specific practical application of constrained scaling.

accompanying the training stimuli. This random noise prevents rote memorization of the scale, a possible factor in learning categorical scale values. Finally, Experiment 10 decreases the feedback ratio to determine whether the customary high ratio of training trials is necessary to ensure scale learning.

While this quintet of refinement experiments may not immediately appear central to the course of theory building, these experiments reveal important theoretical insights into constrained scaling. For example, Experiment 2 makes the theoretical assumption that constrained scaling of loudness works in a variety of settings, even those without background noise reduction. If this is not the case, the current model of constrained scaling needs to be adjusted to restrict the training and testing environment. Experiments 8 and 9 deal with fundamental issues regarding the type of scale. Does constrained scaling work solely with continuous magnitude scales? Does it also apply to partition or poikilitic[22] measurement scales? If constrained scaling fails to work across scale types, this could serve as confirmation for the strict segregation of scale types that Stevens (1975) proposes.

Experiment 3 generalizes the type of light source that may be used in constrained brightness scaling experiments and sets the stage for subsequent scaling of brightness with a variety of color light sources. This basic refinement in the method first outlined in West et al. (2000) enables an array of experiments on color brightness scaling and further

[22] Poikilitic is used by Stevens (1975) to mean a continuous function. The term implies that the function includes a certain amount of variation or scatter, as is characteristic of human scaling on a continuous magnitude dimension.

simplifies the apparatus and instrumentation necessary for conducting constrained scaling experiments.

Experiment 10 reveals crucial information about the learnability of a scale. While there existed anecdotal evidence for how many training trials were necessary before a person learns a constrained scale, there had previously been no systematic investigation about the optimal number of training trials. Understanding the learnability of a scale is important not only for developing a streamlined method for administering constrained scaling as a tool but also for understanding the amount of cognitive effort required during scaling exercises. Constrained scaling rests on the premise that people can be trained to scale magnitude using a naturalistic scale. The naturalness of scales may vary, and a primary way to determine the naturalness of a particular scale is by the ease with which it is learned. Experiment 10 provides an indication of the learnability of a scale that conforms to Stevens' Power Law (Equation 3, p. 29). This learnability can serve as a baseline for future research on the naturalness of other scales.

Experiments 4 – 7 and 11 – 13 are *extension* experiments, in which the cognitive model of constrained scaling is further developed. Extension experiments aim to show that constrained scaling is not only a useful tool for perceptual research (as in Experiments 4 – 7) but also for the types of subjective scaling that underlie psychometric research (as in Experiments 11 – 13). While such an extension obviously adds to the versatility of constrained scaling as a tool, it also significantly advances the cognitive model. These experiments determine if there is a common cognitive scaling process that applies cross-modally to both perceptual and subjective mental phenomena.

Experiments 4 – 7 involved training and testing stimuli of various forms of brightness instead of the loudness scaling conventionally used in constrained scaling experiments. In Experiment 4, participants were trained on a basic brightness scale consisting of grayscale squares on the screen. In turn, they used the learned numeric scale from the brightness stimulus presentations to scale loudness stimuli. Experiment 5 presented a variation of the technique in Experiment 4, in which the duration of the brightness stimuli was shortened to more closely match the characteristics of the loudness stimuli. Experiment 6 served as a baseline measure for grayscale, red, green, and blue brightness scaling. Participants performed a standard magnitude estimation of the brightness of the monochrome and color stimuli, without a prior constrained scaling training session. In Experiment 7, participants were trained to use the basic grayscale brightness scale and subsequently scaled the brightness of red, green, and blue squares.

Experiments 11 – 13 are the theory centerpieces of this dissertation. These experiments compare different scaling modalities in a novel triangulation method. The aim of the experiments is to show that two different constrained scales function identically. Experiment 11 seeks to confirm that a common cognitive process is involved when using different perceptual scales. Participants learned a perceptual brightness scale and applied it to novel brightness stimuli. Participants then learned a second perceptual brightness scale and applied it to novel brightness stimuli. The ratio of the scaling slopes between the trained stimulus scale and the untrained stimulus scale was compared between the two experimental sessions. Experiment 13 is a variation of Experiment 11, instead using a learned scale to assess a subjective scaling dimension. Again, training was compared across two time periods using two different learned scales. Using the same participants in Experiments 11 and 13, it was demonstrated that learned scales were applied in the same manner for both perceptual (i.e., psychophysical) and subjective (i.e., psychometric) scaling modalities. Experiments 11 and 13 serve as an important bridge between the psychophysical and psychometric literature, in that they put forth constrained scaling as a valid unifying model to account for both perceptual and subjective mental magnitudes. Finally, an important cornerstone of this psychometric-psychophysical triangulation is Experiment 12, which provides a baseline for subjective scaling through magnitude estimation, which is compared to the scaling results obtained through constrained scaling in Experiment 13.

The final two experiments are *application* experiments, demonstrating the real-world utility of constrained scaling. Experiment 14 moves the constrained subjective scaling framework pioneered in Experiments 11 – 13 to an applied research domain. Experiment 14 applies a constrained subjective scale to the domain of affective computing. Affect has been found to be an important contributing factor in the use of computers. Within the field of affective computing, aesthetics represents a topic of growing research prominence (Norman, 2004). Since a standard scaling methodology has not been adopted for aesthetics research, there is an excellent and timely opportunity to test constrained scales vs. conventional scales for assessing aesthetics in computing. In the final experiment, Experiment 15, I looked at using constrained scaling as a tool for selecting the quality of service for streamed video broadcasting. In the face of current criticisms about the standard scales for assessing video quality (Watson & Sasse, 1998;

West, Boring, Dillon, & Bos, 2001), a critical comparison of constrained scaling to other video quality scaling methods is important. Together, Experiments 14 and 15 provided a real-word test case for constrained scaling as a tool as well as a glimpse of other possible practical uses for constrained scaling.

EXPERIMENT 1[23]

Introduction

The custom-built apparatus used in West et al. (2000) was not available for use in the present experiments. Consequently, a new apparatus was created to match as closely as possible the apparatus used in the earlier experiments. This experiment served as a replication of the general design and apparatus used in Experiments 1a and 1b in West et al. The intent of this replication was to ensure that the new apparatus produced comparable results to the earlier published results.

Method

Participants

Five university students with self-reported normal hearing volunteered as participants in the experiment. The volunteers were remunerated $10 for their participation.

Apparatus

The experimental control software is described in detail in Appendix A. The experiment was conducted in an Eckoustic sound attenuating chamber using a Windows 2000-based personal computer with a Creative Labs Sound Blaster Audigy card coupled to sealed circumaural headphones by Sennheiser. The software interface provided a customized scrollbar that allowed the user to select values between 0.0 and 99.9, with one

[23] A preliminary summary of the findings from Experiment 1 was presented at the Twelfth Annual Meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science, May 30 – June 1, 2002, in Vancouver, BC (West & Boring, 2002).

decimal point of accuracy. Using a mouse, participants selected the scale value by moving the slider on the scrollbar or by using on-screen buttons to increment or decrement the scale value by 0.1, 1.0, and 10.0. The scale value was displayed in a textbox above the scrollbar. To hear a tone, participants used the mouse to press a button marked PLAY TONE on the screen. For feedback trials, after participants selected their scale value, the scrollbar slider was automatically moved to the correct scale value position and the textbox displayed the correct scale value with five decimal points of accuracy. An on-screen button labeled NEXT allowed participants to advance to the next trial.

Stimulus Materials

The creation of loudness stimuli as well as the calibration of loudness levels is described in detail in Appendix B. The loudness stimuli consisted of 65 and 1000 Hz pure tones played for 1 second through the right earpiece of the headphones. The tones ranged in amplitude from 33 dB to 100 dB and were stepped at 1 dB intervals, with 6 ms ramp-up and ramp-down times.

Design and Procedure

As in Experiments 1a and 1b in West et al. (2000), the participants were first trained to estimate the magnitude of the 1000 Hz stimuli. The participants were presented with 50 training trials during which they heard 1000 Hz pure tones of varying amplitudes. The amplitude values were randomly selected for each trial in order to ensure participant exposure to a range of loudness stimuli. The participants were instructed to estimate how loud the tone was for each trial using the custom scrollbar. After the participants had selected a value, they were shown the correct response and instructed to make a mental note of the value. The correct response values were displayed with five decimal points of accuracy to encourage the participants to make full use of the numerical precision available to them. The five decimal points of accuracy also discouraged rote memorization of the scale. The correct scale values for the 1000 Hz tone amplitudes were calculated according to the following equation:

$$R = 16.6P^{0.60}, \tag{11}$$

where *R* was the correct scale response and *P* was the amplitude as measured in dynes/cm^2. As in West et al. (2000), the exponent of the correct response was set to 0.60 in order to conform to Stevens' sone scale (1975). Using a multiplier of 16.6 allowed the scale to range between 1 and 94, a near full utilization of the 100-point scale available to the participants.

Figure 5 depicts the design of Experiment 1. Participants were initially trained on 50 iterations of randomly selected loudness values of the 1000 Hz tone. In each training trial, they received feedback about the correct loudness value. Next, the participants received a testing block of 100 trials in which they rated the loudness of the 1000 Hz tones with feedback and the 65 Hz tones without feedback. Following a short break to prevent fatigue, participants received another block of 50 training trials in which they were presented a 1000 Hz tone and asked to scale it. In these training trials, the participants received feedback about the correct response. After the block of training trials, the participants again received a testing block of 100 trials. The second testing

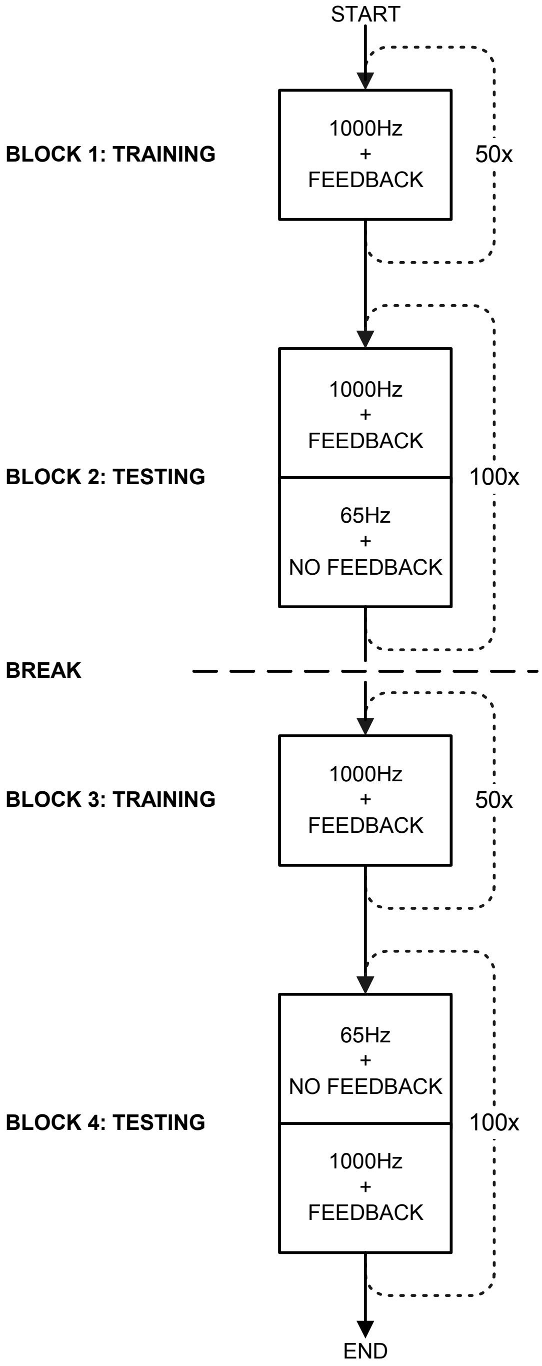


Figure 5. Schematic flow of the constrained scaling experiment in Experiment 1.

block was identical to the first one with the exception that the order of presentation for feedback and no feedback trails was counterbalanced.

Results and Discussion

As in West et al. (2000), the logarithm of the participant response (*R*) was regressed against the logarithm of the sound pressure (*P*) in dynes/cm^2 according to the following equation (Equation 5 from West et al.):

$$\log R = m \log P + \log a + e, \qquad (12)$$

where *m* represents the slope of the resulting line, *a* represents the y-axis intercept, and *e* represents residual error. The residual error term *e* may be removed to produce the general form of the equation:

$$\log R = m \log P + \log a. \qquad (13)$$

Transforming Equation 21 from a logarithmic scale to a standard scale produces the familiar Power Law form:

$$R = aP^m. \qquad (14)$$

The results of the present experiment are summarized in Table 3. The graphs for each participant according to Equation 14 are found in Figure 6 for 1000 Hz tones with feedback and in Figure 7 for 65 Hz tones without feedback. Those values that were more than two standard deviations from the regression line were discarded as outliers. Outliers represented 1.16% of the data and were typically associated with response values during initial trials or for low amplitude stimuli. It was assumed that during initial trials, participants had not yet developed a correspondence between the stimuli and the

Table 3. Summary of participants' loudness scaling for 65 and 1000 Hz tones in Experiment 1.

	1000 Hz + Feedback			**65 Hz + No Feedback**			**Ratio of 1000:65**		
P	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***
1	0.511	1.295	0.754	0.607	1.564	0.776	0.841	0.828	0.972
2	0.513	1.193	0.822	0.614	1.472	0.815	0.836	0.810	1.009
3	0.451	1.423	0.658	0.563	1.588	0.691	0.800	0.896	0.954
4	0.516	1.219	0.757	0.591	1.461	0.774	0.873	0.835	0.978
5	0.514	1.182	0.823	0.605	1.476	0.796	0.849	0.801	1.033
M	0.501	1.262	0.763	0.596	1.512	0.770	0.840	0.834	0.989
SD	0.028	0.100	0.067	0.020	0.059	0.048	0.026	0.037	0.032

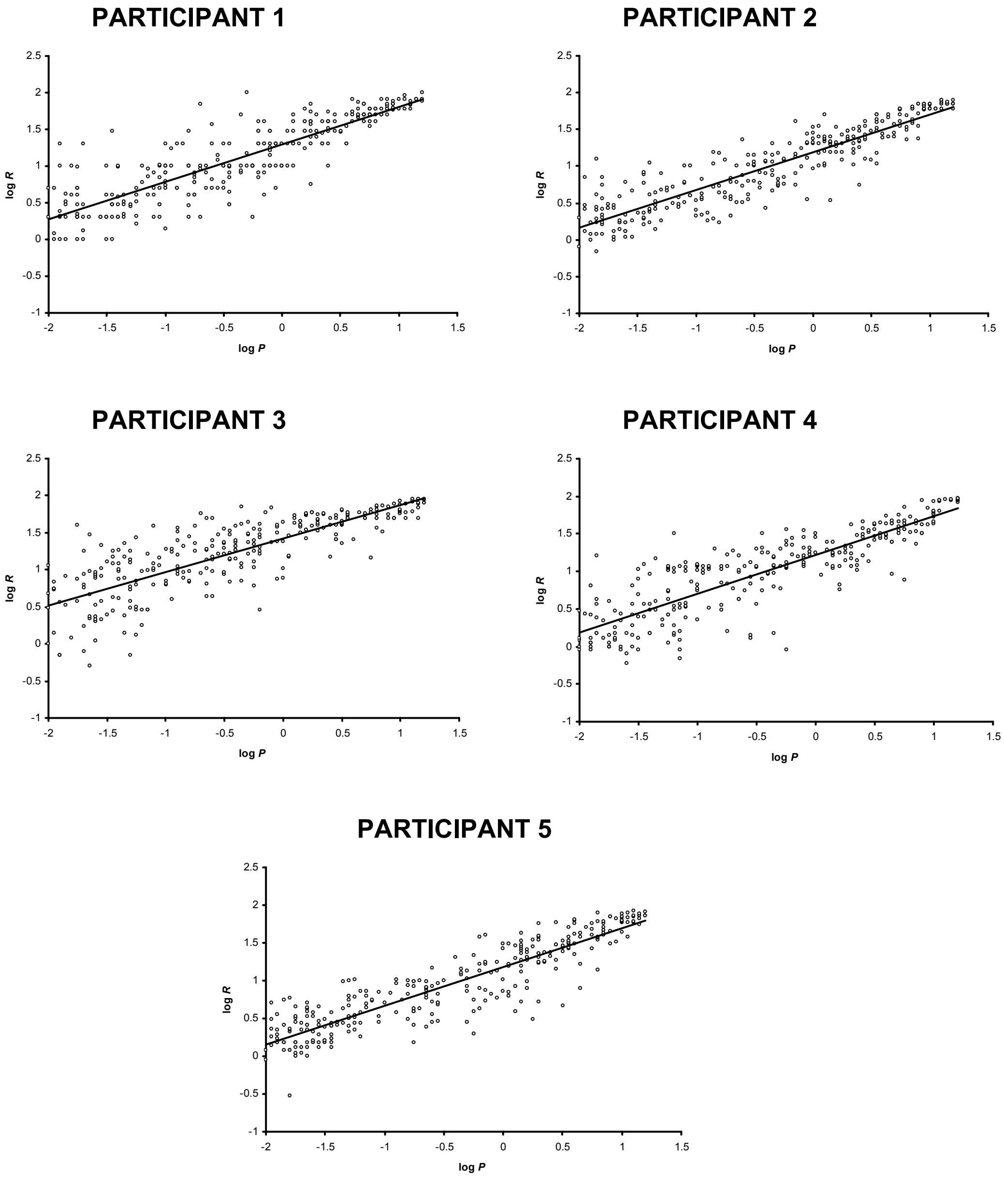
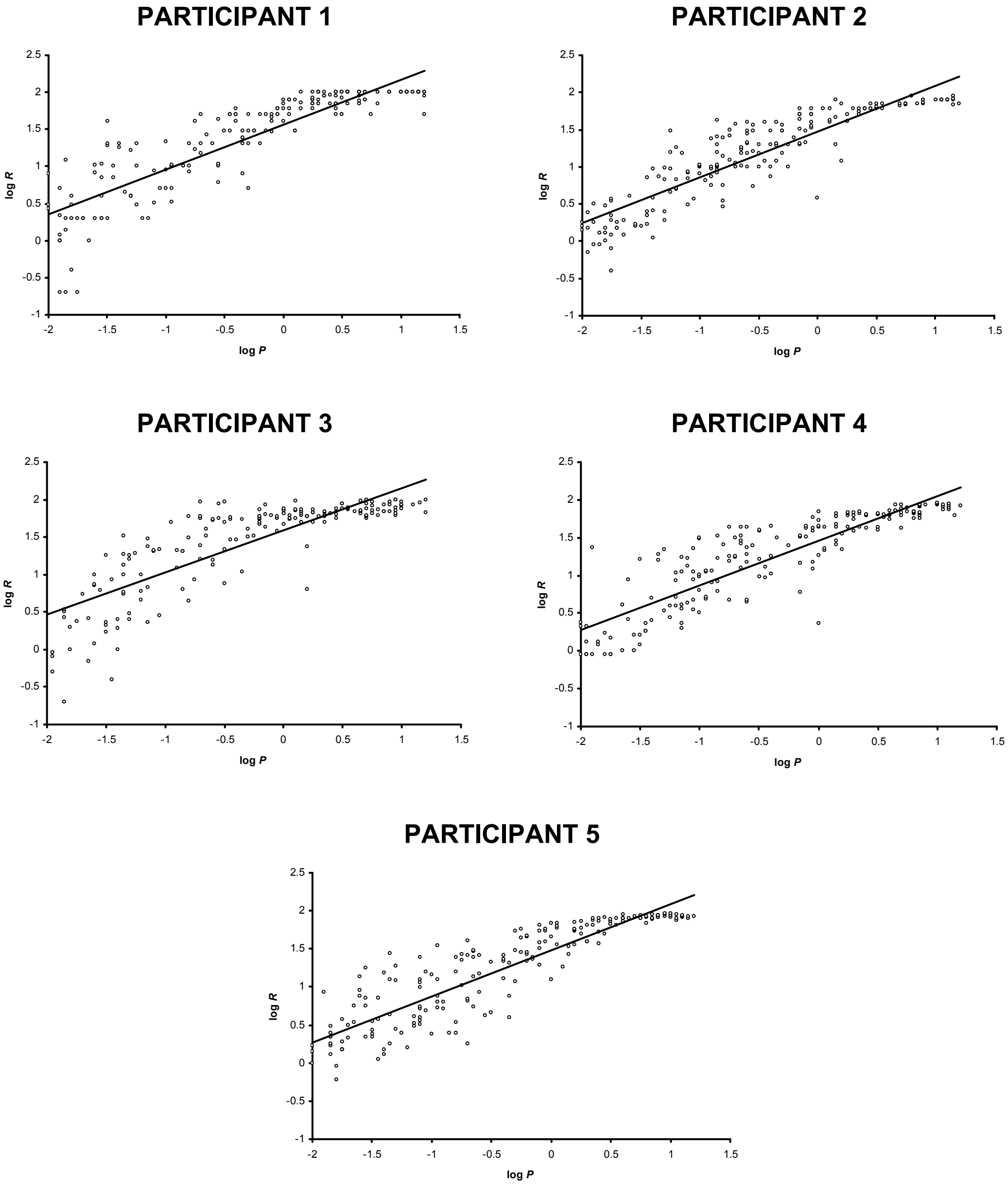


Figure 6. Logarithmic scatterplot and regression line for loudness in dynes/cm^2 (P) and participant response (R) for 1000 Hz tones with feedback in Experiment 1.


PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 7. *Logarithmic scatterplot and regression line for loudness in dynes/cm^2 (P) and participant response (R) for 65 Hz tones without feedback in Experiment 1.*

response scale, thus resulting in aberrant response values. For low amplitude stimuli, it was assumed that the stimuli were near or below the hearing threshold of the participants, resulting in a cluster of values at the response scale nadir.

Exponent Values

For the 1000 Hz tones with feedback, the mean exponent value, *m*, was 0.501, the coefficient of variation (standard deviation divided by the mean, or *SD*/*M*) was 0.056, and the highest-to-lowest (*H*:*L*) exponent ratio was 1.145:1. These results closely replicated the results from Experiments 1a and 1b in West et al. (2000). For 1000 Hz tones with feedback in Experiment 1a, the participants in West et al. had a mean exponent value equal to 0.56, an *SD*/*M* equal to 0.075, and an *H*:*L* ratio equal to 1.23:1.[24] For Experiment 1b, the mean exponent value was equal to 0.56, the *SD*/*M* was equal to 0.112, and the *H*:*L* ratio was equal to 1.35:1. The present experiment revealed slightly lower exponent values but also lower *SD*/*M* and *H*:*L* ratios, suggesting lower variability comparable to that found in West et al.

For the 65 Hz tones without feedback, the present experiment revealed a mean exponent value of 0.596 with an *SD*/*M* equal to 0.034 and an *H*:*L* ratio equal to 1.090:1. For the participants in West et al. (2000), in Experiment 1a the mean exponent value was

[24] West et al. (2000) used two significant digits for the mean and *H*:*L* and three significant digits for *SD*/*M*. Where values are explicitly stated in West et al., the number of significant digits has been transferred as in the source. Where the values are not directly provided in West et al. but have been calculated using available information, I have used three significant digits.

0.77, the *SD*/*M* was 0.132, and the *H*:*L* ratio was 1.33:1; in Experiment 1b, the mean exponent value was 0.67, the *SD*/*M* was 0.096, and the *H*:*L* was 1.37:1. The present experiment's results for 65 Hz tones without feedback closely matched the results from the earlier experiment for 65 Hz tones without feedback. As in West et al., there was a slight increase in the exponent value for lower pitched tones without feedback.

It should be noted that curves for the 65 Hz tones without feedback exhibit slight nonlinearity, primarily at the upper end of the stimulus range. This slight curvature at the endpoints of the data is a deviation from Stevens' Power Law, which may be accounted for by Norwich (1993) as noted earlier in Equations 5 and 6. Despite this slight deviation from linearity, the linearly fitted regression line serves as a useful approximation of the data, since this measure allows comparison to earlier data sets. The nonlinearity is also accounted for in the R^2 values and noted in terms of scaling reliability.

Intercept Values

The average *y*-axis intercept value, *a*, is not reported in West et al. (2000). In fact, the line intercept is rarely reported in psychophysical studies.[25] The intercept is, nonetheless, important, because constrained scaling should effectively reduce the variability of the line intercept across participants, just as it reduces the variability of the slope values.

[25] Stevens' (1975) seminal work on the Power Law makes only cursory mention of the intercept and fails to elaborate on its function.

The intercept values reported in Table 3 represent the values for log *a*.[26] For the 1000 Hz tones with feedback, the mean logarithmic intercept value, log *a*, was 1.262, with *SD/M* = 0.079 and *H*:*L* = 1.204:1. For the 65 Hz tones without feedback, the mean logarithmic intercept value was 1.512, with *SD/M* = 0.039 and *H*:*L* = 1:087:1. Taking the anti-log of these values suggests that the mean intercept value, *a*, was 18.281 for 1000 Hz tones and 32.509 for 65 Hz tones. These average response values were higher than the 16.6 multiplier used for training in Equation 11, particularly for the 65 Hz tones without feedback.[27] As expected from equal loudness contours, this value suggests that 65 Hz tones may require adjustment of the intercept at lower amplitudes to match the learned scale for 1000 Hz tones. The slope or exponent values indicate that despite the perceptual difference in loudness between the two sets of stimuli, the participants used the response scale consistently across both stimulus ranges. The findings suggest that the exponent is the more fundamental measure, indicating that participants learned the relationship between stimuli in terms of the exponent and used the intercept to map their perception to a number scale.

Goodness of Fit Coefficients

The goodness-of-fit coefficient, R^2, measures the degree to which the data conform to the regression line. Technically, R^2 is the proportion of variance the response

[26] To arrive at the actual intercept value, *a*, it is necessary to take the anti-logarithm of the value provided in Table 3.

[27] Recall that the logarithmic *y*-axis intercept is equivalent to the multiplier in the Power Law.

scores, *R*, share with the stimulus scores, *P* (Cohen & Cohen, 1983). The higher the degree of shared variance is, the higher the conformity of data points to the regression line is.

For 1000 Hz tones with feedback, the mean R^2 value was 0.763, *SD*/*M* = 0.088, *H*:*L* = 1.145:1. In Experiment 1a in West et al. (2000), the mean R^2 value for 1000 Hz tones was 0.857, *SD*/*M* = 0.031, *H*:*L* = 1.072:1. In Experiment 1b, the mean R^2 value for 1000 Hz tones was 0.864, *SD*/*M* = 0.083, *H*:*L* = 1.274:1. Overall, the goodness of fit was slightly lower in the present experiment than in Experiments 1a and 1b in West et al.

For 65 Hz tones without feedback, the mean R^2 value was 0.770, *SD*/*M* = 0.062, *H*:*L* = 1.181:1. These results are similar to the results found by West et al. (2000) in Experiments 1a and 1b. In Experiment 1a, the mean R^2 value for 65 Hz tones was 0.847, *SD*/*M* = 0.061, *H*:*L* = 1.200:1. In Experiment 1b, the mean R^2 value for 1000 Hz tones was 0.791, *SD*/*M* = 0.103, *H*:*L* = 1.319:1. As with 1000 Hz tones, the goodness of fit in the present experiment was slightly lower for 65 Hz tones than that found in West et al., but the level of variability was comparable.

Ratio Values

As in West et al. (2000), there is a ratio comparison between the values for the 1000 Hz tones with feedback and the 65 Hz tones without feedback. This ratio measures the relationship between the responses on the training and test stimuli. As discussed in West et al., this ratio—when comparing scaling exponents—approximates magnitude matching of two scaling continua. Assuming an individual's inherent scaling bias is retained across training and test responses and that this scaling bias is largely constant within a scaling modality (i.e., loudness in this case), the resultant ratio reflects the true perceptual ratio of loudness exponents for 65 and 1000 Hz tones.

In West et al. (2000), the mean ratio of 1000 Hz to 65 Hz exponents in Experiment 1a was 0.79, *SD*/*M* = 0.086, *H*:*L* = 1.28:1. In Experiment 1b, the equivalent mean ratio was 0.84, *SD*/*M* = 0.082, *H*:*L* = 1.19:1. The findings from the present experiment replicate these earlier experiments. The mean ratio of 1000 Hz to 65 Hz exponents was 0.840, *SD*/*M* = 0.031, *H*:*L* = 1.091:1.

This result is compatible with differing hearing sensitivity to different frequencies. Equal loudness contours (Stevens, 1975; Ward, 1990) demonstrate that the listener's sensitivity to 65 Hz tones is less thanhis or her sensitivity to 1000 Hz tones at low dB levels. As the dB level increases, the listener's hearing sensitivity to 65 and 1000 Hz tones equalizes. Thus, perception for 65 Hz tones across the normal hearing loudness range results in a steeper rise and consequent greater slope or exponent than hearing perception for 1000 Hz tones. As in West et al. (2000), the results from the present experiment confirm the expected larger exponent value for 65 Hz tones compared to 1000 Hz tones.

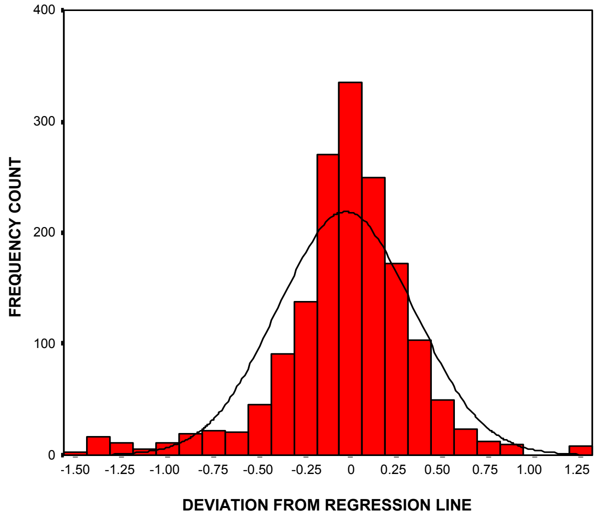
West et al. do not provide ratios for the mean *y*-axis intercept or the goodness-of-fit coefficient. The equivalent ratios are included in Table 3 for the present experiment. It is assumed that the intercept ratios provide a comparison of threshold sensitivity to 1000 Hz and 65 Hz tones. The mean intercept ratio of 1000 Hz to 65 Hz tones equals 0.834, with *SD*/*M* = 0.045 and *H*:*L* = 1.119:1. These results suggest that the participants exhibited less threshold sensitivity to 1000 Hz tones than to 65 Hz tones. The goodness-of-fit ratios provide a measure of how well the data conform to the regression line for the training stimuli vs. the testing stimuli. The mean goodness-of-fit ratio of 1000 Hz to 65 Hz tones equals 0.989, with *SD*/*M* = 0.032 and *H*:*L* = 1.084:1. These results suggest that there was very little difference between the training stimuli (i.e., the 1000 Hz tones) and the testing stimuli (i.e., the 65 Hz tones).

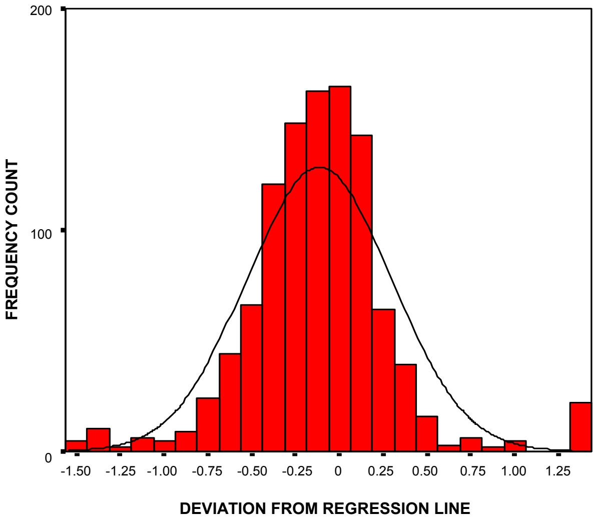
Score Distributions

Figure 8 presents histograms for the frequency distributions of participant scores about the averaged regression line for 1000 Hz and 65 Hz tones. As demonstrated, the scores follow approximately normally shaped distributions about the regression line, suggesting that participants overall did not display a particular score bias in scaling their perception of the stimuli. Both histograms exhibit high kurtosis with heavy tails. For the 65 Hz tones, there is a slight skew on the right tail, which is indicative of the apparent curvature in the scatterplot for high stimulus scores. This ceiling effect in scale values is indicative of an upper bound in perceived loudness for the 65 Hz tones.

General Discussion

The results from Experiment 1 compare favorably to the earlier results reported in West et al. (2000). Using constrained scaling, participants exhibited good mastery of the training scale and good application of that scale to a novel set of stimuli. This was evidenced by the interparticipant variability rates that were comparably low to earlier findings in West et al. Additionally, the ratios of 1000 Hz tones to 65 Hz tones were generally low, further demonstrating the participants' ability to use the learned 1000





NOTE: Frequency counts differ because there were more presentations of the training stimuli (1000 Hz tones) than the testing stimuli (65 Hz tones). The line through the histograph depicts the normal distribution.

Figure 8. Frequency distributions for scores about the average regression line for 1000 Hz stimuli (top) and 65 Hz stimuli (bottom) in Experiment 1.

Hz scale for reporting their perception of 65 Hz tones. Experiment 1 was a successful replication of Experiments 1a and 1b in West et al.

Also, importantly, Experiment 1 was a successful replication using the new experimental apparatus. The new apparatus featured a standard personal computer configuration with a sound card. This experiment marked an important transition for constrained scaling from specialty psychoacoustic equipment to a readily available and easily deployable experimental apparatus.

EXPERIMENT 2

Introduction

One encumbrance to the utility of constrained scaling is the difficulty with which the experimental apparatus is replicated. Especially loudness experiments can be difficult to replicate, given the need to control the acoustic environment in which the experiment is conducted. Background noise abatement is typically accomplished in sound attenuating chambers. Such chambers are not typically portable, requiring a designated psychoacoustic laboratory to house the equipment.

In an effort to overcome the need for a psychoacoustic laboratory in constrained scaling research, the present experiment sought to replicate the method and findings from Experiment 1 outside a sound attenuating chamber. The equipment used to conduct Experiment 1 was moved outside the sound attenuating chamber into a conventional psychological laboratory. This laboratory featured acoustic properties similar to a contemporary office, with ambient noise in the 40 – 45 dB range. To minimize confounding factors such as the variability in ambient noise associated with background conversation, no people other than the participant were present in the laboratory during the experiment.

The goal of this experiment was to establish the applicability of the constrained scaling of loudness is a minimally controlled environment. A cornerstone of constrained scaling's utility as a general purpose scaling method is the ability of researchers to use the method in research settings that do not feature the tight controls exercised in Experiment 1. If it is possible to use constrained scaling of loudness outside a sound attenuating chamber, there is considerable promise that constrained scaling may be employed in a wider array of experiments than those published to date. In this context, Experiment 2 represents a clear refinement of the existing constrained scaling methology.

Method

Participants

Five university students with self-reported normal hearing volunteered for the experiment and were remunerated $10 for their participation.

Apparatus

The apparatus was identical to the apparatus used in Experiment 1, with the exception that the experiment was not carried out in a sound attenuating chamber. The experiment was conducted in a psychological laboratory with ambient noise in the 40 – 45 dB range. The Sennheiser sealed circumaural headphones offered approximately 15 dB in sound isolation.[28]

The loudness of sounds over the headphones was calibrated in the sound isolating container described in Appendix B. The use of the sound isolating container for the headphones was necessary in order to identify the loudness of tones without the additive effect of the background noise level. At the low dB range of sounds, it was otherwise impossible for the dB meter to isolate the amplitude of the 65 and 1000 Hz pure tones amid louder ambient noise. To the human ear, however, these tones were discernable

[28] The level of sound isolation varied depending on the frequency of the background noise. The level of sound isolation is averaged to be 15 dB, based on average performance across a variety of background noise situations.

from the ambient noise, except as these tones approached the threshold of human hearing.

Design and Procedure

The design and procedure of Experiment 2 were identical to those in Experiment 1. Participants were trained on the loudness scale in Equation 11, with a multiplier of 16.6 and an exponent of 0.60 for 1000 Hz pure tones. The participants subsequently applied the learned scale to novel 60 Hz pure tones of varying amplitudes.

Results and Discussion

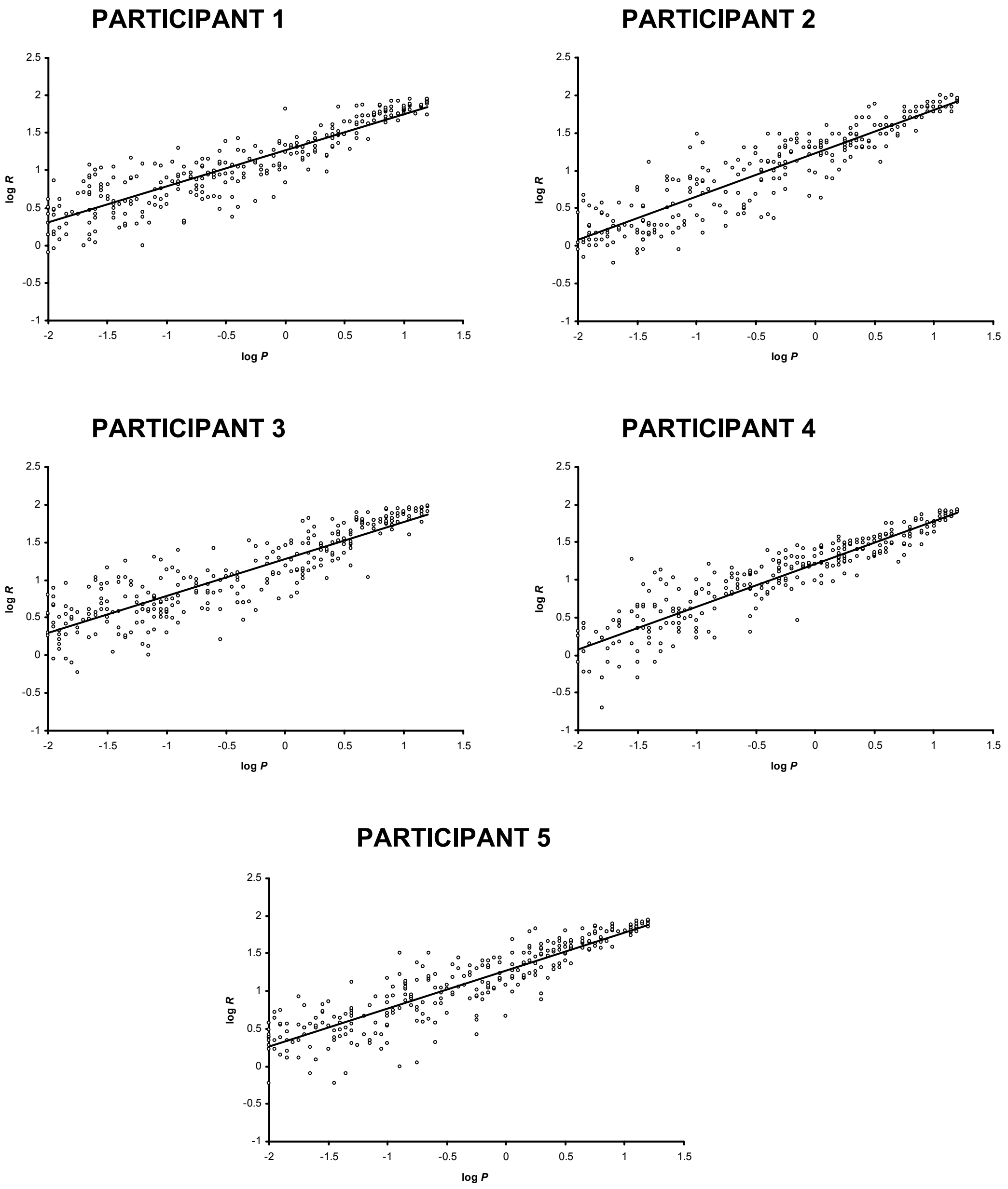
The data were analyzed as in Experiment 1 and are presented in Table 4 and Figures 9 – 11. Those values that were more than two standard deviations from the regression line were considered outliers and were discarded. In the present experiment, 2.36% of response values were discarded as outliers. This outlier rate was slightly more than double the outlier rate for Experiment 1. Analysis of the specific outliers revealed that the majority were at the bottom of the loudness stimulus presentation range. These low-amplitude sounds typically received a loudness rating of "0.0," implying that participants were unable to here the tones above the background noise. Figure 11 presents a comparison of the average regression lines across participants for Experiments 1 and 2.

Exponent Values

For 1000 Hz tones with feedback, the average exponent value, *m*, equaled 0.524, with $SD/M = 0.084$ and $H{:}L = 1.193{:}1$. For 65 Hz tones without feedback, the average

Table 4. Summary of participants' loudness scaling for 65 and 1000 Hz tones in Experiment 2.

	1000 Hz + Feedback			**65 Hz + No Feedback**			**Ratio of 1000:65**		
P	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***
1	0.480	1.268	0.820	0.591	1.347	0.837	0.812	0.942	0.980
2	0.573	1.229	0.849	0.626	1.374	0.855	0.914	0.895	0.994
3	0.490	1.277	0.775	0.598	1.303	0.760	0.819	0.980	1.020
4	0.569	1.214	0.833	0.629	1.333	0.705	0.905	0.911	1.182
5	0.506	1.274	0.809	0.638	1.313	0.720	0.794	0.970	1.123
M	0.524	1.252	0.817	0.616	1.334	0.775	0.849	0.940	1.060
SD	0.044	0.029	0.028	0.020	0.028	0.068	0.056	0.037	0.088



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 9. Logarithmic scatterplot and regression line for loudness in dynes/cm^2 (P) and participant response (R) for 1000 Hz tones with feedback in Experiment 2.

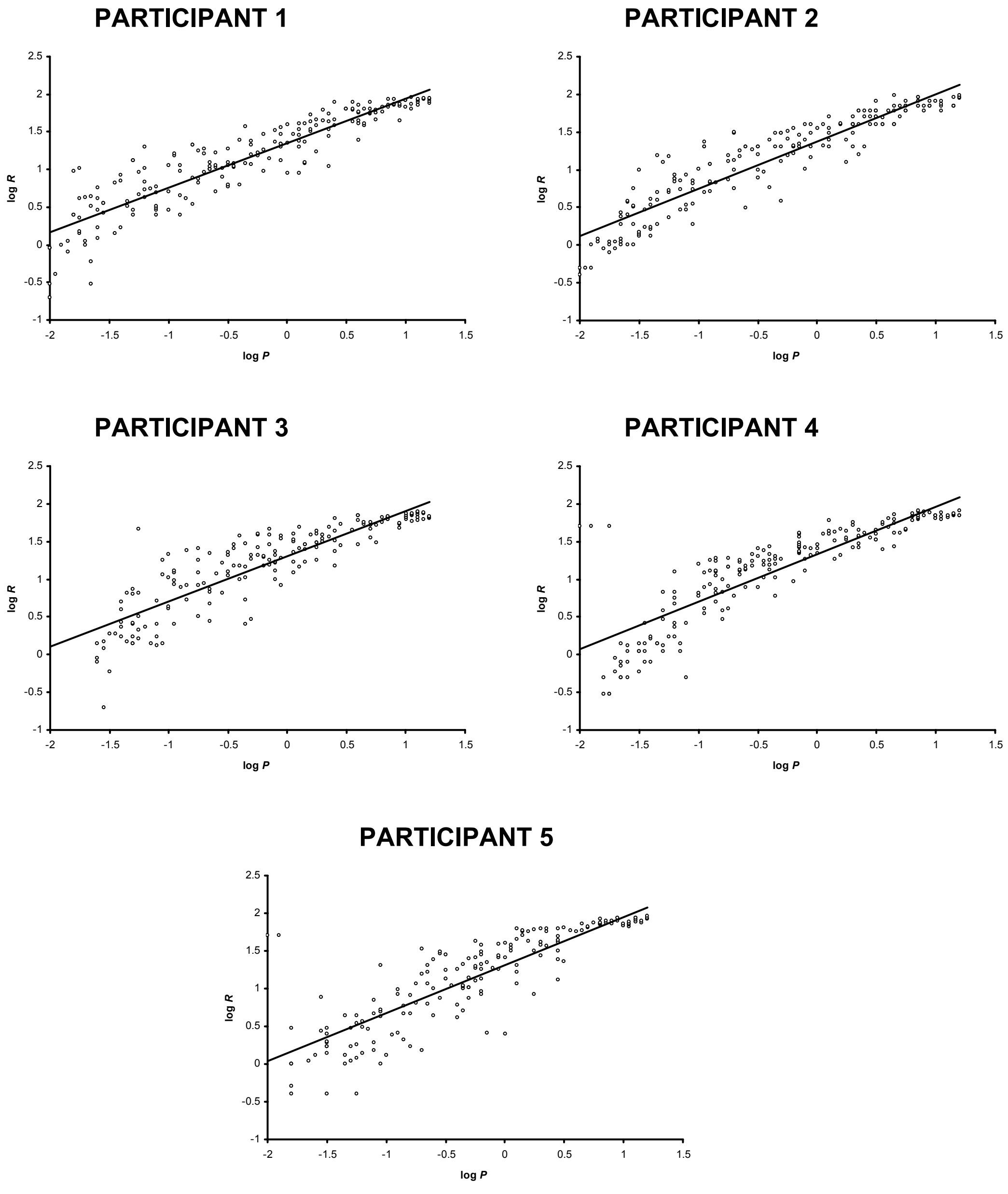


Figure 10. Logarithmic scatterplot and regression line for loudness in dynes/cm^2 (P) and participant response (R) for 65 Hz tones without feedback in Experiment 2.

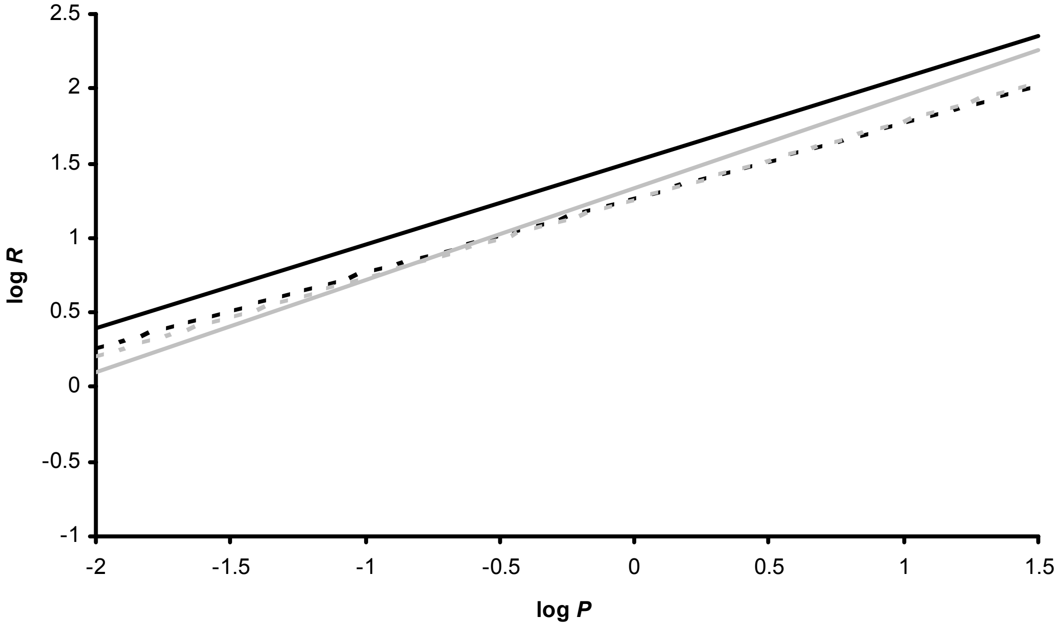


Figure 11. A comparison of 65 Hz (solid line) and 1000 Hz (dotted line) tones for Experiment 1 (black line) and Experiment 2 (grey line).

exponent value equaled 0.616, *SD*/*M* = 0.033 and *H*:*L* = 1.078:1. These exponent values compared favorably to the values obtained in Experiment 1, with a slight, two-tenths of a decimal point increase in the average exponent value in the present experiment. The average ratio of 1000 to 65 Hz tones was nearly identical across experiments (0.840 for Experiment 1 vs. 0.849 for Experiment 2).

Intercept Values

For 1000 Hz tones with feedback, the average intercept value, *a*, was 1.252, with *SD*/*M* = 0.023 and *H*:*L* = 1.051:1. For 65 Hz tones without feedback, the average intercept value was 1.334, with *SD*/*M* = 0.041 and *H*:*L* = 1.109:1. These values compared favorably to the values in Experiment 1. For 1000 Hz tones with feedback, the average intercept was one-tenth of a decimal point lower in Experiment 2 than in Experiment 1. For 65 Hz tones without feedback, the average intercept was a little more than two-tenths of a decimal point lower in Experiment 2 than in Experiment 1. The average intercept ratio of 1000 to 65 Hz tones was higher in Experiment 2 than in Experiment 1 (0.939 vs. 0.834).

These slight differences in intercept values between Experiment 1 and Experiment 2 are attributed to the background noise present in Experiment 2. The intercept is a reflection of the perceived loudness of tones for low amplitude sounds. The background noise in Experiment 2 exhibited minimal interference for 1000 Hz tones, resulting in a close match in average intercepts between the two experiments. However, the background noise interfered with the perceived loudness of the 65 Hz tones. The background noise, characterized primarily by the low-pitched whir of a computer cooling fan, served to mask the perceived loudness of 65 Hz tones.

The intercept differences are most clearly seen in Figure 11. The average regression lines are nearly parallel, suggesting that the exponent (i.e., line slope) values were comparable across the experiments. In contrast, the intercepts were quite different across conditions. For 1000 Hz tones (see the dotted lines in Figure 11), the average regression lines aligned closely across both experiments, suggesting that the exponent and intercept values were nearly identical. In contrast, the average regression lines for 65 Hz tones (see the solid lines) varied considerably between Experiment 1 (see the black line) and Experiment 2 (see the grey line). The average regression line for the 65 Hz tones in Experiment 2 (see the solid grey line) deviated slightly from the general parallelism of the regression lines. The higher amplitude tones were perceived with loudness comparable to the loudness perception found in Experiment 1. As expected, this line demonstrates that the masking of loudness for the 65 Hz tones was more prevalent for lower amplitude sounds than for higher amplitude sounds. Once the tones exceeded a certain amplitude, the effect of the background noise was diminished.

Goodness of Fit Coefficients

The average goodness of fit coefficient, R^2, was 0.817 for 1000 Hz tones, with SD/M = 0.034 and H:L = 1.096:1. For 65 Hz tones, the average R^2 value was 0.775, with SD/M = 0.087 and H:L = 1.212:1. The average R^2 ratio of 1000 to 65 Hz tones was 1.060. These values were nominally higher than the values reported in Experiment 1, and it may be concluded that the goodness of fit values in Experiment 2 closely mirrored those values in Experiment 1.

General Discussion

The results in Experiment 2 closely replicated the results in Experiment 1, with one exception. Background noise may mask the experimental stimuli when it overlaps the frequency of the experimental stimuli, especially for lower magnitude stimuli. With this caveat in mind, the constrained loudness scaling method produces robust results, even outside a sound attenuating chamber. This experiment clearly demonstrates the extension and generalizability of the constrained scaling method beyond the rigid controls of a psychoacoustic laboratory. It is possible for an experimenter to use constrained scaling methods to measure loudness perception even in a conventional laboratory setting without noise abatement.

EXPERIMENT 3

Introduction

Initial results are very promising in terms of constrained scaling as a technique to reduce scaling variability between individuals. However, as a technique, constrained scaling is still relatively new. Most research involving constrained scaling has so far centered on loudness scaling, one of the most frequently investigated and best understood areas of psychophysics (Stevens, 1975). While loudness is an excellent starting point for a model of psychological scaling, it is exactly that–a starting point. This experiment extends previous research cross-modally by addressing constrained scaling as a measurement tool for the visual sensory modality, namely the scaling of brightness.

West et al. (2000) briefly touched upon the scaling of brightness. In Experiment 4 of their paper, participants were trained to scale loudness according to the familiar Equation 11 in this dissertation, $R = 16.6P^{0.60}$, where R was the response feedback and P was the sound pressure in dynes/cm^2. After the participants had learned the loudness scale, they were instructed to use the learned scale to rate the brightness of a green, 565 nm wavelength LED ranging from 0.044 to 272.024 cd/m^2 at six luminous intensity levels.[29] The mean exponent value was 0.33, which corresponds to Stevens' (1975) natural scaling exponent for brightness light sources in the dark.[30] The measures of

[29] West et al. (2000) presented their values in footlamberts (fL). One fL is equivalent to 3.426 cd/m^2.

[30] Stevens' (1966, 1975) scaling exponents for brightness varied from 0.33 to 1.2, depending on the area of brightness source, the brightness contrast between the light source and the background, and the duration of the light source presentation.

variability were slightly higher than the results typically obtained for loudness measures in other constrained scaling experiments, but they were generally lower than the variability found in magnitude estimation or cross-modality matching experiments. The coefficient of variation, *SD*/*M*, equaled 0.152 and the highest-to-lowest exponent ration, *H*:*L*, equaled 1.59:1.

West et al. (2000) used a calibrated LED to generate the brightness stimuli. The LED afforded a limited number of selectable luminance values on a logarithmic scale, potentially confounding the experiment by introducing a categorical stimulus scale. To achieve a continuous brightness scale akin to the continuous loudness scale, the present experiment used a cathode ray tube (CRT) display for stimulus presentation. The CRT display offered a wider gamut of color and luminance than is possible with a single LED as well as a wider display area than the point light source of a LED. Grayscale, the level at which the CRT's three phosphor guns (representing red, green, and blue colors) fire at the same luminous intensity, was selected instead of color for training so that the stimuli might be neutral with respect to color sensitivity in participants.

Method

Participants

Five university students with self-reported normal hearing and normal color vision were enlisted as participants for the experiment. The participants did not overlap with the participants from the previous experiments. As in previous experiments, each volunteer received $10 for his or her participation in this experiment.

Apparatus

The experimental control software from Experiments 1 and 2 was modified to display the brightness stimuli in addition to the loudness stimuli. The loudness stimuli were identical to the 1000 Hz pure tones used in Experiments 1 and 2. The brightness training stimuli consisted of achromatic squares of 4º of visual field displayed on the screen directly in front of the participant.

The brightness stimuli used in the present experiment were displayed on a Samsung 19-inch SyncMaster 950P CRT display attached to an ATI Radeon VE graphics display adapter at a screen refresh rate of 100 Hz. The CRT was pre-warmed for one hour to stabilize the color luminance levels. The graphic display adapter was configured to 24-bit color resolution, corresponding to 8 bits (256 levels) for each of the red, green, and blue color channels. The CRT display was calibrated using a Spyder colorimeter puck to a black point luminance at 0.00 cd/m2 and a white point luminance at 95.0 cd/m2. The brightness stimuli used in the experiments consisted of 15 squares ranging in luminance from 0 to 100 cd/m^2 along an equal log-spaced luminance continuum. The grayscale squares were calibrated to be color neutral in accordance with CIE color standards (Commission Internationale de L'Eclairage, 1931), with average CIE Yxy chromaticity coordinates of x=0.278 and y=0.293. Appendix C provides the full details of the apparatus used in Experiment 3 as well as the calibration technique used for the brightness stimuli.

Design and Procedure

Figure 12 depicts the design of Experiment 3. The design and procedure closely

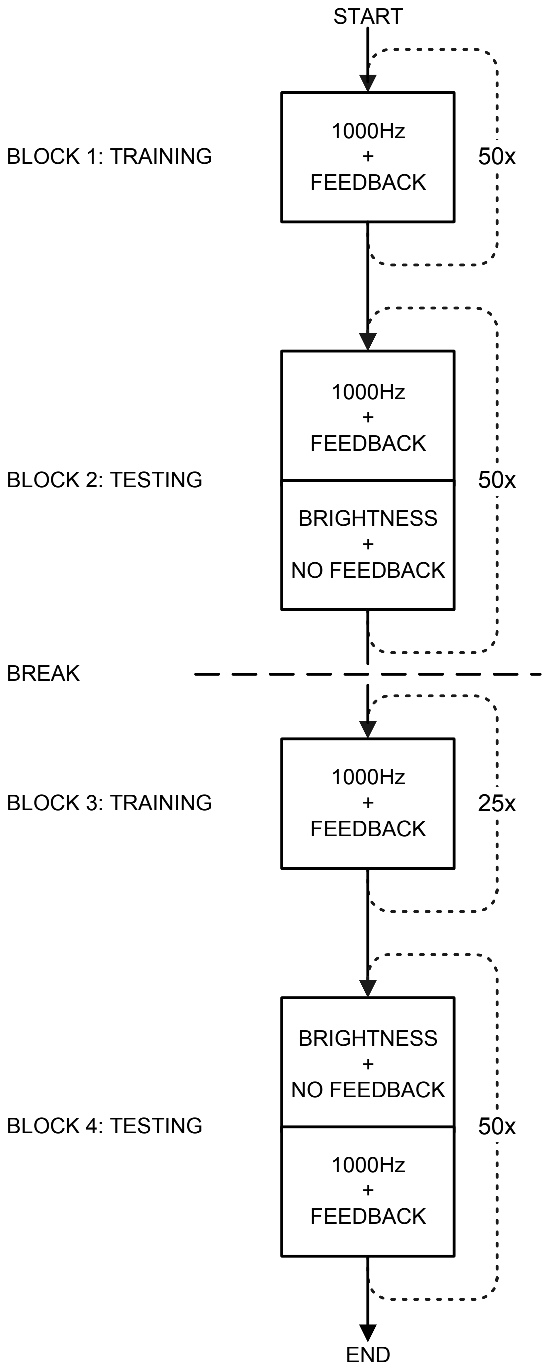


Figure 12. Schematic flow of the constrained scaling experiment in Experiment 3.

follow Experiments 1 and 2. As in Experiment 2, the experiment was not conducted in a sound attenuating chamber. Participants were initially trained on 50 iterations of the 1000 Hz loudness stimuli with feedback about the correct loudness value. Subsequently, they were presented 100 trials of loudness stimuli with feedback paired with brightness stimuli without feedback. The participants were instructed to scale the brightness stimuli according to the learned loudness scale. Following a short break, the participants repeated the same procedure, with 50 additional loudness training trials and 100 pairs of loudness and brightness stimuli. The order of the 100 pairs of brightness and loudness stimuli was counterbalanced to the order of the 100 pairs in the first part of the experiment.

Results and Discussion

The results were analyzed as in Experiments 1 and 2. For brightness, the logarithm of the participant responses (*R*) was regressed against the logarithm of the luminances (*L*) in cd/m^2:

$$\log R = m \log L + \log a \, , \tag{15}$$

which was transformed to the Power Law form:

$$R = aL^m \, . \tag{16}$$

The mean, coefficient of variation, and highest-to-lowest ratio were calculated for the exponent (*m*), the intercept (*a*), and the goodness of fit coefficient (R^2). The ratios of loudness to brightness for these values were also calculated. A summary of the results is found in Table 5. The summary graphs for loudness and brightness scaling are found in Figures 13 – 15.

Table 5. Summary of participants' loudness and brightness scaling in Experiment 3.

	Loudness + Feedback			**Brightness + No Feedback**			**Ratio of Loudness to Brightness**		
P	**Exponent**	**Intercept**	***R*²**	**Exponent**	**Intercept**	***R*²**	**Exponent**	**Intercept**	***R*²**
1	0.524	1.220	0.881	0.277	1.502	0.840	1.889	0.812	1.048
2	0.563	1.243	0.825	0.485	1.108	0.785	1.161	1.122	1.051
3	0.520	1.279	0.843	0.218	1.602	0.702	2.386	0.798	1.201
4	0.510	1.300	0.818	0.493	1.071	0.764	1.034	1.214	1.071
5	0.547	1.280	0.858	0.604	0.815	0.845	0.906	1.570	1.015
M	0.533	1.264	0.845	0.415	1.220	0.787	1.475	1.103	1.077
SD	0.022	0.032	0.025	0.162	0.326	0.059	0.636	0.319	0.072

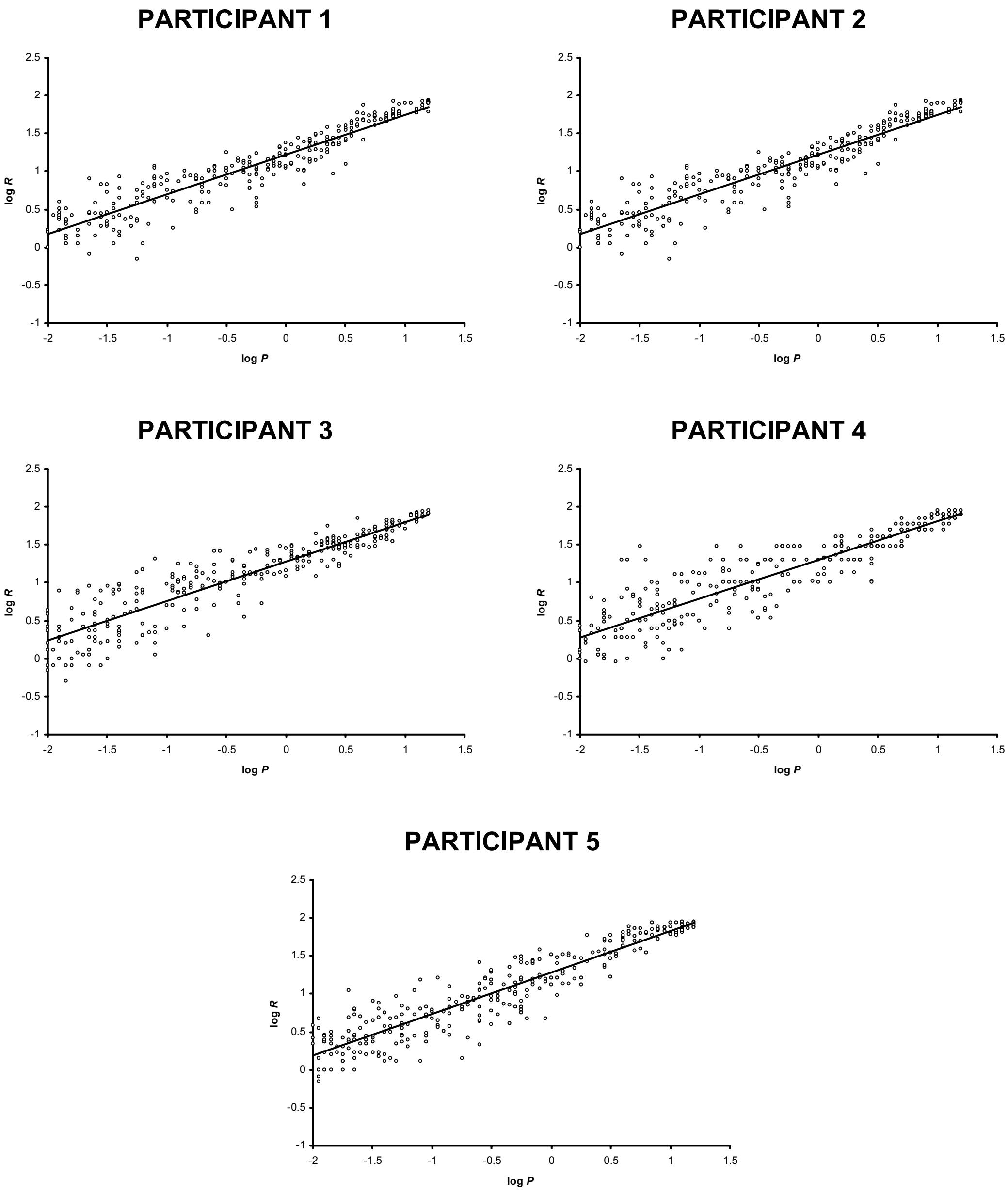


Figure 13. Logarithmic scatterplot and regression line for loudness in dynes/cm^2 (P) and participant response (R) for 1000 Hz tones with feedback in Experiment 3.

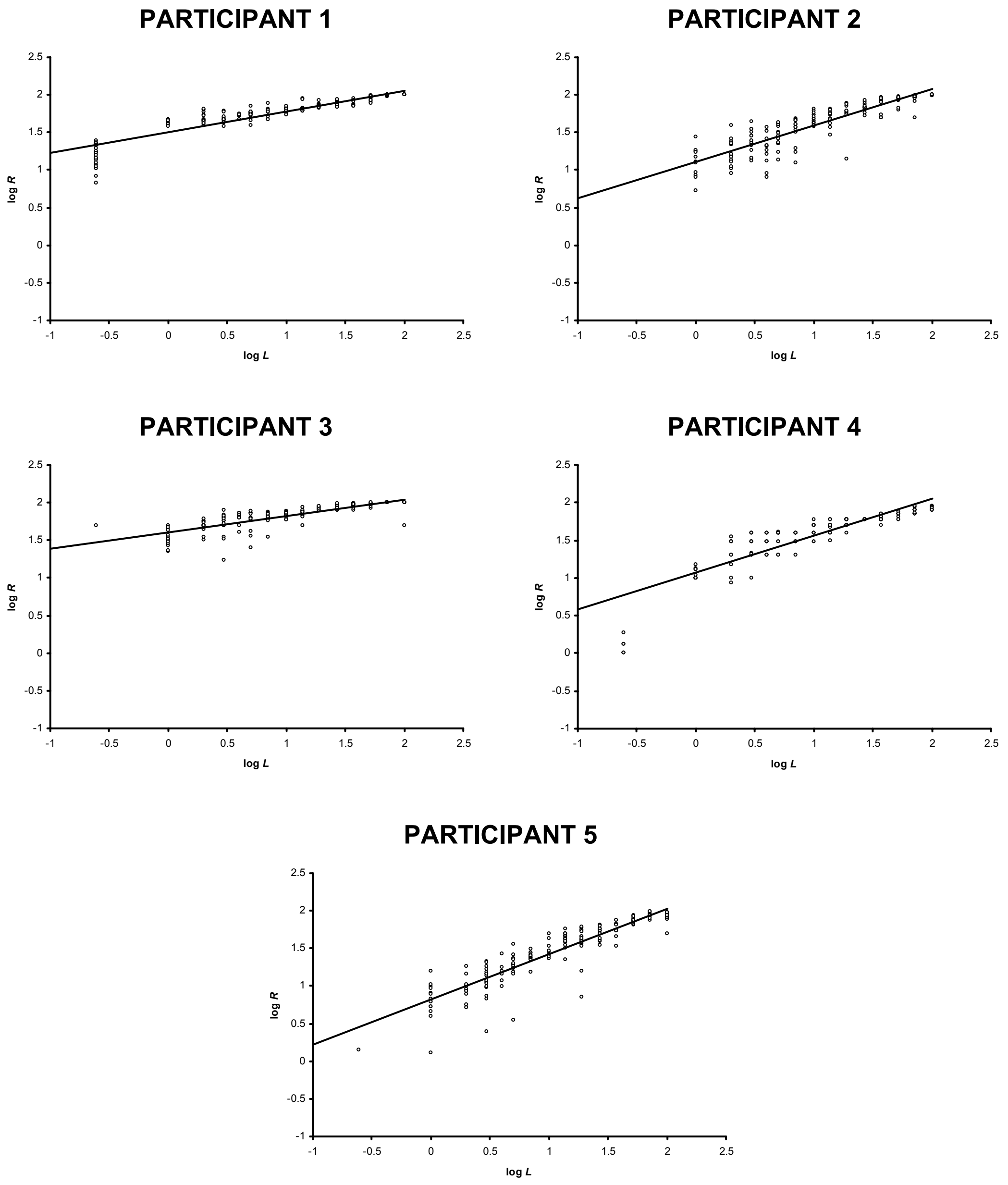
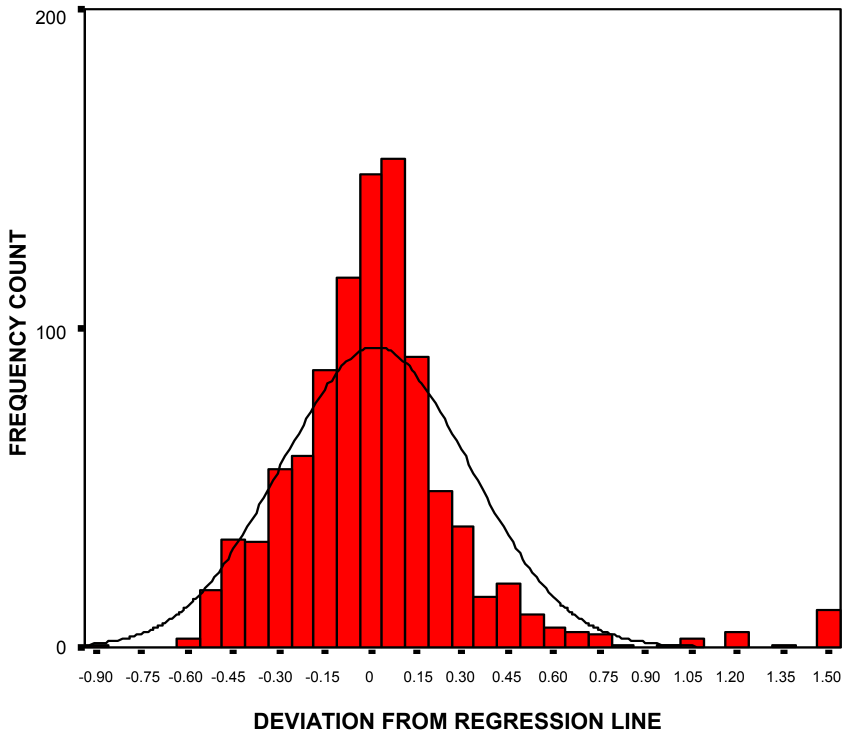


Figure 14. Logarithmic scatterplot and regression line for brightness in cd/m^2 *(L) and participant response (R) for grayscale squares in Experiment 3.*



NOTE: The line through the histograph depicts the normal distribution.

Figure 15. Frequency distributions for scores about the average regression line for brightness stimuli in Experiment 3.

Exponent Values

For the training stimuli of 1000 Hz tones with feedback, the average exponent, *m*, was 0.533, with *SD*/*M* = 0.041 and *H*:*L* = 1.103:1. This exponent value was congruent to the values obtained in Experiments 1 and 2. For the grayscale brightness testing stimuli, the average exponent was 0.415, with *SD*/*M* = 0.389 and *H*:*L* = 2.772:1. The average brightness scaling exponent was lower than the value reported for brightness scaling in West et al. (2000).[31] There was considerable interparticipant variability in the brightness scaling results, comparable to the magnitude estimation and cross-modal matching but not the constrained scaling results reported in West et al. The average exponent ratio of loudness to brightness in the present experiment was 1.475:1.

Intercept Values

For the 1000 Hz tones with feedback, the average intercept, *a*, was 1.264, with *SD*/*M* = 0.026 and *H*:*L* = 1.066:1. Experiments 1 and 2 provided nearly identical intercept and variability results for the loudness training stimuli. For the grayscale brightness testing stimuli, the average intercept was 1.220, with *SD*/*M* = 0.267 and *H*:*L* = 1.965:1. These brightness scaling intercept variability measures were considerably higher than the loudness scaling intercept variability values obtained in Experiments 1

[31] West et al. (2000) removed one participant's results as an extreme outlier. With and without the outlying participant in West et al., the present experiment exhibits higher variability. Note that the present exponent value closely matches the brightness exponent suggested in Stevens and Hall (1966).

and 2.[32] The average intercept ratio of loudness to brightness in the present experiment was 1.103:1.

Goodness of Fit Coefficients

For the loudness training stimuli, the average goodness of fit coefficient, R^2, was 0.845, with *SD*/*M* = 0.030 and *H*:*L* = 1.077:1. For the brightness testing stimuli, the average goodness of fit coefficient was 0.781, with *SD*/M = 0.075 and *H*:*L* = 1.204:1. The goodness of fit for both loudness and brightness scaling was comparable to that found in Experiments 1 and 2. The goodness of fit ratio of loudness to brightness was 1.072, suggesting nearly equal goodness of fit for the regression line across loudness and brightness scaling conditions.

Score Distribution

As in Experiment 1, it is useful to review the distribution of the participants' scores to determine if they follow a normal distribution. Figure 15 shows the histogram for the frequency distribution of scores about the averaged regression line for brightness stimuli. As with loudness in Experiment 1, there is high kurtosis. There is a slight skew to the upper tail, implying some significant deviations indicative of slight scale curvature. Overall, however, the curve follows a roughly normally shaped frequency distribution.

Range Effects

By design, constrained scaling calibrates individuals to use scaling in a natural manner. For example, it would be expected that a person who is trained on a loudness

[32] A direct comparison to West et al. (2000) is not possible, since West et al. do not report intercept values.

scale with an exponent equal to 0.60 has learned a natural mapping between loudness and a numerical scale. It is also assumed that this individual learns a more general mapping between perceptual continua and a numerical scale. In Experiment 4 in West et al. (2000), it was demonstrated that participants trained on a loudness scale did not repeat the exact scale when subsequently scaling brightness. Instead, they scaled loudness according to the natural brightness exponent, 0.33. This finding was not replicated in the present experiment.

The effectiveness of cross-modal scaling is contingent on a number of factors, including the range of the scale. Teghtsoonian and Teghtsoonian (1997) suggest that Stevens' natural scaling exponent, *m*, for any modality is the ratio of the subjective response scaling dynamic range (log R_R max) to the stimulus dynamic range (log R_S max):

$$m = \frac{\log R_R \max}{\log R_S \max}, \tag{17}$$

whereby log R_R max is a constant across stimulus modalities.[33] In most experiments, R_S max has a dynamic range around 2 log units, which is equivalent to a 100-point subjective response scale. In Teghtsoonian and Teghtsoonian, however, the utilized subjective scaling range was 1.53 log units. Equation 17 was put into practice for five stimulus modalities (i.e., loudness, heaviness, sniff vigor, handgrip, and shock intensity). Using a magnitude production method to allow participants to select the range of

[33] The variable notation used by Teghtsoonian and Teghtsoonian (1997) has been changed to concur with the notation used in this dissertation. The present Equation 17 is functionally equivalent to Equation 2 in Teghtsoonian and Teghtsoonian.

acceptable stimulus intensities, they found log R_S max equal to 2.22 for loudness, 1.12 for heaviness, 0.89 for sniff vigor, 0.85 for handgrip, and 0.50 for shock intensity. The resulting exponent values were 0.69 for loudness, 1.37 for heaviness, 1.72 for sniff vigor, 1.81 for handgrip, and 3.06 for shock intensity. These exponent values generally matched exponent values reported in previous scaling literature.

Stevens (1975) suggests that the natural scaling exponent for brightness is 0.33. Given Equation 17 and Teghtsoonian and Teghtsoonian's (1997) subjective scaling range equal to 1.53, the stimulus range would be calculated as follows:

$$0.33 = \frac{1.53}{\log R_S \text{ max}} \tag{18}$$

$$\therefore \log R_S \text{ max} = \frac{1.53}{0.33} = 4.64. \tag{19}$$

For Teghtsoonian and Teghtsoonian's subjective scaling range, the range of acceptable stimulus intensities for brightness would be 4.64 log units.[34]

An important implication of Teghtsoonian and Teghtsoonian's exploration of the dynamic stimulus and subjective scaling range across modalities is that the dynamic range may be prescriptive for constrained scaling. Constrained scaling provides training on the subjective scaling range given a range of stimulus intensities. When using a standard scaling dynamic range of 2 log units and the range of acceptable stimulus

[34] Teghtsoonian and Teghtsoonian (1997) do not provide an explanation of the underlying stimulus scales used in their experiments. For example, the log R_S max for loudness is reported as being 2.22. The antilog of this number is 165.96. Presumably this is not the range of acceptable stimulus intensities along the decibel scale! An upper bound of 165 dB, for example, would considerably exceed the point of pain and hearing damage.

intensities, the resulting exponent value would be expected to match Stevens' (1975) natural scaling exponent values for a given modality. However, what is the expected outcome when using a stimulus range that is smaller than the full range of acceptable stimulus intensities? One possibility is that scalers will use the full response scale even for a restricted stimulus range. If the response scale remains constant across different stimulus scales, the stimulus range has a potentially large impact on the resultant exponent. A restricted stimulus range would potentially result in a steeper slope (and larger exponent) than would the maximized range of acceptable stimulus intensities. Due to constraints with the maximum luminous intensity output by the computer display used in the present experiment, the maximum brightness displayed was 4.25 times dimmer than the maximum brightness displayed using the LED in West et al. (2000).

Figure 16 shows a comparison of the theoretical brightness plots if participants scaled the full range of response, *R*, across the available stimulus range, *L*. For West et al., the brightness stimuli ranged from 0 to 274 cd/m^2; in the present experiment, the brightness stimuli ranged from 0 to 64 cd/m^2. Assuming participants would use the full response range (0 to 100) to scale brightness, participants in the experiment by West et al. would be expected to respond with a scaling value around 100 for the brightest stimulus at 274 cd/m^2. Likewise, participants in the present experiment would be expected to respond with a scaling value around 100 for the brightest stimulus at 64 cd/m^2. Since the loudness training stimuli had a *y*-intercept around 16.6, this intercept is assumed to be transferred to brightness scaling. The assumption made in Figure 16 is that participants adjust their scaling relative to the range of stimulus intensities they encounter.

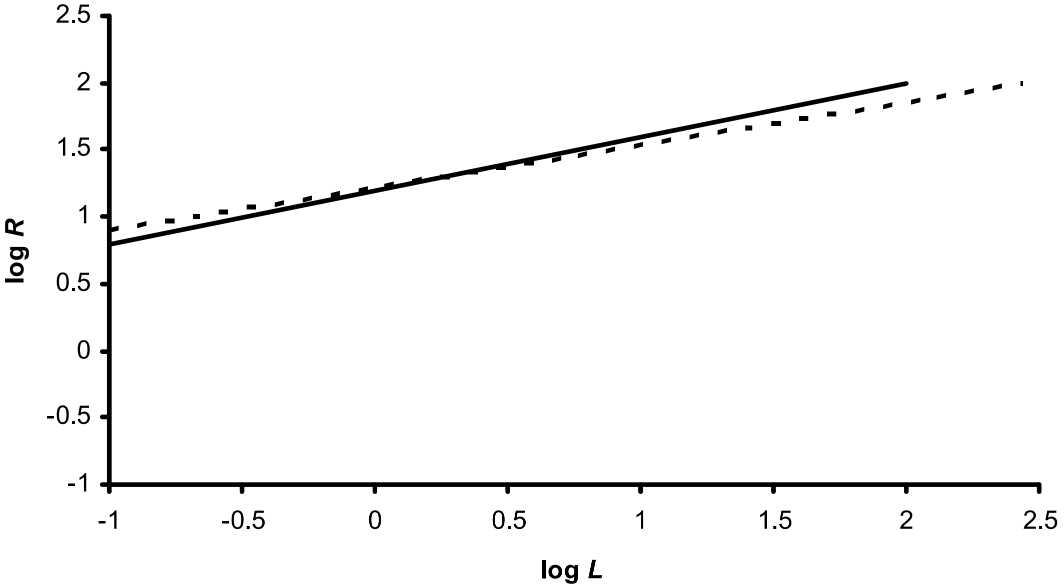


Figure 16. *Comparison of theoretical brightness scaling results in West et al. (dotted line) and in Experiment 3 (solid line).*

Regressing the line for the theoretical range distribution data[35] produces the following equation for West et al.:

$$\log R = 0.32 \log L + \log 16.6, \tag{20}$$

which is equivalent to:

$$R = 16.6 L^{0.32}. \tag{21}$$

Similarly, regressing the line for the theoretical range distribution data produces the following equation for the present experiment:

$$\log R = 0.40 \log L + \log 16.6, \tag{22}$$

which is equivalent to:

$$R = 16.6 L^{0.40}. \tag{23}$$

Note that Equations 21 and 23 nearly exactly match the actual results in West et al. (2000) and in the present experiment, respectively.

One interpretation of this finding is that participants trained on constrained scaling matched the response range to the stimulus range. In the case of West et al., this matching was masked by the stimulus range that produced an exponent value in line with Stevens' natural brightness scaling exponent (Stevens, 1975). One may assume that range effects vary between participants, because the mapping of one stimulus range to another is subject to individual differences (Teghtsoonian, 1989). Support for this interpretation is found in the increased interparticipant variability of brightness scaling

[35] It was necessary to regress the data instead of calculating the exponent value according to Equation 17, because Teghtsoonian and Teghtsoonian (1997) did not provide guidance on the use of the underlying stimulus scale for this equation.

compared to other forms of constrained scaling.[36] Another interpretation of these results is that they are coincidentally related to a range effect but represent a natural scaling exponent for onscreen grayscale, which is different from the natural brightness scaling exponent reported in Stevens (1975) and West et al. (2000). The remaining experiments in this dissertation will shed further light on this phenomenon and offer suggestions for resolving these two interpretations.

General Discussion

The present experiment demonstrated one method of refining constrained scaling methodology. The use of constrained scaling cross-modally from loudness to brightness was not supported by the results, in which first training on a loudness scale did not seem to reduce the scaling variability beyond the variability expected in a brightness magnitude estimation study. In order to explore brightness scaling, the subsequent series of experiments look at the relationship between brightness training and loudness testing (Experiments 4 and 5). Experiment 6 explores brightness scaling across colors, in which participants are trained on grayscale brightness and tested on the brightness of red, green, and blue colors. Experiment 7 collects baseline magnitude estimation measures for grayscale, red, green, and blue colors. This series of experiments allows a comparison of scaling results, in particular scaling variability, between cross-modal loudness-brightness scaling in Experiments 3 – 5 and intramodal brightness scaling in Experiments 6 and 7.

[36] The increased variability was more prevalent for the present experiment than for West et al. (2000).

EXPERIMENT 4

Introduction

Whereas Experiment 3 used loudness as the training scale for brightness, the present experiment reversed this design. Brightness was used as the training scale for loudness. The purpose of this experiment was to determine if training scales other than loudness would prove effective at generating scaling responses with low interrater variability. Since no experiment has been reported in the constrained scaling literature involving training on brightness stimuli and testing on loudness stimuli, this experiment is considered an extension of constrained scaling methodology—it presents a novel domain for constrained scaling.

Method

Participants

As in previous experiments, five participants with self-reported normal hearing and normal color vision volunteered for this experiment. The volunteers were paid $10 for their participation in the experiment.

Apparatus and Stimulus Materials

The apparatus was identical to the apparatus developed and employed in Experiment 3. The experimental control software featured precise calibration of sound card amplitude as measured in decibels (see Appendix B) and of the brightness of onscreen squares as measured in lumens (see Appendix C).

Design and Procedure

The schematic flow of Experiment 4 is presented in Figure 17. The design mimicked the design of Experiment 3, except training involved grayscale squares of various brightness levels with feedback, and testing involved 1000 Hz tones without feedback. The participants were trained on grayscale squares according to the following equation:

$$R = 21.8L^{0.33} \tag{24}$$

where L is the luminous intensity in cd/m^2, 0.33 is Stevens' natural exponent for brightness scaling, and the coefficient 21.8 adjusts the scale to fit a 100-point scale range. L ranged from 0 to 100 cd/m^2 along 15 logarithmically spaced stimulus points (see Table C-1 in Appendix C).

Results and Discussion

The results of the present experiment are summarized in Tables 6 – 7 and in Figures 15 – 22. Note that the present experiment features a comparison of reaction times across brightness and loudness scaling trials, which is featured in Figure 22 and Table 7, and is explained later in this section.

Exponent Values

The mean exponent value for the brightness training stimuli with feedback was 0.314, with *SD*/*M* equal to 0.054 and *H*:*L* equal to 1.100:1. The participants demonstrated good learning of the brightness training stimuli and exhibited low variability in line with constrained loudness scaling experiments. The mean exponent value for the loudness testing stimuli without feedback was 0.326, with *SD*/*M* equal to

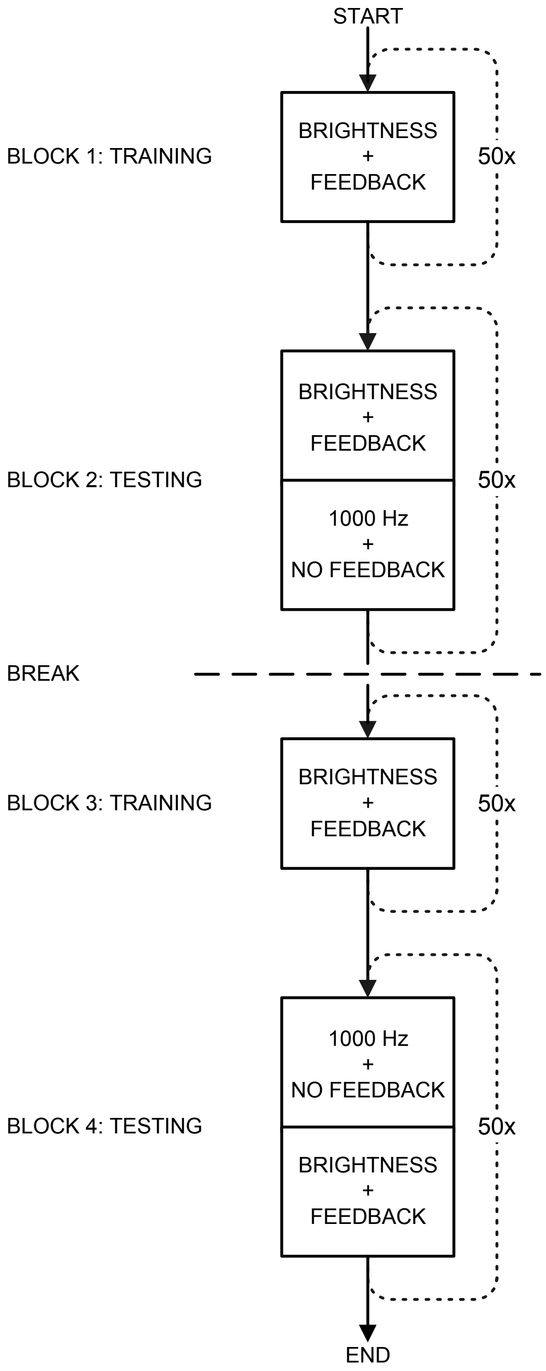


Figure 17. *Schematic flow of the constrained scaling experiment in Experiment 4.*

Table 6. Summary of participants' brightness and loudness scaling in Experiment 4.

	Brightness + Feedback			**Loudness + No Feedback**			**Ratio of Brightness to Loudness**		
P	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**
1	0.331	1.352	0.820	0.294	1.733	0.538	1.126	0.780	1.524
2	0.329	1.339	0.911	0.328	1.589	0.759	1.004	0.842	1.200
3	0.301	1.371	0.915	0.418	1.587	0.713	0.720	0.864	1.283
4	0.301	1.365	0.922	0.247	1.711	0.847	1.216	0.798	1.088
5	0.306	1.370	0.888	0.343	1.730	0.803	0.893	0.792	1.106
M	0.314	1.359	0.891	0.326	1.670	0.732	0.992	0.815	1.240
SD	0.017	0.015	0.048	0.072	0.078	0.191	0.216	0.039	0.185

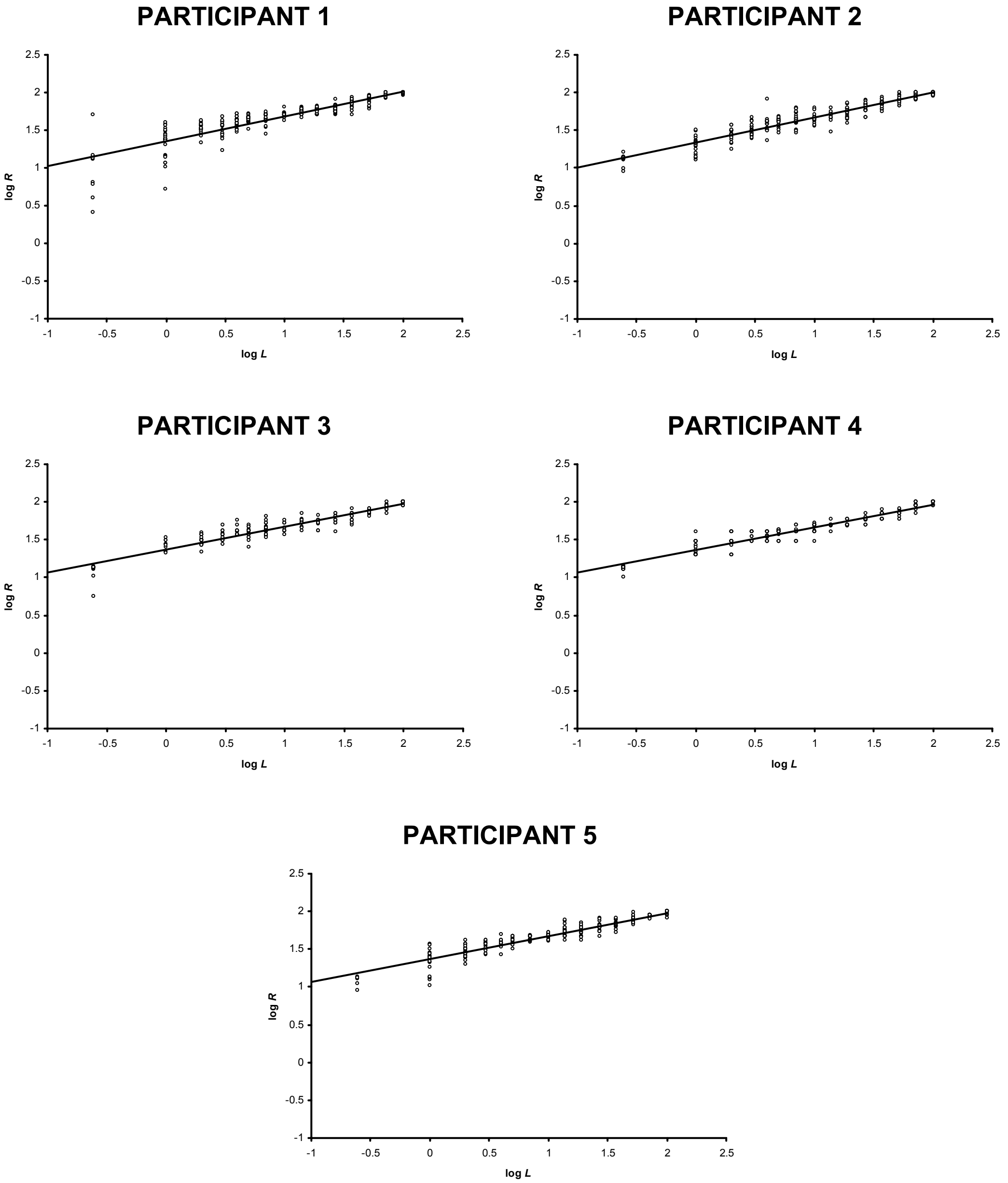


Figure 18. Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for grayscale squares with feedback in Experiment 4.

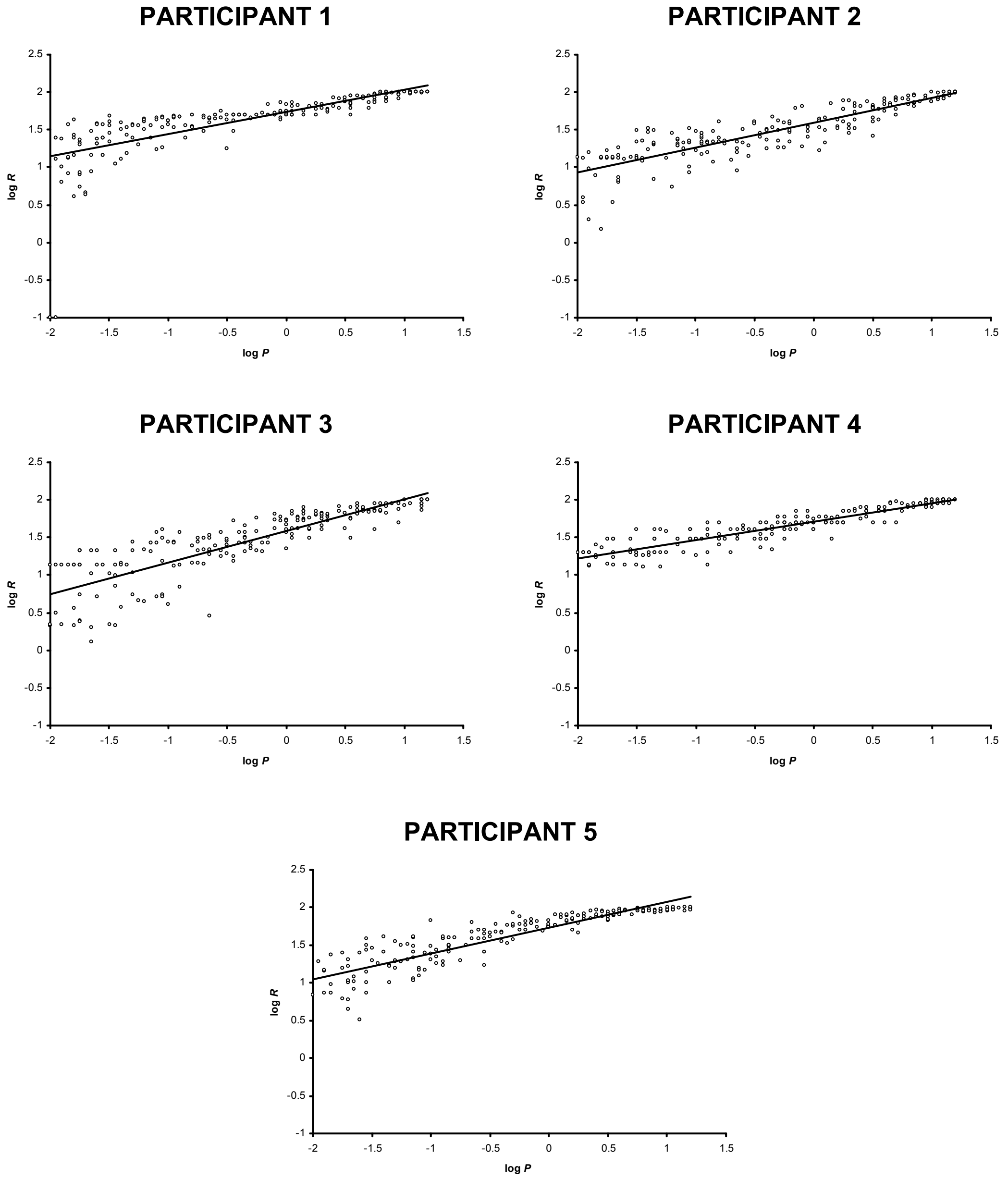


Figure 19. *Logarithmic scatterplot and regression line for loudness in dynes/cm^2 (P) and participant response (R) for 1000 Hz tones without feedback in Experiment 4.*

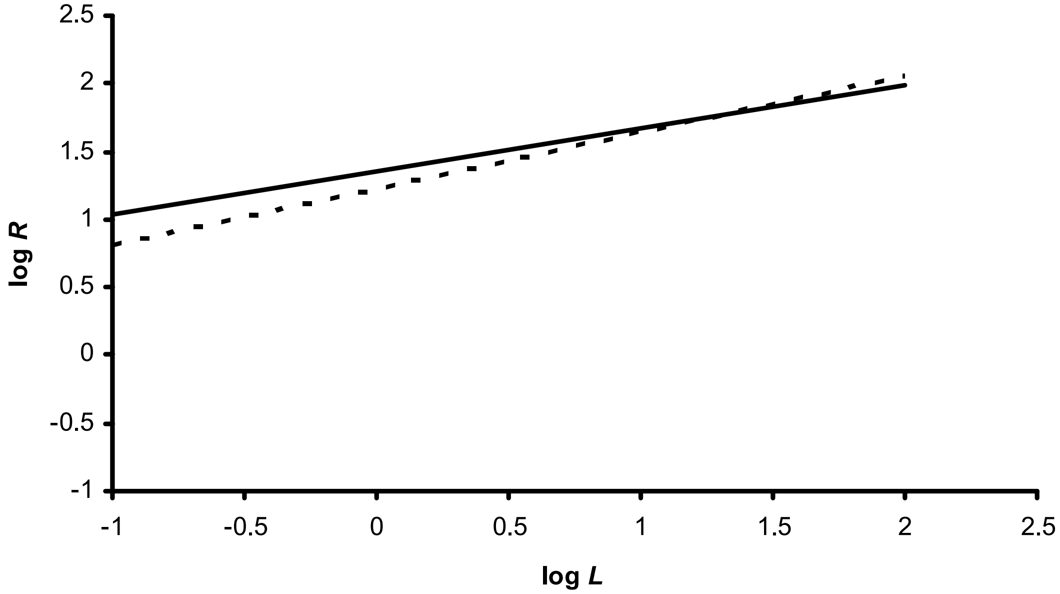


Figure 20. Comparison of constrained brightness scaling for Experiment 3 (dotted line) and Experiment 4 (solid line).

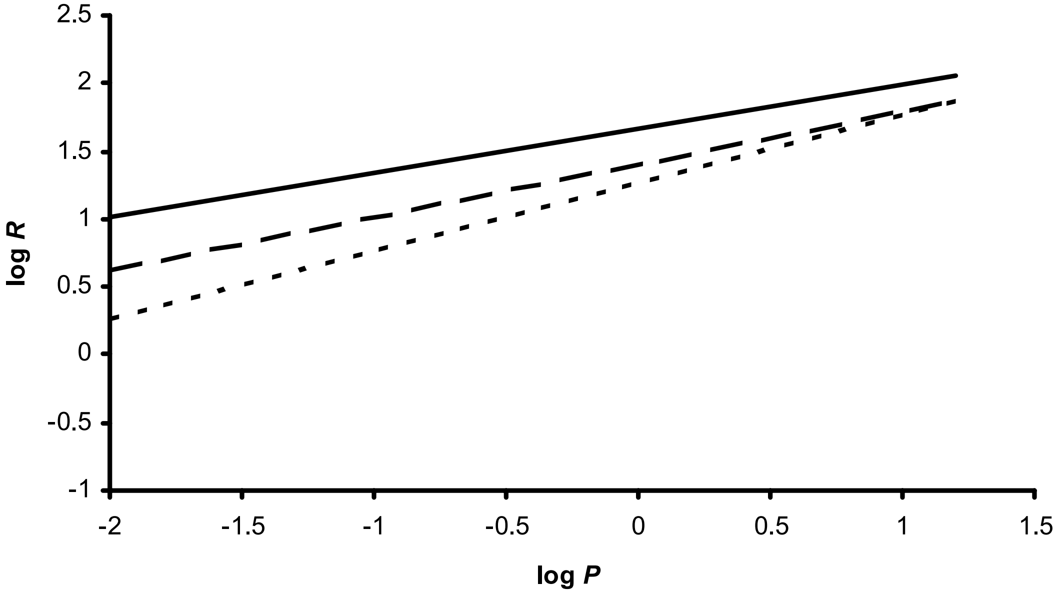


Figure 21. Comparison of constrained loudness scaling for Experiments 1 (dotted line), Experiment 4 (solid line), and theoretical loudness with a brightness intercept (dashed line).

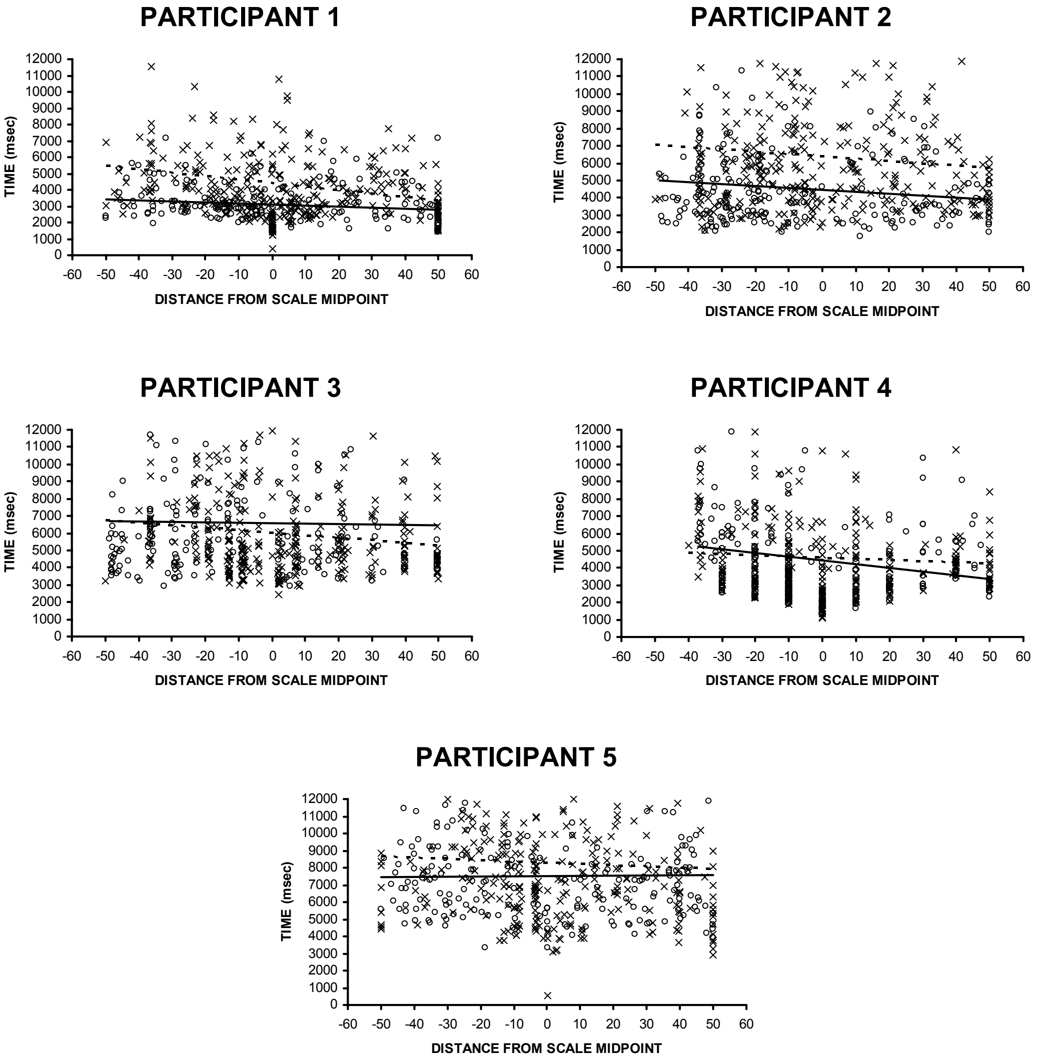


Figure 22. *Reaction time for training (×'s with dotted regression line) and testing (○'s with solid regression line) stimuli according to distance from the scale midpoint in Experiment 4.*

Table 7. Average reaction times (ms) for each participant for brightness training and loudness testing in Experiment 4.

P	**Training**	**Testing**
1	4339	3087
2	6386	4560
3	5998	6625
4	4612	4433
5	8303	7513
M	5928	5244
SD	1590	1791

0.222 and *H*:*L* equal to 1.691:1. This exponent value was lower than the exponent found in constrained loudness scaling experiments. In fact, the average loudness exponent value matched the brightness training exponent, which deviates significantly from the expected natural scaling exponent of 0.60 for loudness. The average ratio of brightness to loudness exponents equaled 0.992, whereas the expected ratio would be 0.55.[37]

Intercept Values

The average intercept for the grayscale training stimuli with feedback was 1.359, with *SD*/*M* equal to 0.011 and *H*:*L* equal to 1.024:1. This value was very close to the training intercept of 21.8, the log of which equals 1.338. The average intercept for the loudness testing stimuli without feedback was 1.670, with *SD*/*M* equal to 0.047 and *H*:*L* equal to 1.092:1. This intercept was higher than the training intercept and curiously higher than the intercepts reported for loudness scaling in previous experiments. This finding is discussed further in the section entitled *Range Effects*. The intercept ratio of brightness to loudness was 0.815, whereas in Experiment 3, the brightness and loudness intercept were nearly identical. The cross-modal verbatim transfer of the training intercept to the testing intercept in Experiment 3 was not found in the present experiment.

Goodness of Fit Coefficients

The average goodness of fit coefficient for the brightness training stimuli with feedback equaled 0.891, with *SD*/*M* equal to 0.054 and *H*:*L* equal to 1.125:1. This value

[37] Using Stevens' (1975) natural scaling exponents of 0.33 for brightness and 0.60 for loudness, the ratio of brightness to loudness exponents would equal 0.33/0.60 or 0.55.

indicated that participants were quite consistent in scaling the brightness stimuli. For the loudness stimuli, the average goodness of fit coefficient equaled 0.732, with *SD*/*M* equal to 0.178 and *H*:*L* equal to 1.575:1. The average goodness of fit coefficient was comparable to that obtained the training scaling conditions in previous experiments, but the variability was somewhat higher, suggesting some inconsistency in scaling across participants. The average ratio of brightness to loudness R^2 values was 1.240.

Reaction Time

In order to offer possible insights into how scaling works when transferring a learned scale to another sensory modality, reaction time data were collected for each participant for all trials. The reaction time data were gathered starting with the presentation of the brightness or loudness stimuli. Reaction time data were collected with a millisecond accurate timer, as described in Appendix A.

During each trial, the starting slider position was always centered on a value of 50. It is assumed that it will take the participant longer to select a value at the endpoints than toward the center of the slider scale. Fitts (1954) suggested that movement time was a logarithmic function of the movement amplitude and the distance to travel:

$$MT = a + b\log_2\left(\frac{2A}{W}\right), \tag{25}$$

where *MT* is the movement time, *a* and *b* are constants related to the regression line for line traveling along a straight line, *A* is the movement amplitude, and *W* is the target width. To consider the applicability of Fitts' Law, the reaction time data were plotted as a function of the response, *R*, distance from the center of the slider scale (see Figure 22).

The figure illustrates that there was no consistent relationship between the response position and the reaction time and that the functional relationship between slider position and reaction time is noisy at best. However, the regression lines do illustrate that there is a very slight tendency for the larger response values to have a faster reaction time. A slight decrease in reaction time for higher intensity brightness stimuli is consistent with the research literature (Pease, 1964).

A comparison of the average reaction times for the brightness training and the loudness testing stimuli is provided in Table 7. For the brightness training stimuli, the average reaction time was 5928 ms (*SD* = 1590), while for the loudness testing stimuli, the average reaction time was 5244 ms (*SD* = 1791). The 628 ms advantage offered for the testing stimuli was not statistically significant, $t(4) = \pm 1.610$, $p = 0.183$.

Range Effects

In Experiment 3, the possibility was considered that the stimulus range might influence the scaling slope or exponent. Figure 20 illustrates the relationship between brightness scaling slopes in Experiment 3 and the present experiment. As can be seen, the two lines follow a similar pattern and are closely related. The line for the present experiment features a larger intercept but a flatter slope, while the line for Experiment 3 features a smaller intercept but a slightly steeper slope. The differences between the two scaling lines are more pronounced at the lower stimulus end than at the upper stimulus end, with an intersection of the two scales in the upper stimuli.

Figure 21 illustrates the relationship between loudness scaling for Experiment 1 and the present experiment. Here are two strikingly different scales. The present experiment has a considerably larger intercept and a much flatter slope than Experiment 1. The flatter loudness slope in the present experiment is not attributable to a stimulus range effect, because the same set of loudness stimuli were used in both experiments. The different loudness scaling slope in the present experiment is also not attributable to a carryover of the brightness scaling intercept to the loudness scale, because the loudness intercept was larger than the brightness intercept.

Figure 21 includes another line to explore a third possibility. The dashed line represents the theoretical scale that would occur if there were a carryover of the brightness scaling intercept yet a subjective scaling upper bound equivalent to the maximum subjective loudness scaling value from Experiment 1. This hybrid scale has an exponent value equal to 0.388, slightly higher than the actual exponent of 0.326. Its scaling intercept is equal to 1.398, which is lower than the actual intercept of 1.670. This theoretical scale offers a hypothetical assuagement for the unexpected loudness exponent, but it fails to explain the elevated loudness intercept in the present experiment.

General Discussion

The loudness stimuli without feedback were scaled in a fashion clearly accounted for by neither constrained scaling theory nor range effects. It appears that participants in the present experiment scaled loudness according to the same exponent on which they were trained to scale brightness. This would be an unnatural scaling exponent for the modality of loudness, which could explain the increase in variability compared to intra-modal constrained scaling experiments.

However, this near perfect match of exponents from brightness to loudness was not found in the reverse direction in Experiment 3. Participants did not appear to use the same scaling exponent for brightness as they had been trained upon for loudness. But, in the case of Experiment 3, the intercept value was clearly carried over from loudness to brightness. In the present experiment, the intercept values were quite different between brightness and loudness. The lack of bidirectionality in the findings of Experiments 3 and 4 muddles the theoretical clarity of cross-modal constrained scaling.

EXPERIMENT 5

Introduction

Stevens (1975) identifies four separate exponent values for brightness. These are: an exponent of 0.33 for a 5º target in the dark, an exponent of 0.50 for a point light source, an exponent of 0.5 for a briefly flashed brightness stimulus, and an exponent of 1.0 for a point light source that is briefly flashed. Generally speaking, according to Stevens' findings, flashing a light source has the effect of increasing the exponent.

The present experiment is a direct follow-on to Experiments 3 and 4 and a near replication of Experiment 4. In Experiments 3 and 4, the brightness stimuli were presented for the duration of the scaling trial. The grayscale square was left on the screen until the participant had selected the scaling response value. In contrast, in West et al. (2000), the brightness stimuli were displayed for 1 second and then removed. To help account for the differences in exponents between West et al. and Experiments 3 and 4, a variation of Experiment 4 was implemented in which the grayscale boxes were flashed on the screen for 1 second.

Method

Participants

As in previous experiments, five paid participants with self-reported normal hearing and color vision were enlisted for the experiment.

Apparatus and Stimulus Materials

The apparatus was identical to the apparatus in Experiment 4 with the exception that the present experiment incorporated a timer for the display of the grayscale squares.

After 1 second, the grayscale squares were blacked out. The experiment control software displayed a dark gray border around the square to identify its location.

Design and Procedure

The design and procedure were identical to Experiment 4, except participants were given cues to indicate that a square was about to be displayed. The participants could begin selecting a scale value as soon as the grayscale square displayed on the screen, but the grayscale square disappeared automatically after 1 second.

Results and Discussion

The results were analyzed as in Experiment 4 and are summarized in Tables 8 – 9 and Figures 23 – 25. As in Experiment 4, reaction time data were analyzed, and are presented here in Table 9.

Exponent Values

The average exponent value for the flashed brightness stimuli with feedback was 0.306, with *SD*/*M* equal to 0.073 and *H*:*L* equal to 1.195:1. These values closely matched the brightness scaling exponents found in Experiment 4. For the loudness stimuli without feedback, the average exponent value was 0.408, with *SD*/*M* equal to 0.228 and *H*:*L* equal to 1.903:1. The loudness exponent in the present experiment was higher than the exponent in Experiment 4, but variability remained at comparable levels. Although higher than the exponent in Experiment 4, this exponent was still considerably lower than the expected exponent of 0.6 for loudness scaling. The average ratio of brightness to loudness exponents equaled 0.749, which was higher than the predicted exponent value of 0.55 in Experiment 4.

Table 8. Summary of participants' brightness and loudness scaling in Experiment 5.

	Brightness + Feedback			**Loudness + No Feedback**			**Ratio of Brightness to Loudness**		
P	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**
1	0.318	1.353	0.697	0.501	1.380	0.907	0.634	0.980	0.769
2	0.320	1.353	0.928	0.478	1.558	0.847	0.671	0.868	1.096
3	0.299	1.398	0.878	0.263	1.742	0.715	1.135	0.803	1.229
4	0.270	1.403	0.595	0.399	1.315	0.736	0.675	1.067	0.809
5	0.322	1.407	0.830	0.401	1.469	0.736	0.804	0.958	1.128
M	0.306	1.383	0.786	0.408	1.493	0.788	0.749	0.926	0.997
SD	0.022	0.028	0.137	0.093	0.167	0.084	0.240	0.165	1.620

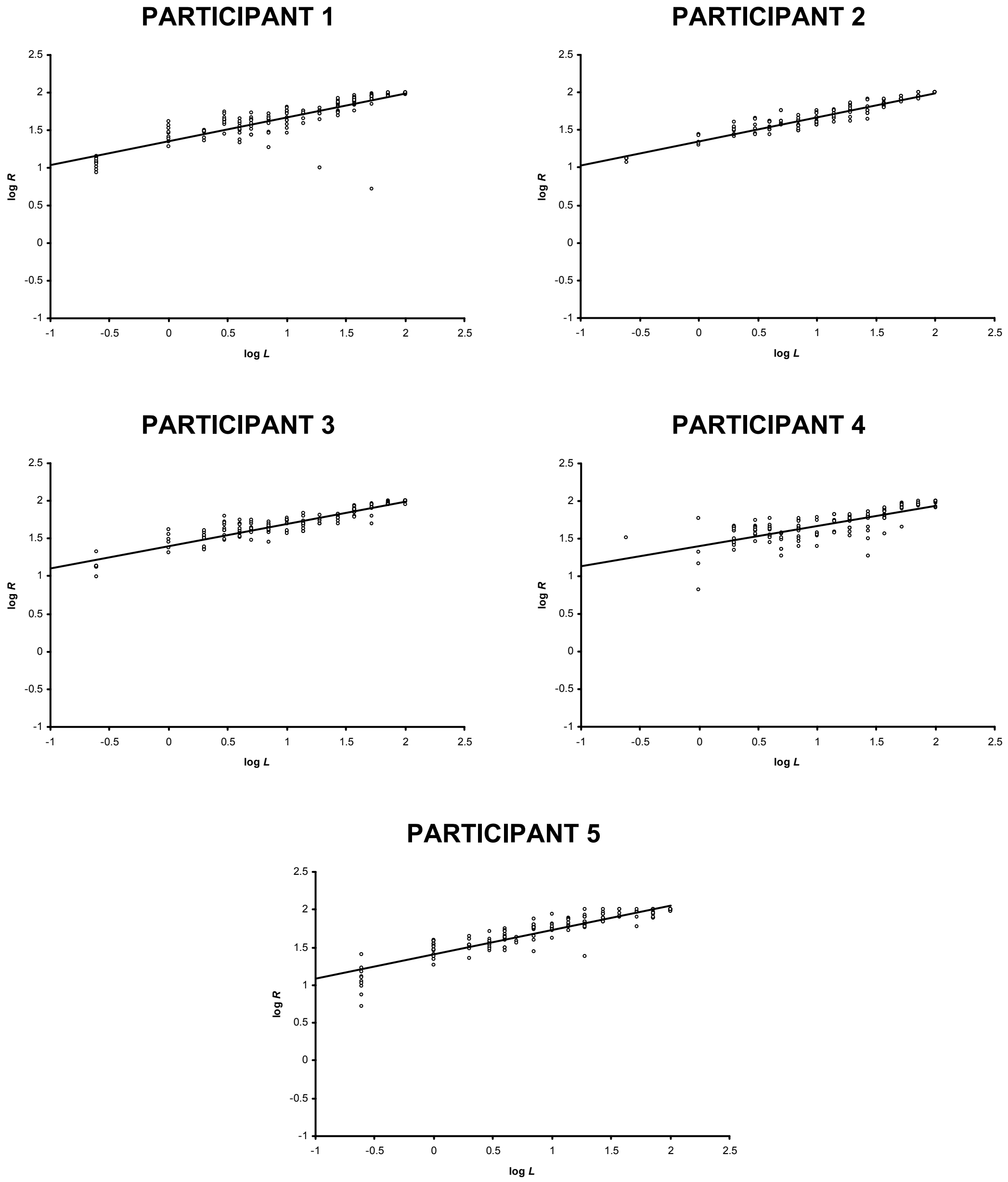
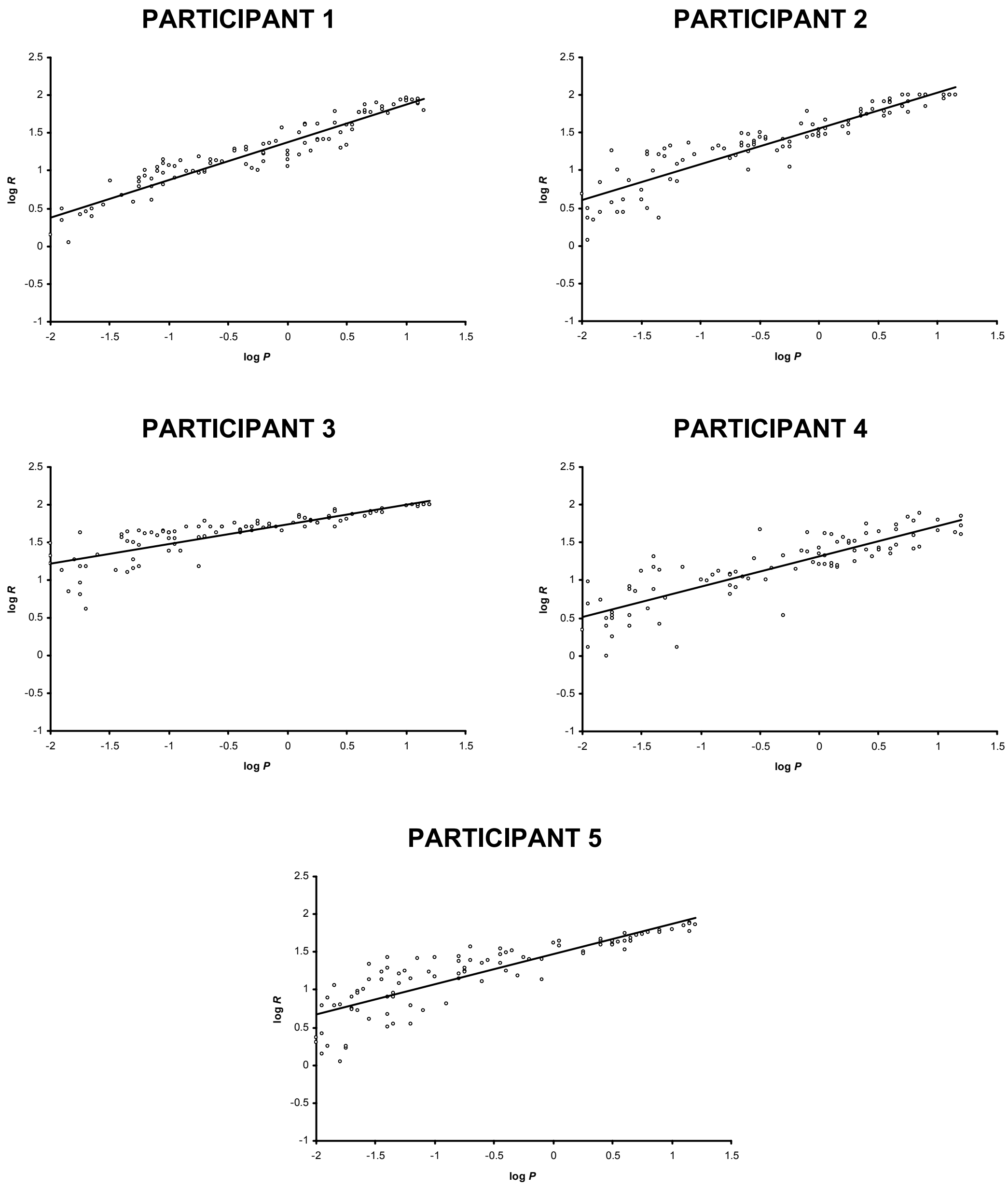


Figure 23. Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for grayscale squares with feedback in Experiment 5.



PARTICIPANT 1

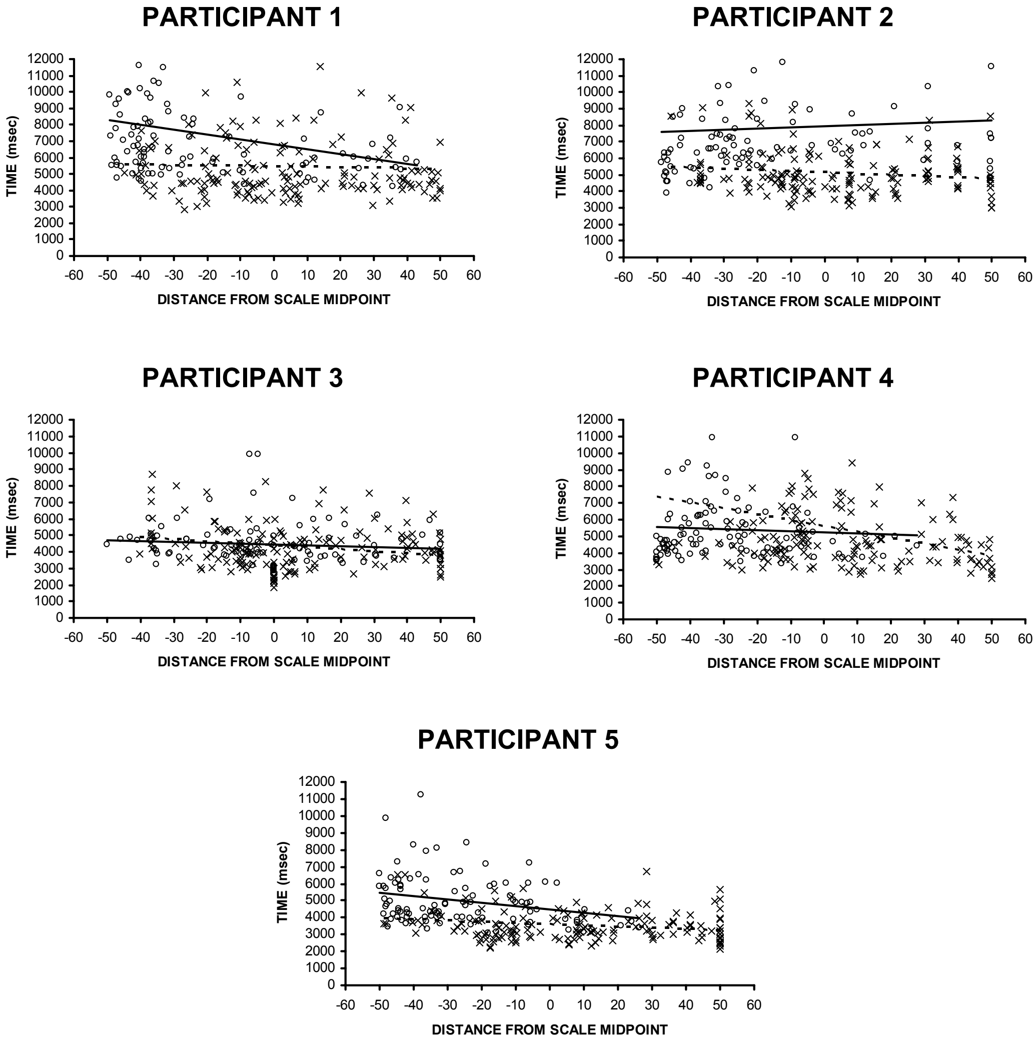
PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 24. Logarithmic scatterplot and regression line for loudness in dynes/cm^2 (P) and participant response (R) for 1000 Hz tones without feedback in Experiment 5.


PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 25. Reaction time for training (×'s with dotted regression line) and testing (○'s with solid regression line) stimuli according to distance from the scale midpoint in Experiment 5.

Table 9. Average reaction times (ms) for each participant for brightness training and loudness testing in Experiment 5.

P	**Training**	**Testing**
1	5435	6479
2	5104	6858
3	4354	3442
4	5431	4400
5	3575	3997
M	4778	5035
SD	803	1535

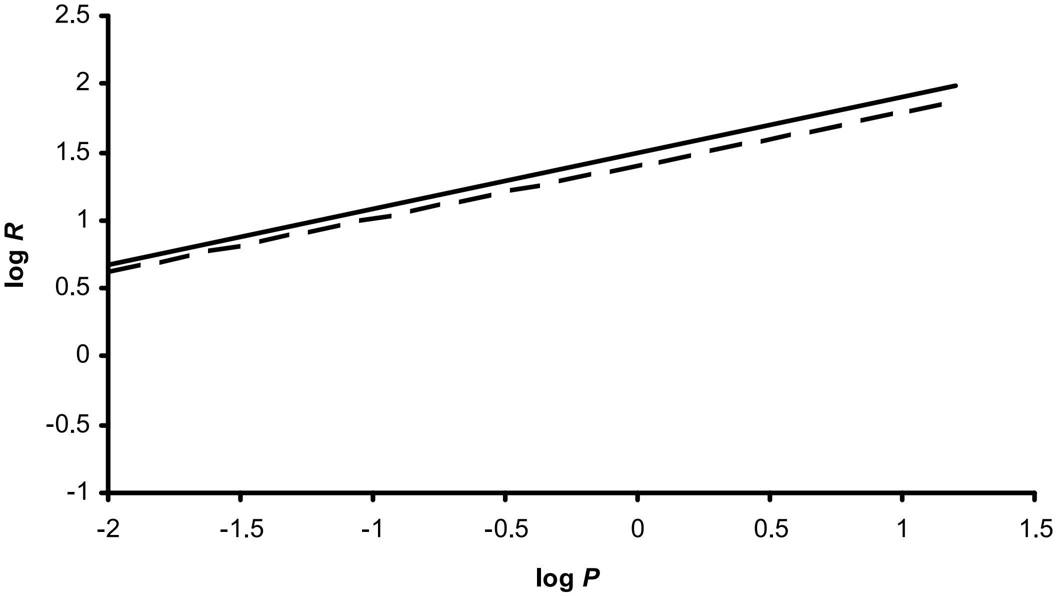


Figure 26. Comparison of constrained loudness scaling for Experiment 5 (solid line) and theoretical loudness with a brightness intercept (dashed line).

Intercept Values

The average scaling intercept value for the brightness stimuli was 1.383, with *SD*/*M* equal to 0.020 and *H*:*L* equal to 1.040:1. These values were comparable to those values found in Experiment 4. For the loudness stimuli, the average scaling intercept was 1.493, with *SD*/*M* equal to 0.112 and *H*:*L* equal to 1.324:1. This intercept value was lower than the intercept found in Experiment 4, although the present experiment featured higher variability. The scaling intercept ratio of brightness to loudness was 0.926, suggesting a strong transference from brightness to loudness.

Goodness of Fit Coefficients

The average goodness of fit coefficient for the brightness stimuli was 0.786, with *SD*/*M* equal to 0.174 and *H*:*L* equal to 1.559:1. This goodness of fit coefficient was 0.105 lower than in Experiment 4, while the variability was considerably higher, where *SD*/*M* was over three times higher. For scaling the loudness stimuli, the average goodness of fit coefficient equaled 0.788, with *SD*/*M* equal to 0.107 and *H*:*L* equal to 1.269:1. These values represented a marginally better goodness of fit and lower variability than loudness scaling in Experiment 4.

Reaction Time

The reaction time data were recorded and analyzed as in Experiment 4. Figure 25 presents reaction time data as a function of the distance from the slider scale midpoint. Table 9 presents the average reaction time data in the present experiment. The average reaction time to select the scaling value was 4778 ms (*SD* = 803) for the brightness stimuli and 5035 ms (*SD* = 1535) for the loudness stimuli. There was no significant difference in scaling times for the brightness training stimuli and the loudness testing stimuli, $t(4) = \pm 0.473$, $p = 0.661$. There was also no significant difference between the reaction times in Experiment 4 and the present experiment [$t(8) = \pm 1.444$, $p = 0.187$ for brightness scaling, and $t(8) = \pm 0.198$, $p = 0.848$ for loudness scaling].

General Discussion

Figure 26 repeats the theoretical loudness scale from Figure 21, in which the dashed line represents the scale that would result from a carryover of the brightness intercept and the maximum subjective loudness scaling value from Experiment 1. The loudness scaling from the present experiment closely follows the slope and intercept of the theoretical line. However, it remains unclear why there was such transference for the flashed brightness stimuli but not for the constant brightness stimuli.

One possible explanation is that light adaptation plays a role in the subjective perception of brightness. Given the longer duration of the brightness stimuli in Experiment 4 than in the present experiment and given the presumed partial dark adaptation in the experimental setting, it is possible that participants may light adapt when the stimuli are maintained on the screen. Light adaptation would have the effect of decreasing the perceived brightness of the grayscale stimuli.[38] Participants would,

[38] Recall that Stevens (1975) found that people generally tend to have larger scaling exponents for briefly flashed brightness stimuli. This effect was documented for stimulus intensities under 1 second (Aiba and Stevens, 1964), which is shorter than the duration of the flashed brightness stimuli in the present experiments. It has been suggested that there is a critical duration under 1 second, over which brightness perception is independent of the stimulus duration (Stevens and Hall, 1966). All brightness stimuli used in the present experiments were presented longer than this critical duration.

nonetheless, learn to associate certain numeric values with the diminished brightness scale due to the training trials. The decrease in perceived brightness coupled with the scale training would have the effect of amplifying the mapping between the perceived intensity of the stimulus and the numeric scale. Participants would need to increase their response values—and, consequently, their intercept or exponent values—to align with the learned scale. When participants are presented with stimuli to which there has been no adaptation, such as loudness stimuli, the expected result would be higher than normal intercept or exponent values.

Participants did, on average, take over 1 second longer to scale brightness in Experiment 4 than in the present experiment (5928 ms for Experiment 4 vs. 4778 ms for Experiment 5). There is nothing in the experimental apparatus that would force participants to take more time scaling brightness in Experiment 4 versus the present experiment. This extra scaling time might therefore be a reflection of extra time spent studying the brightness stimuli, which could have the indirect effect that the increased exposure time could facilitate a moderate degree of light adaptation to the brightness stimuli. Substantial rod light adaptation occurs within a quarter of a second (Adelson, 1982), while cone light adaptation occurs rapidly in the first 3 seconds of exposure to light (Baker, 1949). Both rods and cones continue to light adapt over an 80 second period. Given the short period required for light adaptation, this remains a plausible explanation for the scaling differences encountered in the present experiment for shortened brightness exposure times.

Another possible explanation is that the differences in the results between Experiment 4 and the present experiment may be attributable to separate cognitive processes involved in scaling a live stimulus and scaling a stimulus from memory. The concept of reperception—perceiving something in memory—and its effects on scaling are well documented (for reviews, see Algom, 1992; and Petrusic, Baranski, & Kennedy, 1998). In terms of direct magnitude estimation, the perceived intensity of a remembered stimulus (R') is a power function of the perceived intensity of an actual or live stimulus (R):

$$R' = aR^m. \tag{26}$$

The actual value of the exponent (m) transformation is the subject of discussion. In many cases, R' is less than the original R value, while in other cases R' may be greater than R. The lack of a conclusive, systematic relationship between R' and R has been attributed by some (e.g., Petrusic et al., 1998) to be the byproduct of shortcomings in magnitude estimation. This has led to the use of alternative methods such as similarity comparisons to arrive at the relationship between R' and R.

Given the present data set and experimental design, it is difficult to determine the role that memory might play in the determination of the scaling exponent.[39] It is, however, clear that the memory implicated in the present experiments would not be the type of sustained or long-term memory found in typical experiments on reperception.

[39] To test the effect of memory, it would be necessary to remove the stimulus presentation even for the longer duration exposure in Experiment 4. This would ensure that no experimental condition received the active stimulus during the process of selecting the scale value.

Assuming a robust sensory store and a short-term memory that are capable of holding stimulus magnitude information for several seconds (Atkinson & Shiffrin, 1968), it is doubtful that there would be any discernible difference in stimulus intensity between the flashed and the sustained stimulus, especially if the participant is primed to appraise the stimulus intensity upon first exposure, as in the present experiments. While it is impossible completely to rule out reperception as the determining factor for the scaling differences between Experiments 4 and 5, a more compelling case can be made for the influence of light adaptation on the scaling outcome.

EXPERIMENT 6

Introduction

The previous brightness scaling experiments in this dissertation have addressed achromatic or grayscale stimuli but not color stimuli. Similarly, experiments in the research literature have been conducted to scale the brightness of specific color wavelengths (Judd, 1951; Marks, 1989; Marks & Stevens, 1966; Stevens & Marks, 1980) or to match the perceived brightness of different color wavelengths at a fixed luminance level (Chapanis & Halsey, 1955). However, research has heretofore overlooked the comparative brightness scaling functions of separate color wavelengths.[40]

The present experiment redressed this gap in the research literature by determining the brightness scaling function for red, green, and blue primary luminous stimuli as well as for an achromatic (i.e., grayscale) stimulus. This experiment utilized a magnitude estimation method to determine baseline performance for scaling the brightness of colors. Participants were instructed to rate the brightness of squares flashed on the screen for 1000 ms by using a 100-point slider on the screen. Participants received no feedback on scale usage across the brightness matching trials. Experiment 7, in turn, featured a comparable experimental design within a constrained scaling framework. Like Experiments 4 and 5, the present experiment was classified as an extension of existing scaling methodologies.

[40] Stevens (1975) provides a useful discussion of equal brightness contours but fails to relate this information to exponent values in magnitude estimation experiments.

Method

Participants

Five university students with self-reported normal color vision were enlisted as participants for the experiment. The participants had not previously participated in scaling experiments. The volunteers were remunerated $5.00 for their participation.

Apparatus and Stimulus Materials

As in previous experiments, an experimental control program to display the brightness stimuli and record participant data was developed using Visual Basic 6 running under Windows 2000. The display background was set to a luminous intensity of 0 cd/m^2, and the slider used graphical elements with a maximum luminous intensity of 24 cd/m^2. The colored squares were framed by a border with a luminous intensity of 95 cd/m^2 covering an area equivalent to 4º of the participant's field of vision. Appendix C provides additional details about the experimental apparatus as well as the calibration of the color brightness stimuli.

Design and Procedure

Participants performed the experiment in a dark room illuminated by a single 40 Watt light bulb reflected on a wall behind the participant. Participants were allowed to dark adapt to the room over a 5-minute period prior to the experiment. Subsequently, participants rated the brightness of the squares flashed on the screen. Three iterations of grey, red, green, and blue squares at 14 luminous intensity levels were flashed on the screen for 1000 ms each, resulting in 168 total trials (see Figure 27). The 14 luminous intensity levels were equal across the grayscale and color squares.

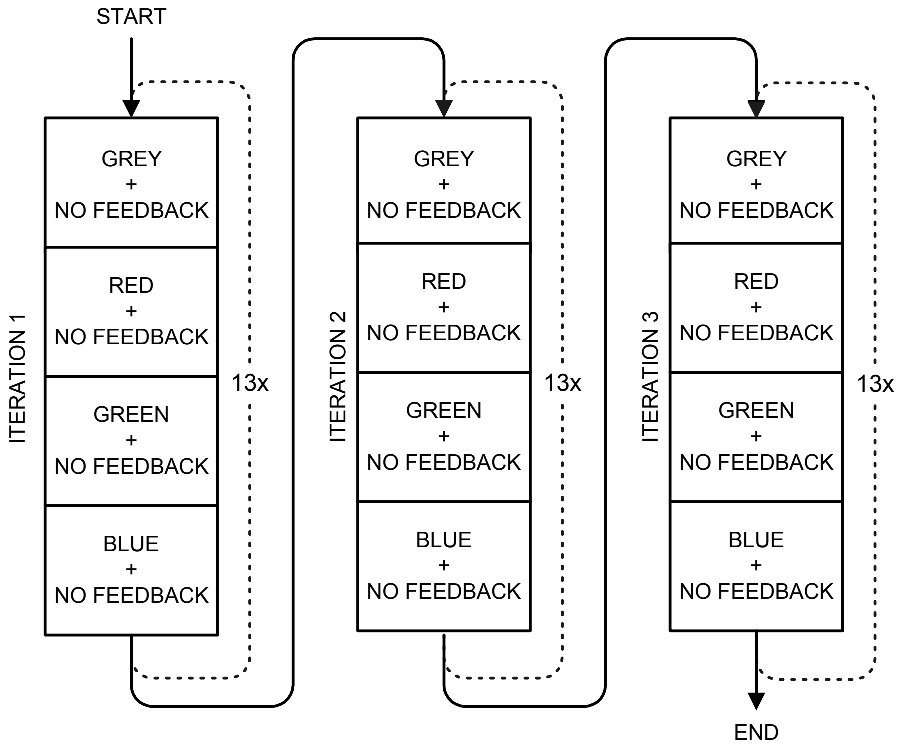


Figure 27. Schematic flow of the design for color brightness magnitude estimation in Experiment 6.

Results and Discussion

The results were analyzed as in Experiments 3 – 5. The mean (*M*), standard deviation (*SD*), coefficient of variation (*SD*/*M*), and highest-to-lowest (*H*:*L*) ratio were calculated for each participant across all colors for the slope of the brightness line (*m*) and the line intercept (log *a*). As a measure of variability, the goodness-of-fit coefficient (R^2) was computed and compared across participants. Tables 10 – 11 summarize the results from the present experiment. Figures 28 – 31 show the individual participant scatterplots and regression lines for the grayscale, red, green, and blue colored squares. For each participant, scores that were two or more standard deviations from the regression line were treated as outliers and were discarded. Reaction time data were not recorded for the present experiment.

Exponent Values

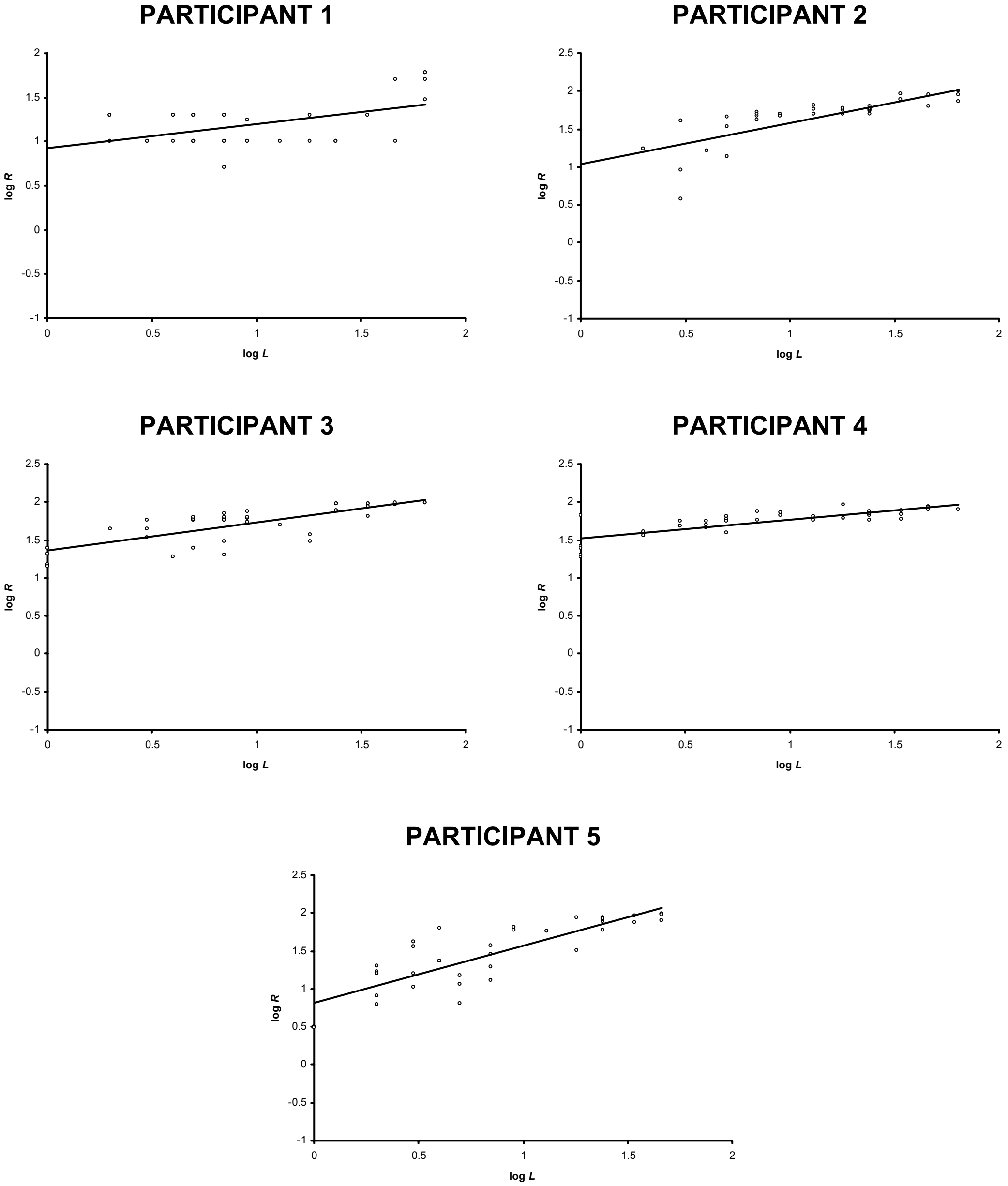
On average, the grayscale squares yielded the largest brightness slope, *M* = 0.437 with *SD*/*M* = 0.485 and *H*:*L* = 3.062:1. This large exponent, indicative of a steep slope, suggesting that participants assigned a broader range of scores between dark and light stimuli for grayscale versus the color stimuli. Red and green squares yielded nearly identical brightness exponents to one another, *M* = 0.363, with *SD*/*M* = 0.544 and *H*:*L* = 3.919:1, and *M* = 0.362 with *SD*/*M* = 0.442 and *H*:*L* = 2.800:1, respectively. Note, however, that red squares produced greater variability in responses than did the green scquares. The blue squares yielded the smallest brightness exponent and the greatest variability, *M* = 0.202 with *SD*/*M* = 0.589 and *H*:*L* = 7.054:1. Using the grayscale

Table 10. Summary of participants' brightness scaling in Experiment 6.

	Grey			Red			Green			Blue		
P	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2
1	0.273	0.929	0.257	0.341	1.428	0.568	0.398	1.236	0.804	0.219	1.563	0.415
2	0.542	1.042	0.596	0.553	1.185	0.692	0.526	1.149	0.671	0.284	1.551	0.797
3	0.368	1.359	0.603	0.198	1.685	0.706	0.202	1.669	0.862	0.119	1.804	0.701
4	0.246	1.524	0.636	0.147	1.729	0.663	0.188	1.666	0.666	0.048	1.835	0.127
5	0.755	0.817	0.724	0.577	1.103	0.653	0.497	1.191	0.730	0.339	1.480	0.630
M	0.437	1.134	0.563	0.363	1.426	0.657	0.362	1.382	0.747	0.202	1.647	0.534
SD	0.212	0.298	0.178	0.198	0.283	0.054	0.160	0.262	0.085	0.119	0.161	0.267

Table 11. Scaling ratios of grayscale stimuli to color stimuli in Experiment 6.

		Red			**Green**			**Blue**	
P	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***
1	0.802	0.651	0.453	0.687	0.752	0.320	1.246	0.752	0.619
2	0.981	0.880	0.861	1.031	0.907	0.888	1.913	0.907	0.748
3	1.859	0.807	0.853	1.823	0.815	0.699	3.105	0.815	0.859
4	1.673	0.882	0.959	1.311	0.915	0.955	5.123	0.915	4.999
5	1.307	0.741	1.108	1.519	0.686	0.992	2.224	0.686	1.150
M	1.324	0.792	0.847	1.274	0.815	0.771	2.722	0.815	1.675
SD	0.447	0.098	0.243	0.438	0.099	0.276	1.499	0.099	1.868



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 28. Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for grayscale squares in Experiment 6.

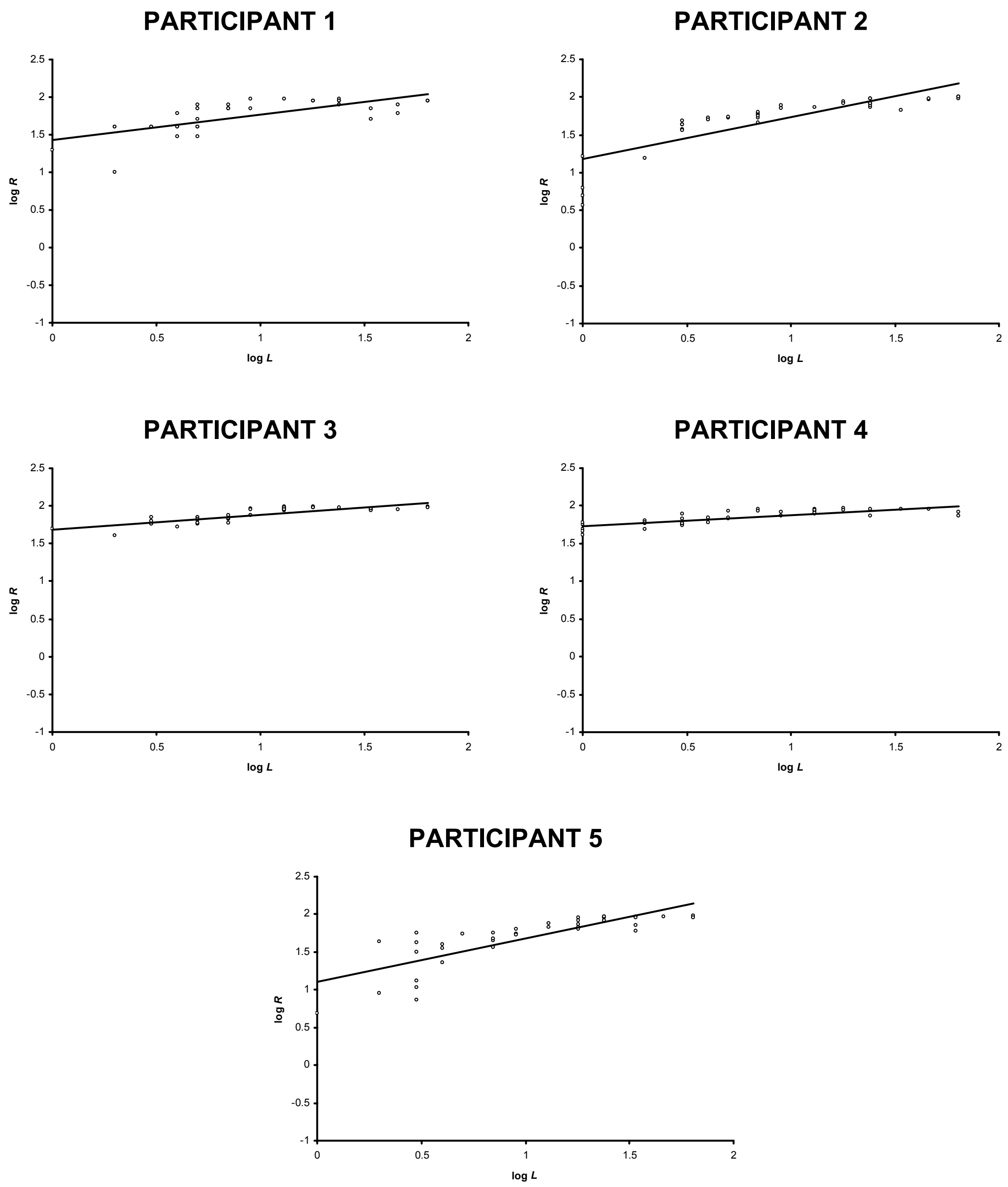
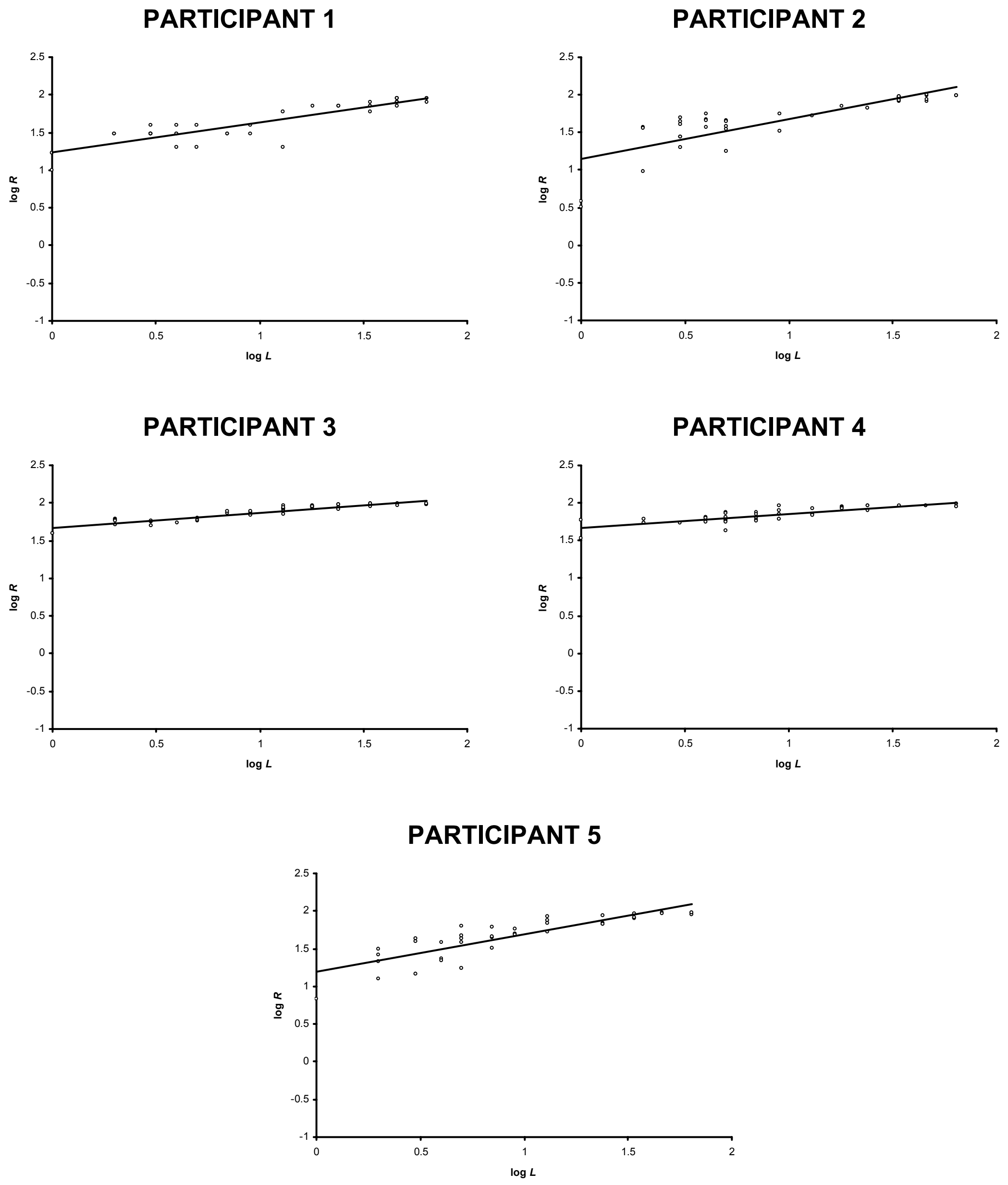


*Figure 29. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for red squares in Experiment 6.*


PARTICIPANT 1

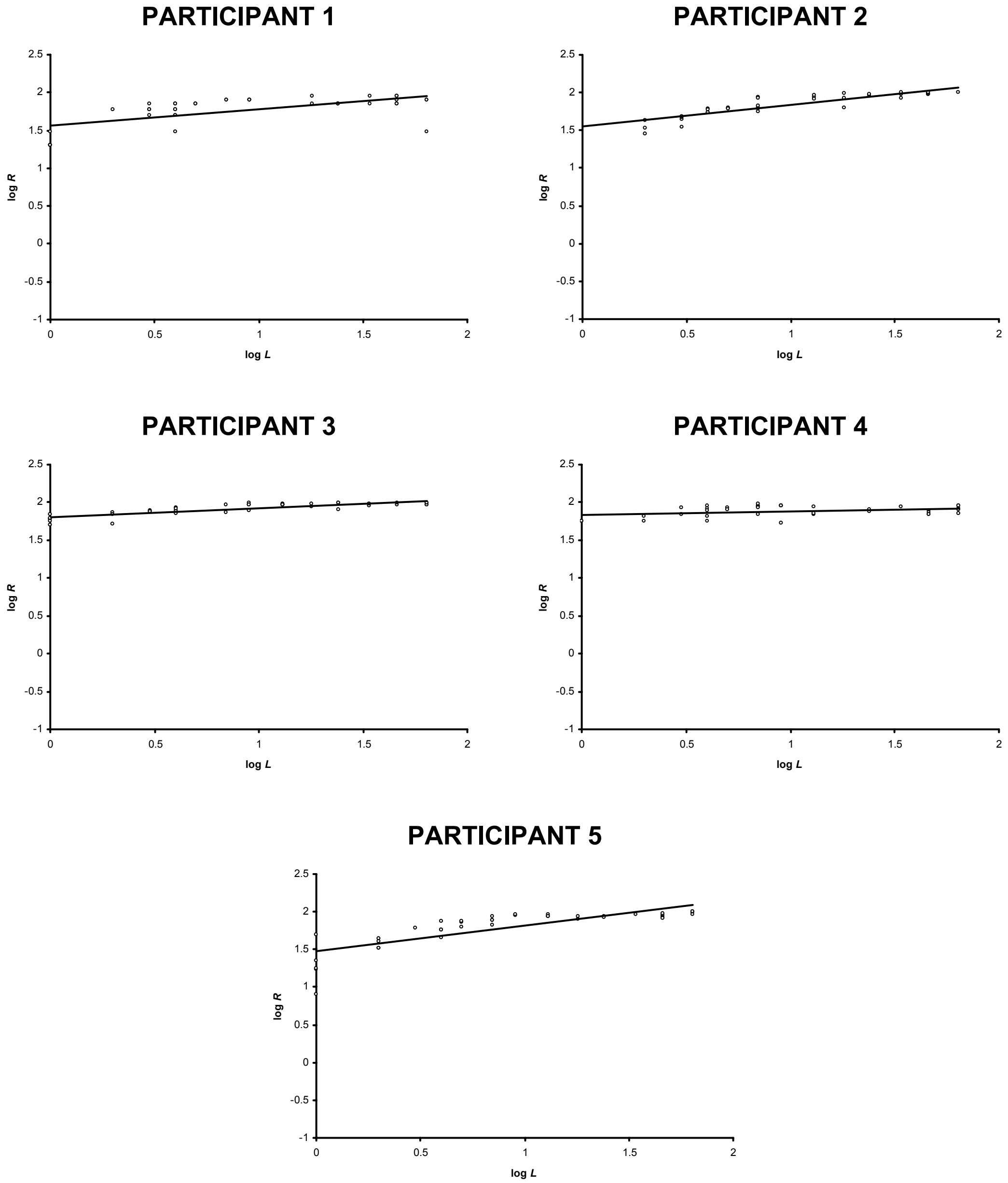
PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 30. Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for green squares in Experiment 6.



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 31. Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for blue squares in Experiment 6.

exponent as a baseline, the ratio of the average brightness exponent for grayscale squares to red squares was 1.324. For green squares, this ratio was 1.274 In the case of the blue squares, the average brightness exponent was over twice the average brightness exponent for grayscale squares, with a grayscale-to-blue exponent-to-exponent ratio equal to 2.722.

Intercept Values

The intercepts for the color brightness lines followed a pattern in which greater exponents produced lower intercepts. The grayscale brightness intercept was the smallest at M = 1.134 with SD/M = 0.262 and $H{:}L$ = 1.867:1. The grayscale brightness intercept was followed by red and green brightness intercepts at M = 1.426 with SD/M = 0.199 and $H{:}L$ = 1.568:1, and M = 1.382, SD/M = 0.190 and $H{:}L$ = 1.453:1, respectively. The average blue brightness intercept was the largest at M = 1.647 with SD/M = 0.098 and $H{:}L$ = 1.24:1. The average intercept scaling ratio of grayscale to color stimuli was 0.792 for red stimuli, 0.815 for green stimuli, and 0.680 for blue stimuli.

Goodness of Fit Coefficients

Variability as measured by R^2 revealed the best average stimulus-to-response fit for green stimuli, R^2 = 0.747 with SD/M = 0.114 and $H{:}L$ = 1.294:1. The red stimuli produced the second best goodness of fit, with R^2 = 0.657 and SD/M = 0.082 and $H{:}L$ = 1.243:1. Grayscale stimuli produced an R^2 equal to 0.563 with SD/M = 0.317 and $H{:}L$ = 2.813:1. Blue stimuli had an R^2 equal to 0.534 with SD/M = 0.500 and $H{:}L$ = 6.265:1. The average R^2 scaling ratio from the grayscale to the color stimuli was 0.847 for red stimuli, 0.771 for green stimuli, and 1.675 for blue stimuli.

General Discussion

The goodness of fit coefficients for the color scaling exponents were considerably larger than those reported for magnitude estimation and cross-modality scaling in West et al. (2000). In West et al., the average goodness of fit coefficient for the 14 cited studies was 0.333, whereas the present experiment produced *SD*/*M* values ranging from 0.437 for grayscale stimuli to 0.589 for blue stimuli. The *H*:*L* ratios revealed similarly high levels of variability for scaling the brightness of colors, ranging from 2.800:1 for green stimuli to 7.054:1 for blue stimuli. The average *H*:*L* ratio was 2.995:1 in West et al. The high *SD*/*M* and, in part, *H*:*L* values suggest that color brightness scaling is particularly susceptible to scaling biases and inconsistencies. As such, it is an ideal test bed for the application of constrained scaling methodology. The next experiment addresses this possibility.

EXPERIMENT 7

Introduction

Experiment 6 revealed a high level of variabily for scaling the brightness of colors. This variability represents an ideal set of circumstances for testing the tenet that constrained scaling significantly reduces scaling variability compared to magnitude estimation (or matching) methods. The present experiment was identical to the previous experiment with the exception that the participants now received initial scale training in accordance with constrained scaling methodology.

Method

Participants

As in Experiment 6, five university students with self-reported normal color vision were enlisted as participants for the experiment and were paid $5.00 for volunteering. The participants had not previously participated in brightness scaling experiments.

Apparatus and Stimulus Materials

The apparatus and stimulus materials were identical to Experiment 6, except the experimental control software provided scaling feedback for grayscale stimuli.

Design and Procedure

The design was identical to Experiment 6, except participants were trained on the brightness of the grayscale squares (see Figure 32). Participants first completed 3 iterations of training on the 14 luminous intensity values for the grayscale squares. After

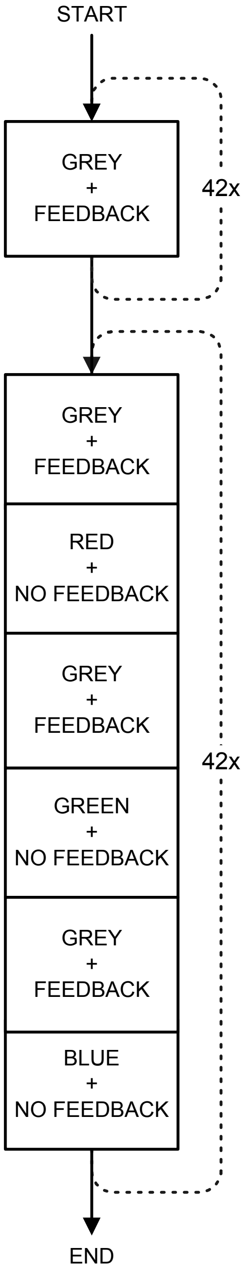


Figure 32. Schematic flow of the design for constrained scaling of color brightness in Experiment 7.

rating the brightness of each grayscale square, participants received feedback on the actual brightness of the square according to the following equation:

$$R = 12.6L^{0.33}, \tag{27}$$

where R was the response or feedback value and L was the stimulus luminance in cd/m^2. As in Experiments 4 and 5, the exponent value of 0.33 was used to represent Stevens' (1975; Marks & Stevens, 1966) measured exponent for the brightness of a 5° target in the dark. The constant, 12.6, was selected such that the maximum feedback value was approximately 50 for the brightest stimulus of 64 cd/m^2.[41] Following completion of the initial 42 training trials, participants completed 3 iterations of the 14 luminance intensities for red, green, and blue squares. Participants received no feedback for their ratings of the colored squares. A grayscale square with feedback was interspersed between each colored square in order to maintain scale learning.

Results and Discussion

The results were analyzed as in Experiment 6 and are summarized in Tables 12 – 13 and Figures 33 – 36. A comparison of the results from Experiment 6 and the present experiment is presented later in this chapter.

Exponent Values

The mean brightness exponent for the grayscale squares with feedback was 0.288 with SD/M = 0.090 and H:L = 1.248:1. For the red squares without feedback, the mean

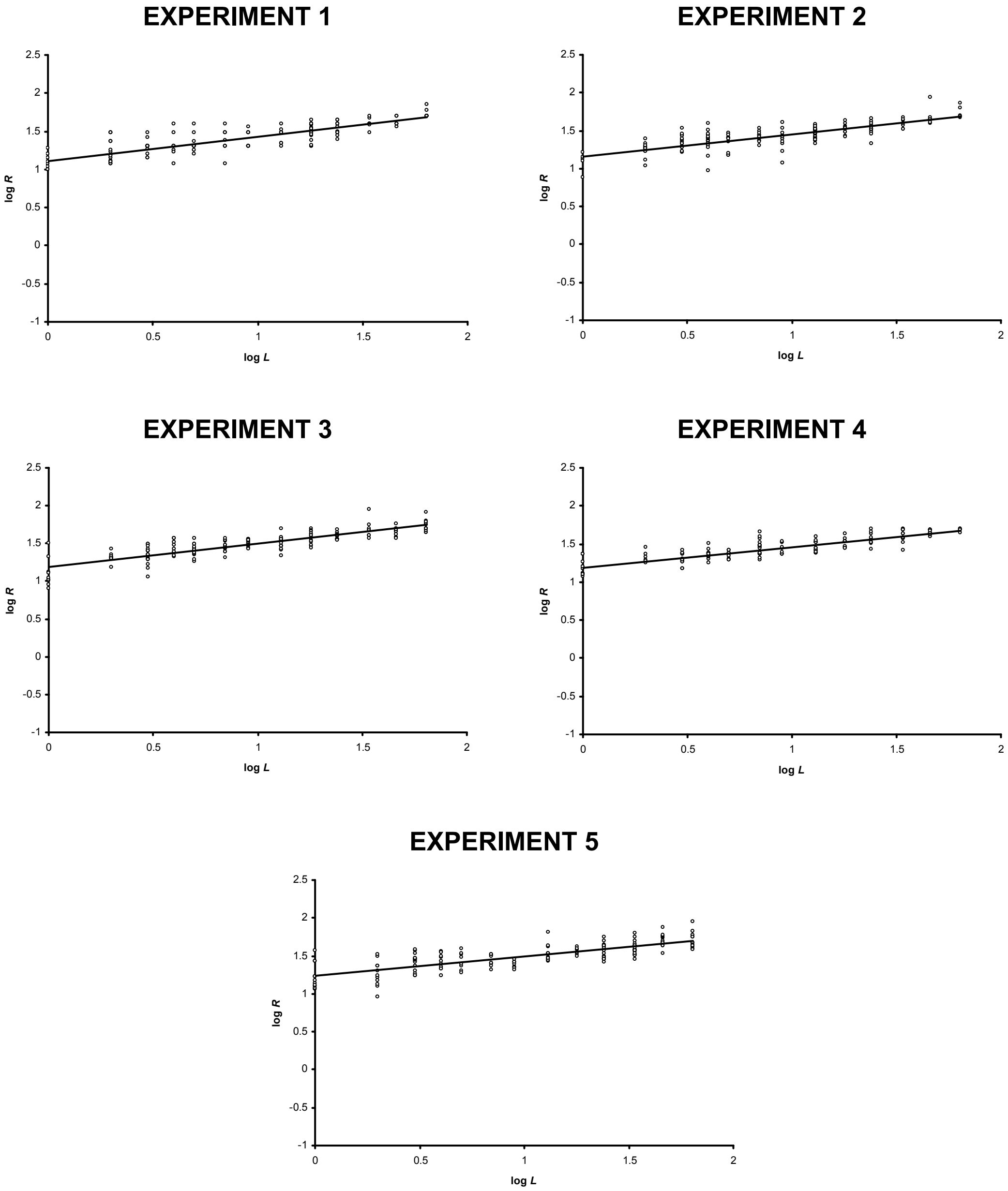
[41] Because the red, green, and blue stimuli could not be produced with enough brightness to match the maximum brightness levels for the grayscale stimuli, the training scale in the present experiment is different than the training scale used in Experiments 4 and 5.

Table 12. Summary of participants' brightness scaling in Experiment 7.

	Grey			**Red**			**Green**			**Blue**		
P	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2
1	0.318	1.111	0.739	0.335	1.173	0.702	0.443	1.011	0.828	0.313	1.285	0.644
2	0.291	1.159	0.646	0.283	1.250	0.666	0.256	1.228	0.785	0.144	1.441	0.506
3	0.306	1.187	0.736	0.404	1.123	0.768	0.376	1.211	0.866	0.255	1.400	0.498
4	0.270	1.189	0.778	0.319	1.185	0.844	0.368	1.073	0.836	0.279	1.280	0.841
5	0.255	1.237	0.613	0.289	1.398	0.684	0.374	1.247	0.826	0.234	1.484	0.519
M	0.288	1.177	0.702	0.326	1.226	0.733	0.363	1.154	0.828	0.245	1.376	0.601
SD	0.026	0.046	0.069	0.048	0.107	0.073	0.067	0.105	0.029	0.064	0.092	0.146

Table 13. Scaling ratios of grayscale stimuli to color stimuli in Experiment 7.

		Red			**Green**			**Blue**	
P	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2
1	0.950	0.948	1.054	0.718	1.098	0.893	1.016	0.865	1.148
2	1.029	0.927	0.970	1.138	0.944	0.823	2.021	0.804	1.277
3	0.758	1.058	0.958	0.814	0.980	0.850	1.201	0.854	1.477
4	0.845	1.003	0.922	0.734	1.108	0.931	0.966	0.929	0.925
5	0.880	0.884	0.897	0.681	0.992	0.742	1.089	0.833	1.183
M	0.892	0.964	0.960	0.817	1.025	0.848	1.258	0.857	1.202
SD	0.103	0.068	0.060	0.186	0.074	0.072	0.435	0.046	0.201



EXPERIMENT 1

EXPERIMENT 2

EXPERIMENT 3

EXPERIMENT 4

EXPERIMENT 5

*Figure 33. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for grayscale squares with feedback in Experiment 7.*

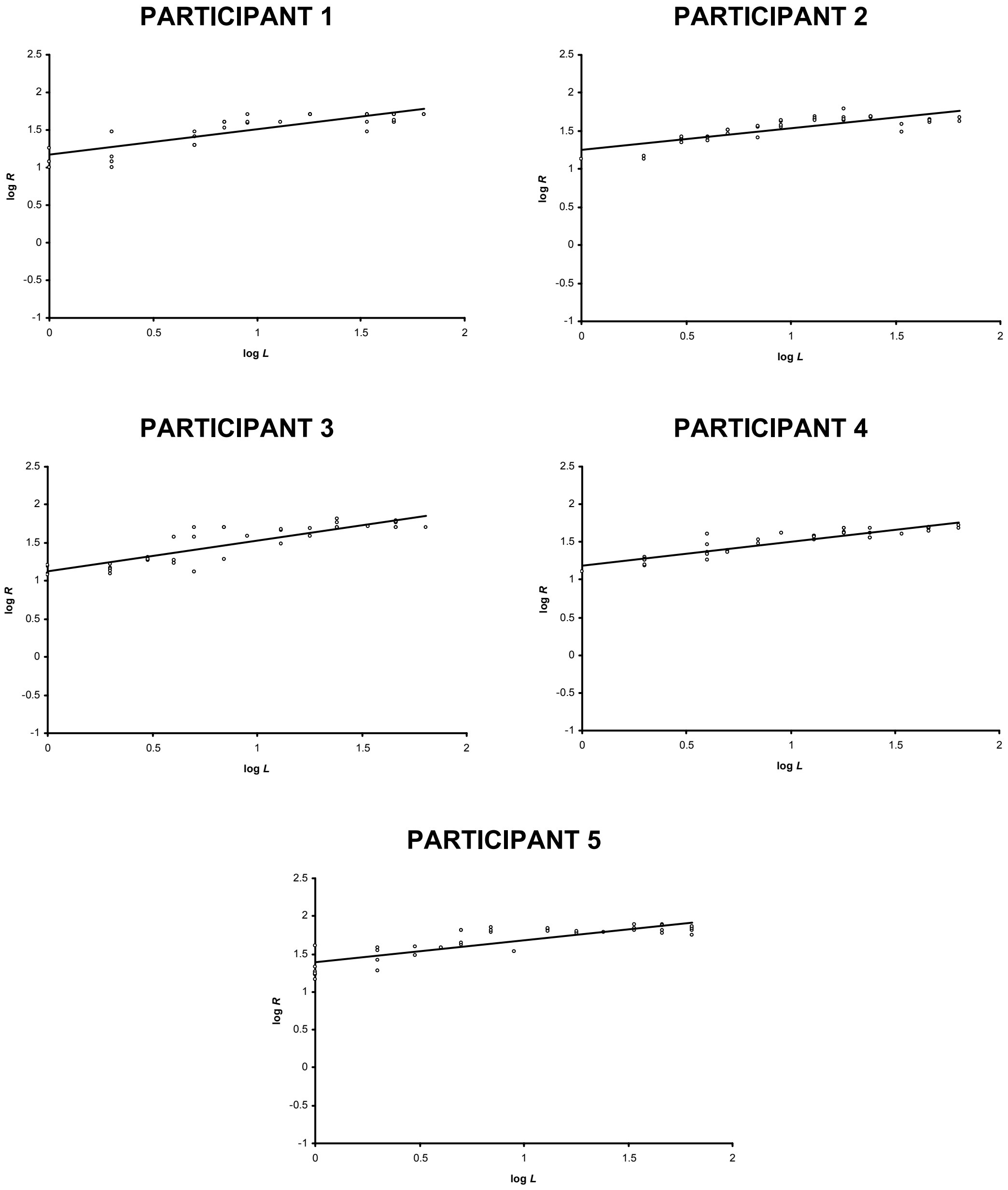


Figure 34. *Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for red squares without feedback in Experiment 7.*

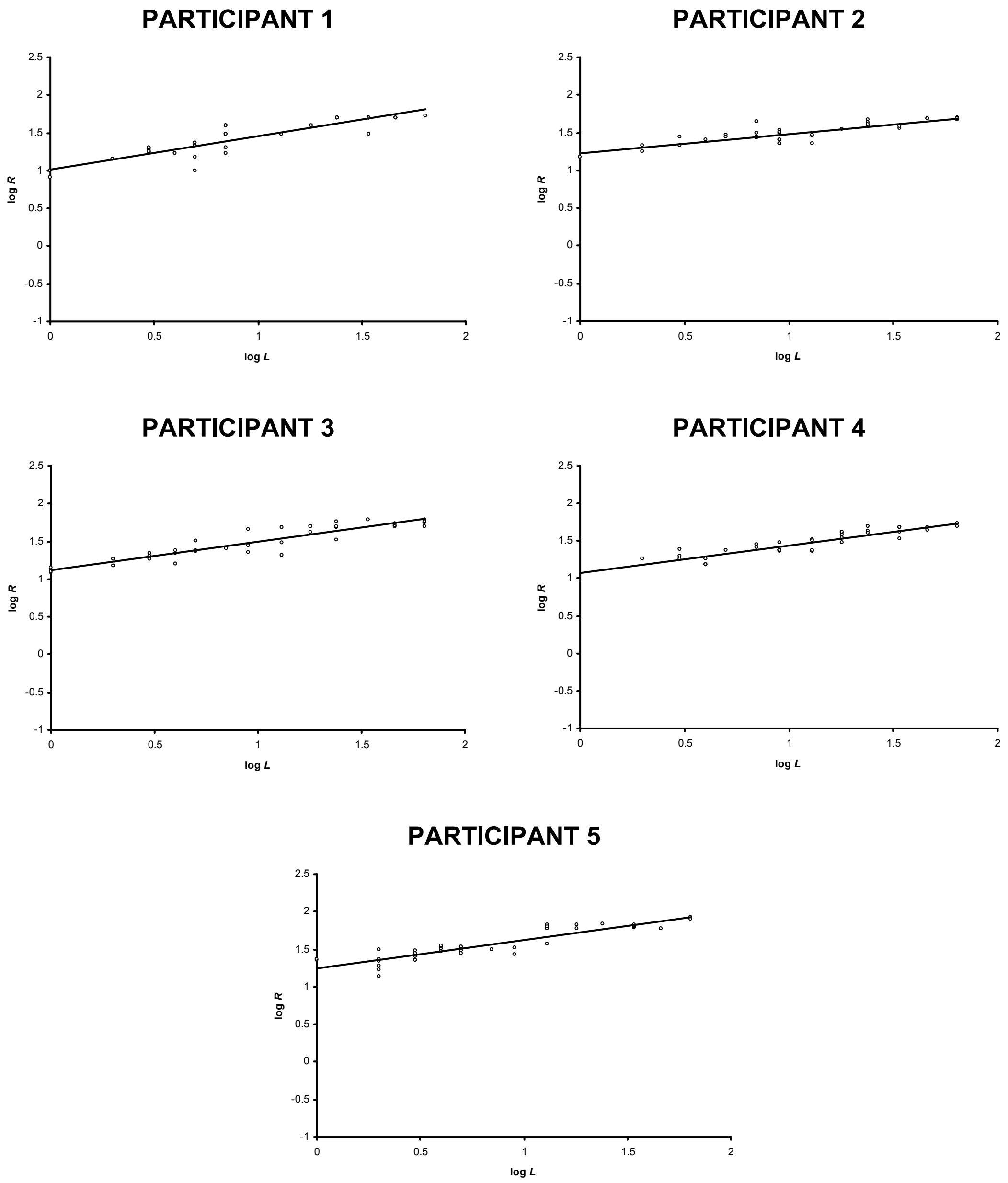


Figure 35. *Logarithmic scatterplot and regression line for brightness in cd/m*^{2} *(*L*) and participant response (*R*) for green squares without feedback in Experiment 7.*

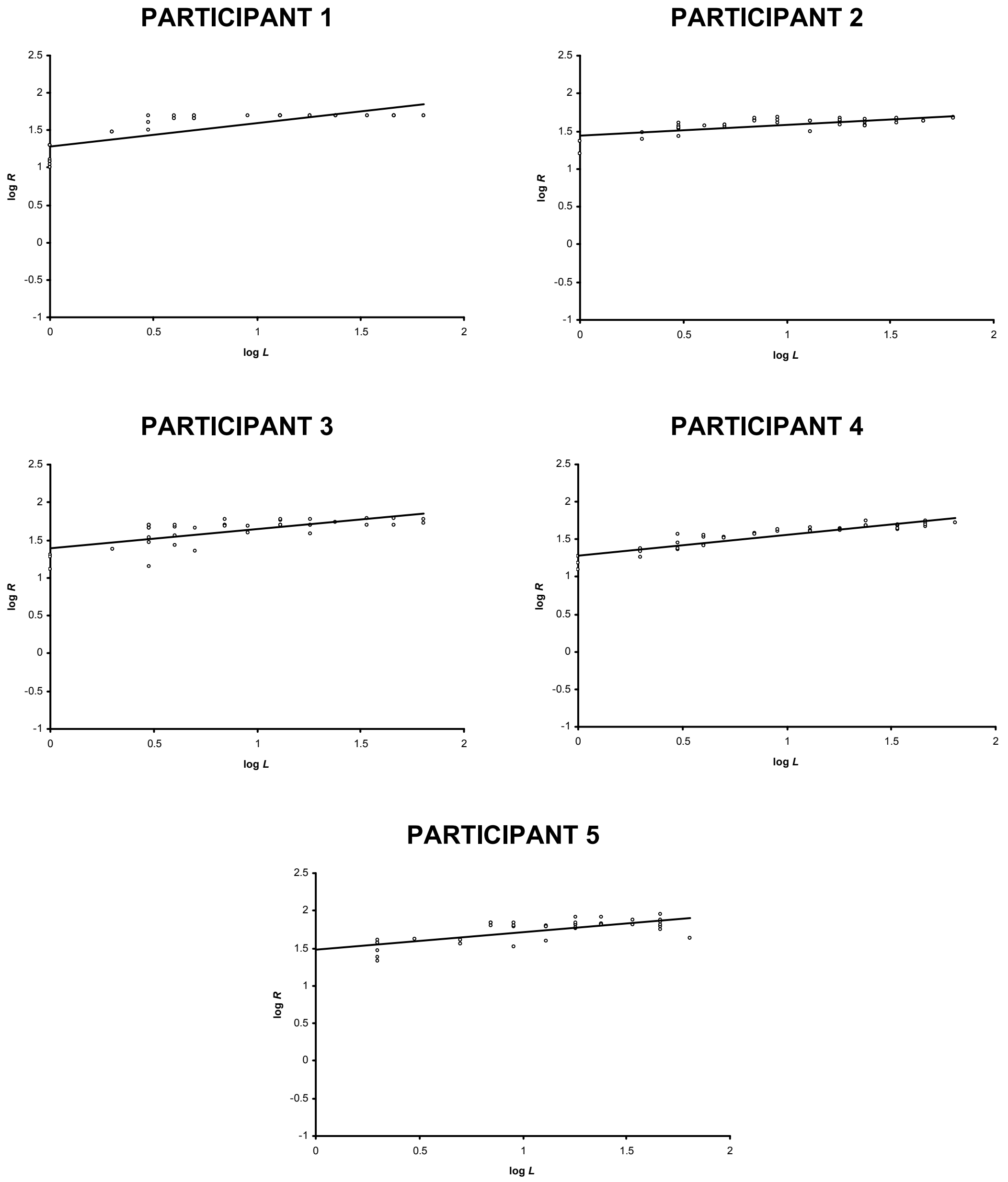


*Figure 36. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for blue squares without feedback in Experiment 7.*

exponent equaled 0.326 with *SD*/*M* = 0.149 and *H*:*L* = 1.428:1; the mean exponent for the green squares equaled 0.363 with *SD*/*M* = 0.185 and *H*:*L* = 1.428:1; the mean exponent for the blue squares equaled 0.245 with *SD*/*M* = 0.260 and *H*:*L* = 2.171:1. The average ratio of grayscale to red brightness scaling exponents was 0.892; for grayscale to green exponents, it was 0.817; and for grayscale to blue exponents, it was 1.259.

Intercept Values

The average brightness intercept for grayscale squares was 1.177 with *SD*/*M* = 0.039 and *H*:*L* = 1.113:1; for red squares, it was 1.226 with *SD*/*M* = 0.087 and *H*:*L* = 1.246:1; for green squares, it was 1.154 with *SD*/*M* = 0.091 and *H*:*L* = 1.232:1; and for blue squares, it was 1.376 with *SD*/*M* = 0.067 and *H*:*L* = 1.159:1. The average ratio of grayscale to red scaling intercepts was 0.964; for grayscale to green intercepts, it was 1.025; and for grayscale to blue intercepts, it was 0.857.

Goodness of Fit Coefficients

For the goodness of fit of the regression line across participants, scaling of grayscale squares with feedback produced an average R^2 value equal to 0.702 with *SD*/*M* = 0.099 and *H*:*L* = 1.268:1; red squares without feedback produced an average R^2 value equal to 0.733 with *SD*/*M* = 0.100 and *H*:*L* = 1.267:1; green squares produced an average R^2 value equal to 0.828 with *SD*/*M* = 0.035 and *H*:*L* = 1.103:1; blue squares produced an average R^2 value equal to 0.601 with *SD*/*M* = 0.243 and *H*:*L* = 1.6881. The average ratio of grayscale to red R^2 values was 0.960; for grayscale to green R^2 values, it was 0.848; and for grayscale to blue R^2 values, it was 1.202.

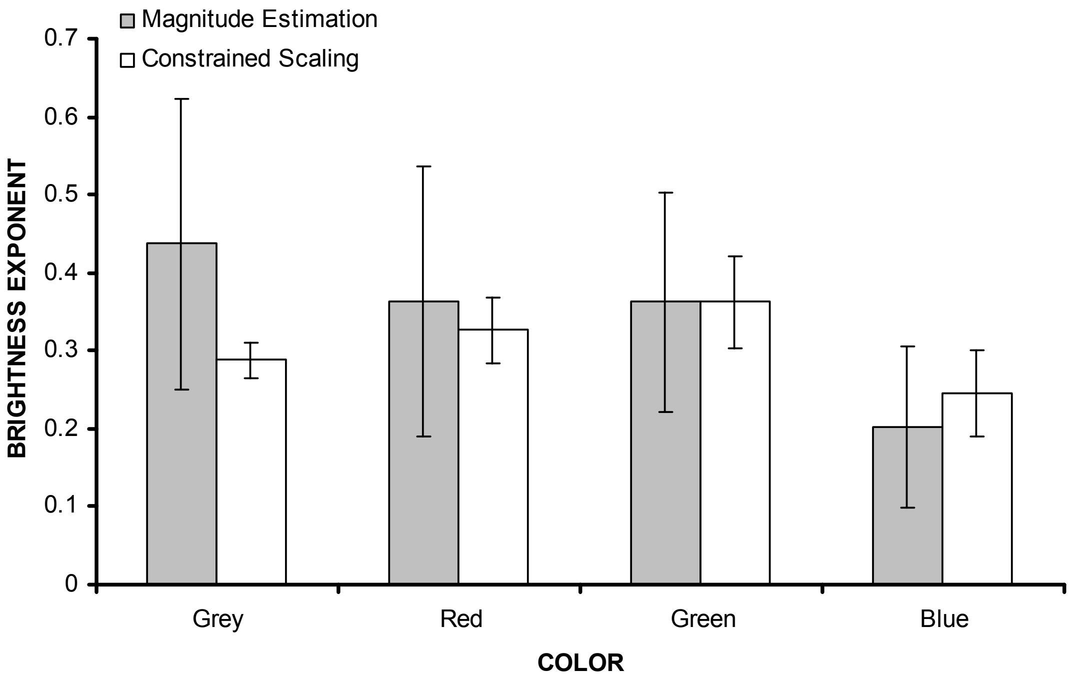
Comparison of Experiments 6 and 7

Whereas Experiment 6 used a conventional magnitude estimation method to scale the brightness of colors, Experiment 7 used an implementation of the constrained scaling methodology. A close comparison of the two experiments (see Figures 37 – 39) reveals that constrained scaling of brightness significantly reduced variability compared to magnitude estimation.

Exponent Values

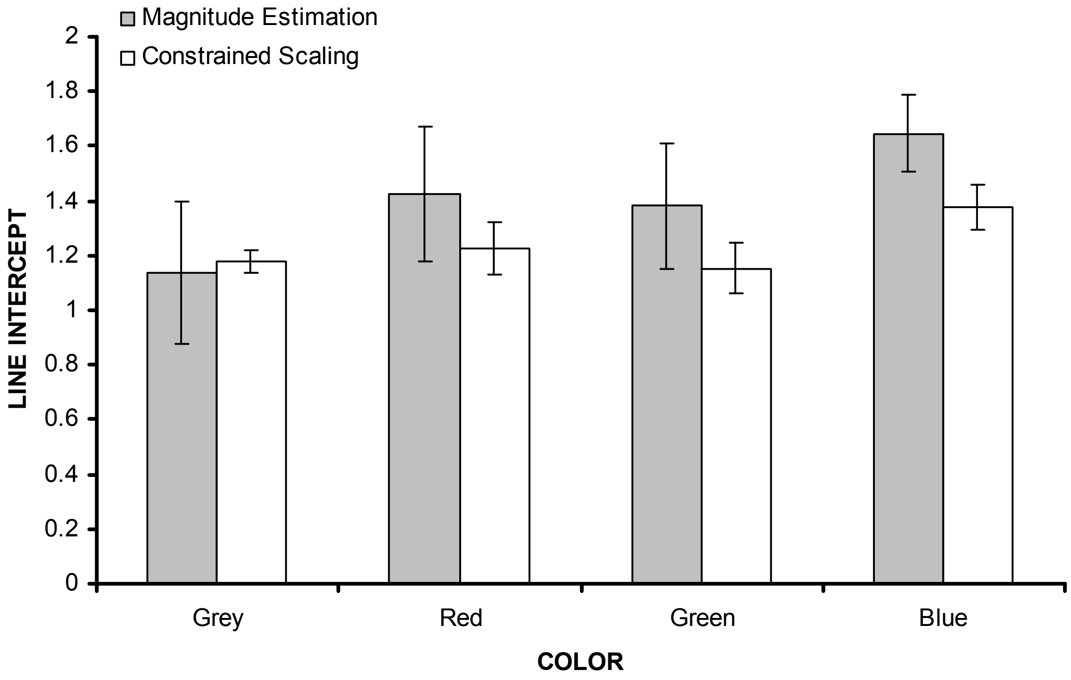
Figure 37 contrasts the average exponent values for the brightness stimuli across colors between magnitude estimation and constrained scaling. The magnitude estimation participants exhibited a marginally higher average grayscale exponent than the constrained scaling participants, $t(8) = 1.561$, $p = 0.079$. Magnitude estimation revealed an average grayscale slope of 0.437, whereas constrained scaling revealed an average slope of 0.288. This finding suggests that the appropriate "natural" exponent value for scaling the brightness of the achromatic squares may be higher than the exponent value of 0.33 adopted from Stevens (1975) and that the value may, in fact, be closer to 0.45 as suggested in J.C. Stevens and Hall (1966). However, magnitude estimation had significantly higher R^2 variability than constrained scaling for scaling the brightness of grayscale squares, $F(1,8) = 67.193$, $p < 0.001$. Given the high level of variability in magnitude estimation, it would be premature to conclude that the "natural" exponent value for brightness scaling should be higher than that prescribed by Stevens.

For the color squares, there was no significant difference in the mean brightness exponent values between magnitude estimation and constrained scaling. Red brightness



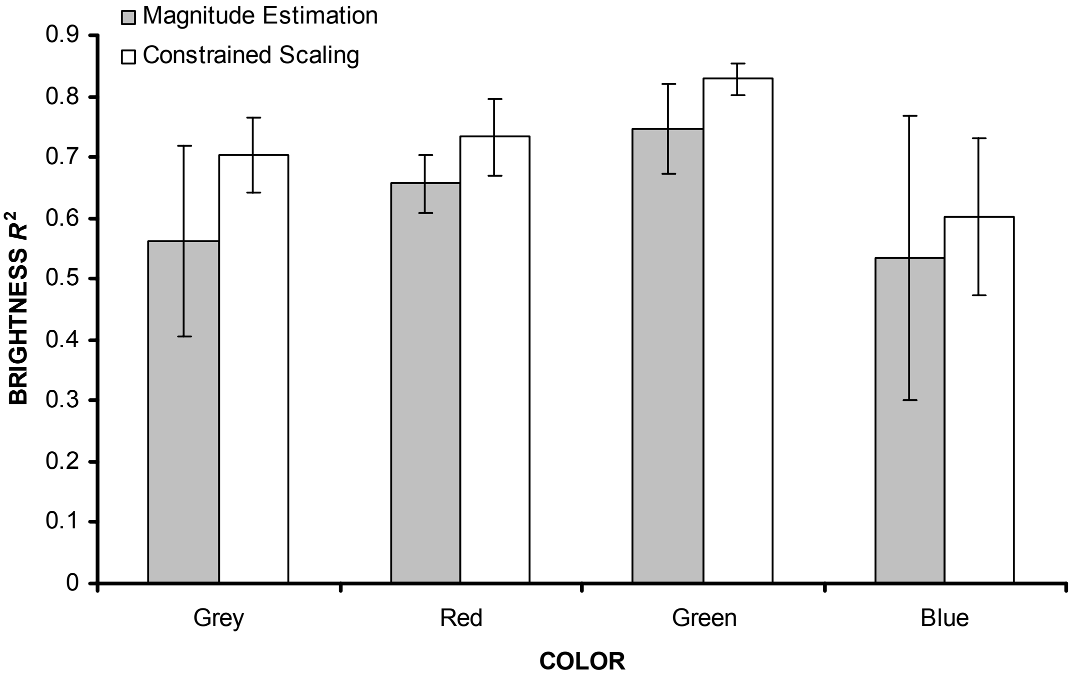
NOTE: The error bars indicate the 95% confidence intervals.

Figure 37. Comparison of brightness exponents between Experiment 6 (magnitude estimation) and Experiment 7 (constrained scaling).



NOTE: The error bars indicate the 95% confidence intervals.

Figure 38. Comparison of brightness line intercepts between Experiment 6 (magnitude estimation) and Experiment 7 (constrained scaling).



NOTE: The error bars indicate the 95% confidence intervals.

*Figure 39. Comparison of brightness goodness of fit coefficients (*R^2*) between Experiment 6 (magnitude estimation) and Experiment 7 (constrained scaling).*

scaling had a *t* value comparing magnitude estimation and constrained scaling equal to 0.410, $p = 0.346$; for green brightness, $t(8)$ equaled -0.015, p = 0.494; and, for blue brightness, $t(8)$ equaled -0.717, $p = 0.247$.

Although the mean brightness exponent values were comparable across colors, constrained scaling exhibited generally lower variability than magnitude estimation as measured by a statistical *F* test of variance between the methods. There was significantly lower variance in the constrained scaling condition for grayscale and red stimuli. For grayscale squares, $F(1,8) = 67.193$, $p < 0.001$; for red squares, $F(1,8) = 16.622$, $p < 0.01$. Variance was marginally reduced in the constrained scaling condition for green and blue stimuli. For green squares, $F(1,8) = 5.654$, $p = 0.061$; and, for blue squares, $F(1,8) = 3.492$, $p = 0.127$.

The generally lower variability for the constrained scaling of brightness is further demonstrated by comparing the coefficients of variation (*SD*/*M*). For the grayscale brightness exponents, *SD*/*M* is 5.389 times higher for magnitude estimation than for constrained scaling, 0.485 vs. 0.090, respectively; for the red brightness exponents, *SD*/*M* is 3.651 times higher for magnitude estimation than for constrained scaling, 0.544 vs. 0.149; for the green brightness exponents, *SD*/*M* is 2.389 times higher for magnitude estimation than for constrained scaling, 0.442 vs. 0.185; for the blue brightness exponents, *SD*/*M* is 2.274 times higher for magnitude estimation than for constrained scaling, 0.589 vs. 0.260.

A similar decrease in variability from magnitude estimation to constrained scaling is found in the highest-to-lowest (*H*:*L*) exponent ratios. For the grayscale brightness exponents, *H*:*L* is 2.454 times higher for magnitude estimation than for constrained scaling, 3.062:1 vs. 1.248:1, respectively; for red brightness exponents, *H*:*L* is 2.744 times higher for magnitude estimation than for constrained scaling, 3.919:1 vs. 1.428:1; for green brightness exponents, *H*:*L* is 1.619 times higher for magnitude estimation than for constrained scaling, 2.800:1 vs. 1.729:1; for blue brightness exponents, *H*:*L* is 3.249 times higher for magnitude estimation than for constrained scaling, 7.054:1 vs. 2.171:1.

Intercept Values

Figure 38 shows the relationship between the regression line intercepts across colors for magnitude estimation and constrained scaling. While magnitude estimation and constrained scaling methods produced nearly identical intercepts for grayscale squares, $t(8) = -0.314$, $p = 0.381$, constrained scaling produced lower intercept values for colored squares. For the red squares, $t(8) = 1.476$, $p = 0.089$; for the green squares, $t(8) = 1.805$, $p = 0.054$; for the blue squares, $t(8) = 3.261$, $p < 0.01$.

Not only did constrained scaling result in lower brightness intercept values on average compared to magnitude estimation, it also resulted in generally lower scaling variance. The constrained scaling intercepts had significantly lower variance than the magnitude estimation intercepts for grayscale brightness, $F(1,8) = 41.872$, $p < 0.01$; for red brightness, $F(1,8) = 7.070$, $p < 0.05$; and for green brightness, $F(1,8) = 6.233$, $p = 0.052$. The difference in variance of the intercepts for scaling the brightness of blue and grayscale squares was nonsignificant, $F(1,8) = 3.098$, $p = 0.150$.

This pattern was confirmed by comparing the coefficients of variation for the brightness intercepts across colors between magnitude estimation and constrained scaling. For the grayscale brightness intercepts, the *SD*/*M* was 6.718 times higher for magnitude estimation than for constrained scaling, 0.262 vs. 0.039, respectively; for the red brightness intercepts, the *SD*/*M* was 2.287 times higher for magnitude estimation than for constrained scaling, 0.199 vs. 0.087; for the green brightness intercepts, the *SD*/*M* was 2.088 times higher for magnitude estimation than for constrained scaling, 0.190 vs. 0.091; for the blue brightness intercepts, the *SD*/*M* was 1.463 times higher for magnitude matching than for constrained scaling, 0.098 vs. 0.067.

To a negligible extent, there was a reduction in the highest-to-lowest ratios for the brightness intercepts between magnitude estimation and constrained scaling. The grayscale *H*:*L* ratio of the brightness intercept was 1.677 times higher for magnitude estimation than for constrained scaling, 1.867:1 vs. 1.113:1, respectively; for red brightness, it was 1.258 times higher, 1.568:1 vs. 1.246:1; for green brightness, it was 1.179 times higher, 1.453:1 vs. 1.232:1; for blue brightness, it was 1.070 times higher, 1.240:1 vs. 1.159.

Goodness of Fit Coefficients

Figure 39 shows the relationship of the average goodness-of-fit coefficient, R^2, between magnitude estimation and constrained scaling. The R^2 values demonstrate that constrained scaling resulted in lower intraparticipant variability for scaling the brightness of most colors in the stimulus gamut. The regression line for grayscale brightness was a marginally better fitted line for constrained scaling than magnitude estimation, $t(8) = -1.627$, $p = 0.071$. For both red and green brightness, constrained scaling resulted in a significantly better fitting line than did magnitude estimation, $t(8) = -1.869$, $p < 0.05$, and $t(8)= – 2.026$, $p < 0.05$, respectively. In contrast, the regression line for scaling of blue brightness was not significantly different in terms of goodness of fit, $t(8)= –0.494$, $p = 0.317$. No F-tests of variance were performed on the R^2 values, because the R^2 values were already inherently measures of intraparticipant variance.

General Discussion

Comparing the results of Experiments 6 and 7, it is clear that constrained scaling succeeded in offering a clear alternative to magnitude estimation for the scaling of brightness. In general, constrained scaling reduced the variability associated with scaling the brightness of colors.

While the results clearly point to constrained scaling as a method for reliable brightness color scaling, they also verify that the efficacy of constrained scaling is not simply an artifact of loudness scaling. The findings from Experiment 7 demonstrate that scaling brightness can produce coefficients of variation and highest-to-lowest exponent ratios comparable to the low levels achieved for scaling loudness using constrained scaling.

The decrease in variability afforded by constrained scaling was not apparent for scaling the brightness of blue colored squares. The reason for this may be attributable to the low luminous intensity of the blue phosphor gun in the CRT. Table C-4 (see Appendix C) shows that the maximum luminous intensity that is possible using solely the blue phosphor gun (i.e., RGB = [0, 0, 255])was approximately 9 cd/m^2). This luminous intensity level contrasts with the much more luminous red and green phosphor guns (see Tables C-2 and C-3), which are individually capable of approximately 20 cd/m^2 (RGB =

[255, 0, 0]) and 60 cd/m^2 (*RGB* = [0, 255, 0]). As the maximum luminous potential of the single phosphor gun is reached, it becomes necessary to add achromatic color to boost the luminous intensity of the color. In adding achromatic color to the blue color signal, the color undergoes a considerable shift in its *Yxy* chromaticity coordinates. Shifting the chromaticity coordinates may have resulted in shifts of saturation or hue, which have been demonstrated to have their own scaling functions (Panek & Stevens, 1966; Indow & Stevens, 1966). While red and green also required achromatic color additions to reach the desired brightest stimulus value of 64 cd/m^2, the red and green pure color signals were brighter, requiring a much smaller overall addition of achromatic color. Consequently, there was a less pronounced shift in the chromaticity coordinates for red and green than for blue. It is hypothesized that the diminished success of constrained scaling to reduce participant variability for scaling the brightness of blue colors is a result of a stimulus confound caused by the chromaticity shift for bright blue colors.

EXPERIMENT 8

Introduction

Previous research on constrained scaling has centered on teaching and testing stimulus intensities that were continuous in nature. In Stevens' terminology (1975), these stimuli were interval or ratio scales, meaning that the stimulus and response intensities followed a more-or-less consistent, continuous sequence from low intensity to high intensity with presumed equidistant spacing between scale units. A wealth of scaling data available and summarized in Stevens suggests that mental magnitudes follow a similar sequence from low magnitude to high magnitude. So, it would seem that the use of interval or ratio scaled stimulus intensities makes for a natural fit to human magnitude processing.

The purpose of this experiment is to explore what happens when the training scale is categorical. Categorical scales, including nominal and ordinal scales, are those scales for which there is no equidistant spacing between units. Nominal scales are scales that do not necessarily represent ascending quantity, while ordinal scales are scales that represent a climb in quantity without consistent units between stepwise increases in the scale. Stevens (1975) cautions against using categorical scales for psychological measures, since they do not map directly to human sensation. In fact, Stevens suggests that using categorical data increases the types of phenomena that Poulton (1989) came to call response biases. In this experiment, I deliberately provided training scales with impoverished ranges in order to establish the effectiveness of constrained scaling when using categorical scaling. As such, this experiment represents a refinement of existing constrained scaling methodology.

Method

Participants

Five participants with self-reported normal color vision served as volunteers for the experiment. Participants who had previously participated in a brightness scaling experiment were excluded from the present experiment. As in previous experiments, the participants were remunerated for their participation in the experiment.

Apparatus and Stimulus Materials

The apparatus and stimulus materials were identical to those used in Experiment 7, except only five brightness levels of the grayscale stimuli with feedback were presented. These levels corresponded to 1, 4, 9, 24, and 64 cd/m^2.[42] The red, green, and blue stimuli were presented and tested at the same 14 levels found in Experiments 6 and 7.

Design and Procedure

The design replicates Experiment 7, with the exception that for learning trials, participants were presented with five levels of brightness for the grayscale stimuli with feedback. As in Experiment 7, participants received 42 initial training trials, except the

[42] This scale was ordinal, in that it represented increasing units for which the distance between successive scale units was not equidistant. However, as a caveat to this experiment, it is important to note that since the scale units were approximately logarithmically equidistant, it is possible that the scale might have been interpreted to be an interval scale by some participants.

five training stimuli were repeated with greater frequency in the current experiment, thereby reinforcing familiarity with the categorical scale. Each brightness level of the grayscale squares was presented an average of over 8 times during training in the present experiment; in Experiment 7, each brightness level was presented only 3 times during training.

Results and Discussion

The data were analyzed as in Experiment 7 and are summarized in Tables 14 – 15 and Figures 40 – 43. The results for categorical brightness scaling are presented briefly below before they are compared to continuous brightness scaling later in this chapter. No reaction time data were recorded for this experiment.

Exponent Values

The average scaling exponent value for the grayscale stimuli with feedback was 0.316 with *SD*/*M* = 0.137 and *H*:*L* = 1.442:1. For red stimuli without feedback, the average exponent equaled 0.249 with *SD*/*M* = 0.158 and *H*:*L* = 1.435:1. For green stimuli without feedback, the average exponent equaled 0.273 with *SD*/*M* = 0.181 and *H*:*L* = 1.623:1. For blue stimuli without feedback, the average exponent equaled 0.124 with *SD*/*M* = 0.718 and *H*:*L* = 4.163:1. The average ratio of grayscale to red stimulus exponents was 1.299; for grayscale to green, it was 1.180; and for grayscale to blue, it was 3.332.

Intercept Values

The average scaling intercept value for the grayscale stimuli was 1.118 with *SD*/*M* = 0.050 and *H*:*L* = 1.138:1. For red stimuli, the average intercept equaled 1.376

Table 14. Summary of participants' brightness scaling in Experiment 8.

	Grey			**Red**			**Green**			**Blue**		
P	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2
1	0.318	1.104	0.854	0.299	1.213	0.664	0.303	1.152	0.699	0.281	1.252	0.643
2	0.308	1.162	0.641	0.223	1.364	0.481	0.258	1.275	0.723	0.079	1.569	0.311
3	0.385	1.036	0.665	0.233	1.418	0.535	0.259	1.371	0.626	0.068	1.586	0.422
4	0.301	1.108	0.787	0.208	1.360	0.490	0.338	1.141	0.722	0.099	1.532	0.263
5	0.267	1.179	0.675	0.282	1.525	0.605	0.208	1.378	0.624	0.092	1.572	0.447
M	0.316	1.118	0.724	0.249	1.376	0.555	0.273	1.263	0.679	0.124	1.502	0.417
SD	0.043	0.056	0.092	0.039	0.113	0.078	0.049	0.114	0.050	0.089	0.141	0.147

Table 15. Scaling ratios of grayscale stimuli to color stimuli in Experiment 8.

		Red			**Green**			**Blue**	
P	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2
1	1.065	0.958	1.287	1.049	0.958	1.223	1.131	0.882	1.329
2	1.386	0.912	1.332	1.194	0.912	0.887	3.884	0.741	2.058
3	1.652	0.756	1.242	1.487	0.756	1.062	5.695	0.653	1.577
4	1.444	0.972	1.607	0.889	0.972	1.091	3.054	0.723	2.997
5	0.946	0.855	1.115	1.281	0.855	1.081	2.895	0.750	1.508
M	1.299	0.891	1.316	1.180	0.891	1.069	3.332	0.750	1.894
SD	0.288	0.088	0.181	0.227	0.088	0.120	1.658	0.083	0.673

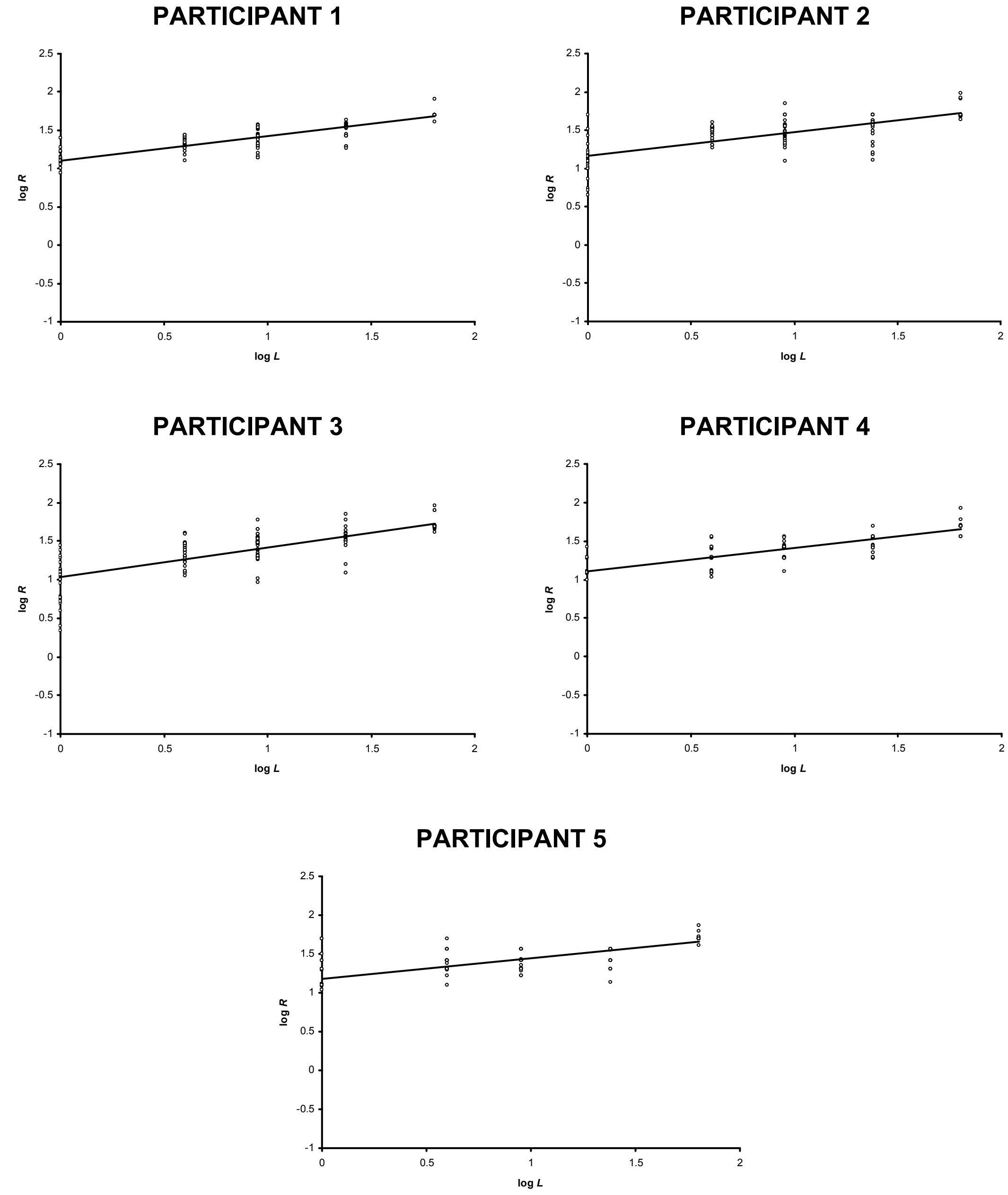
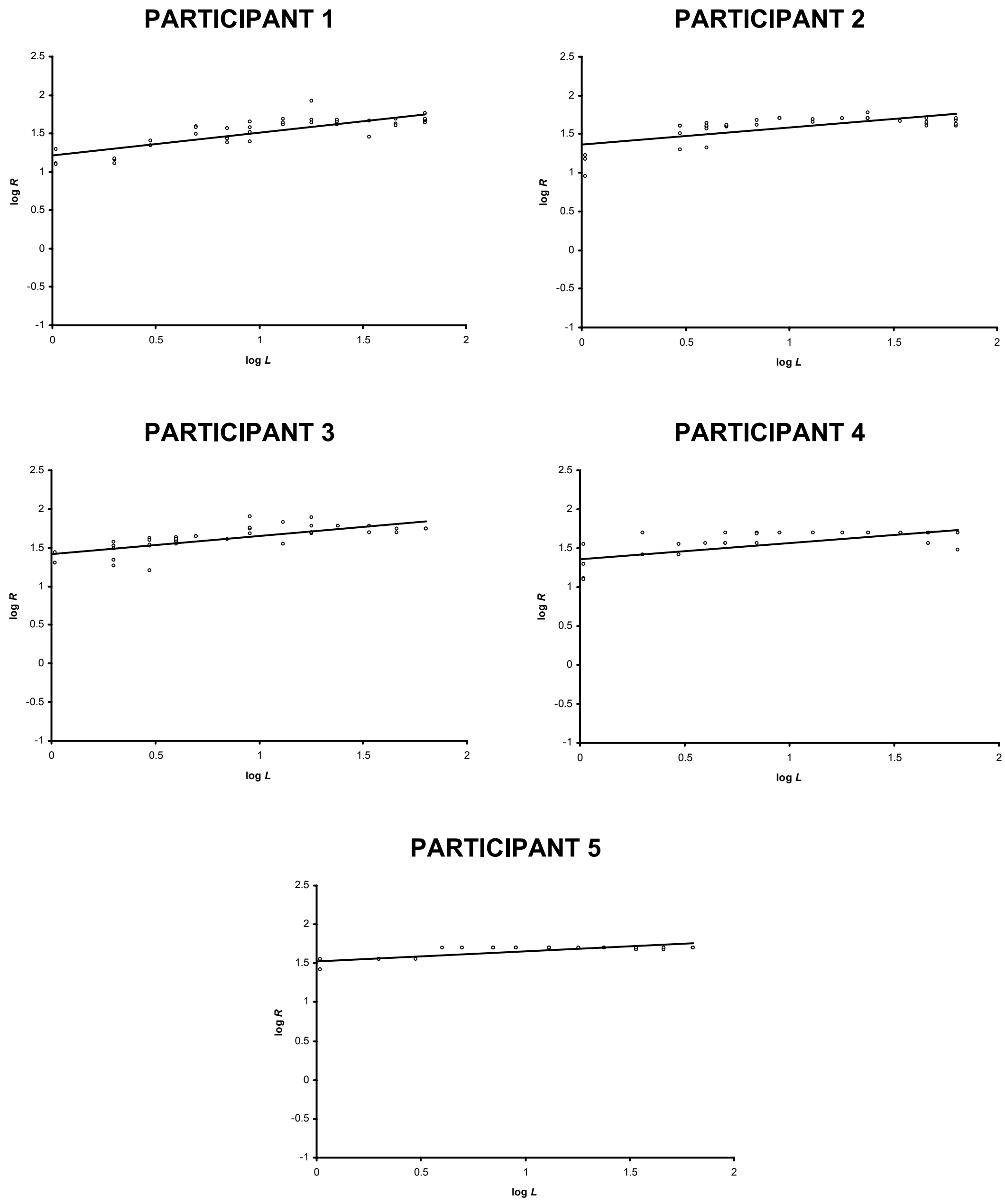


*Figure 40. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for grayscale squares with feedback in Experiment 8.*


PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

*Figure 41. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for red squares without feedback in Experiment 8.*

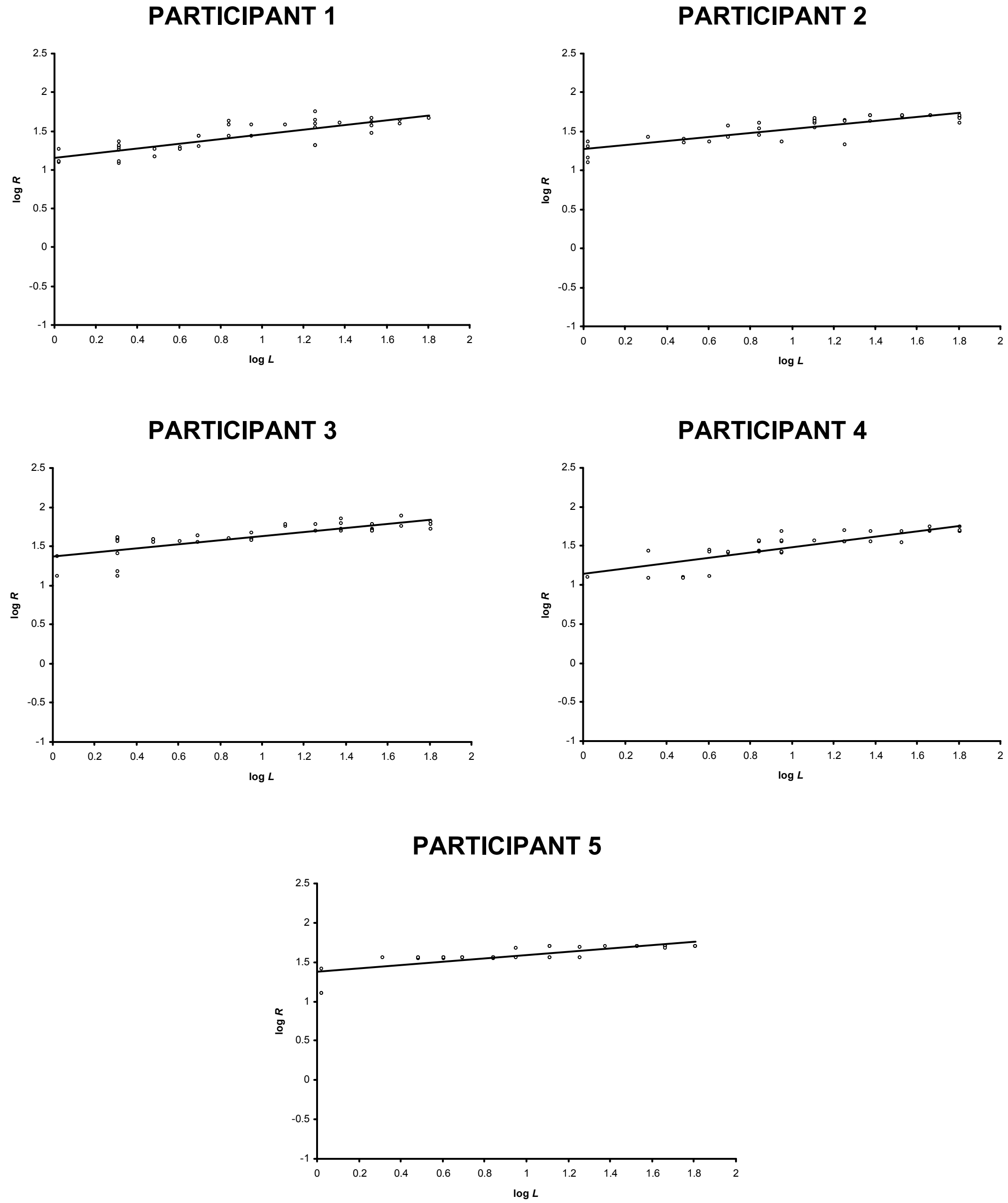


Figure 42. *Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for green squares without feedback in Experiment 8.*

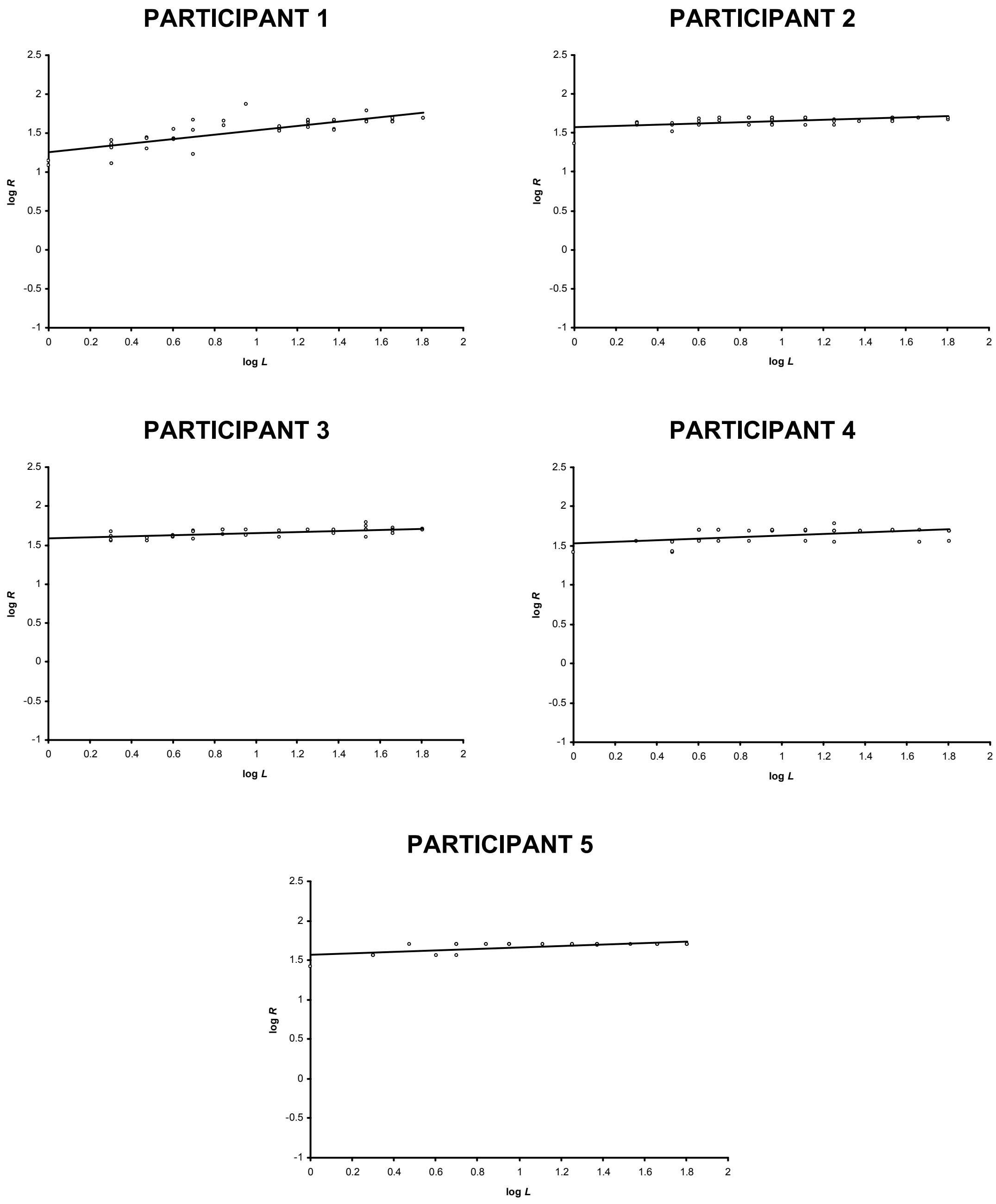


*Figure 43. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for blue squares without feedback in Experiment 8.*

with *SD*/*M* = 0.082 and *H*:*L* = 1.257:1. For green stimuli, the average intercept equaled 1.263 with *SD*/*M* = 0.091 and *H*:*L* = 1.208:1. For blue stimuli, the average intercept equaled 1.502 with *SD*/*M* = 0.094 and *H*:*L* = 1.267:1. The average ratio of grayscale to red stimulus intercepts was 0.816; for grayscale to green, it was 0.891; and for grayscale to blue, it was 0.750.

Goodness of Fit Coefficients

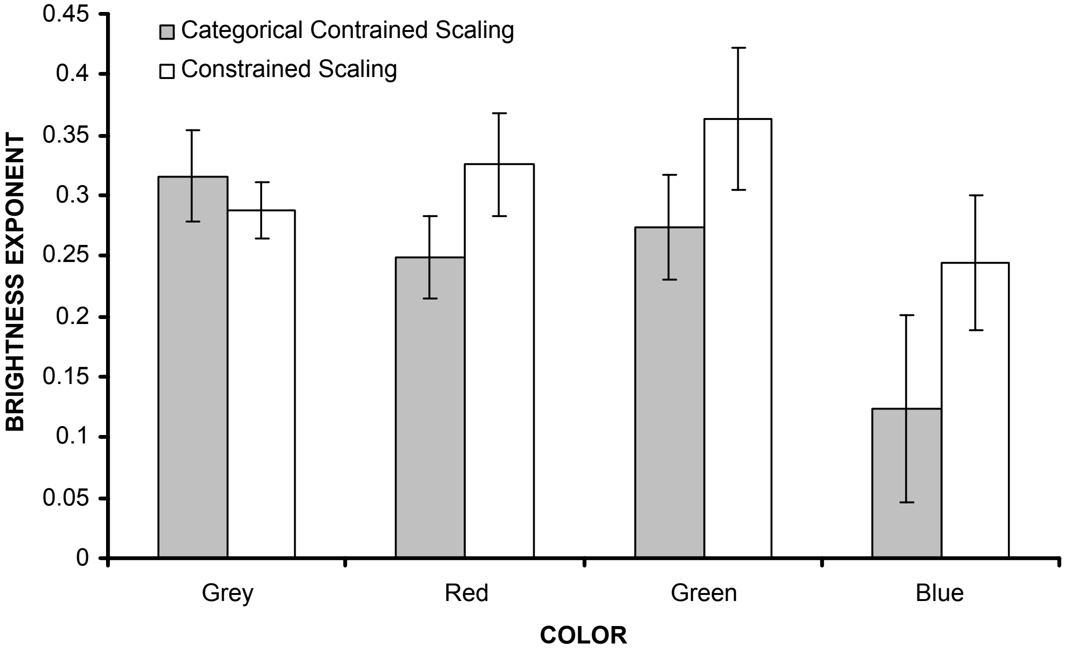
The average R^2 value for scaling grayscale stimuli was 0.724 with *SD*/*M* = 0.127 and *H*:*L* = 1.333:1. For red stimuli, the average R^2 equaled 0.555 with *SD*/*M* = 0.127 and *H*:*L* = 1.380:1. For green stimuli, the average R^2 equaled 0.679 with *SD*/*M* = 0.073 and *H*:*L* = 1.148:1. For blue stimuli, the average R^2 equaled 0.417 with *SD*/*M* = 0.353 and *H*:*L* = 2.448:1. The average R^2 ratio of grayscale to red stimuli was 1.316; for grayscale to green, it was 1.069; for grayscale to blue, it was 1.894.

Comparison of Experiments 7 and 8

Constrained scaling (Experiment 7) and categorical constrained scaling (the present experiment) are compared in Figures 44 – 46. In general, training using a categorical scale resulted in decreased exponent values and goodness of fit coefficients, and increased intercept values.

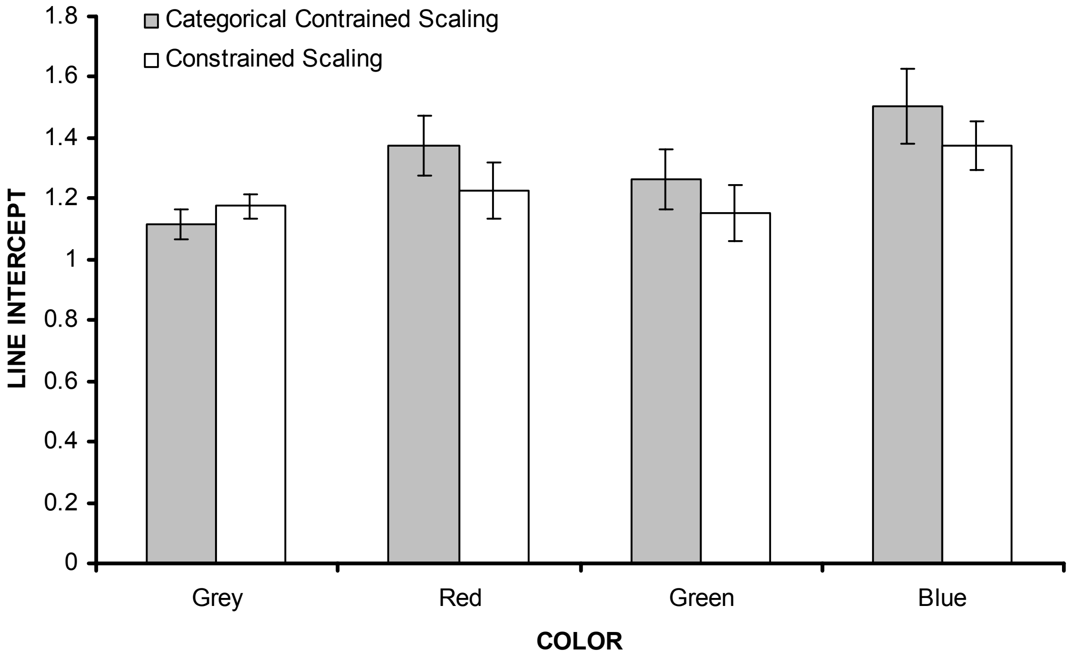
Exponent Values

Figure 44 represents the relationship of exponent values between grayscale, red, green, and blue color stimuli for conventional and categorical constrained scaling. There was no significant difference in scaling exponents between methods for grayscale stimuli [$t(8) = 1.246$, $p = 0.124$], but categorical constrained scaling resulted in significantly



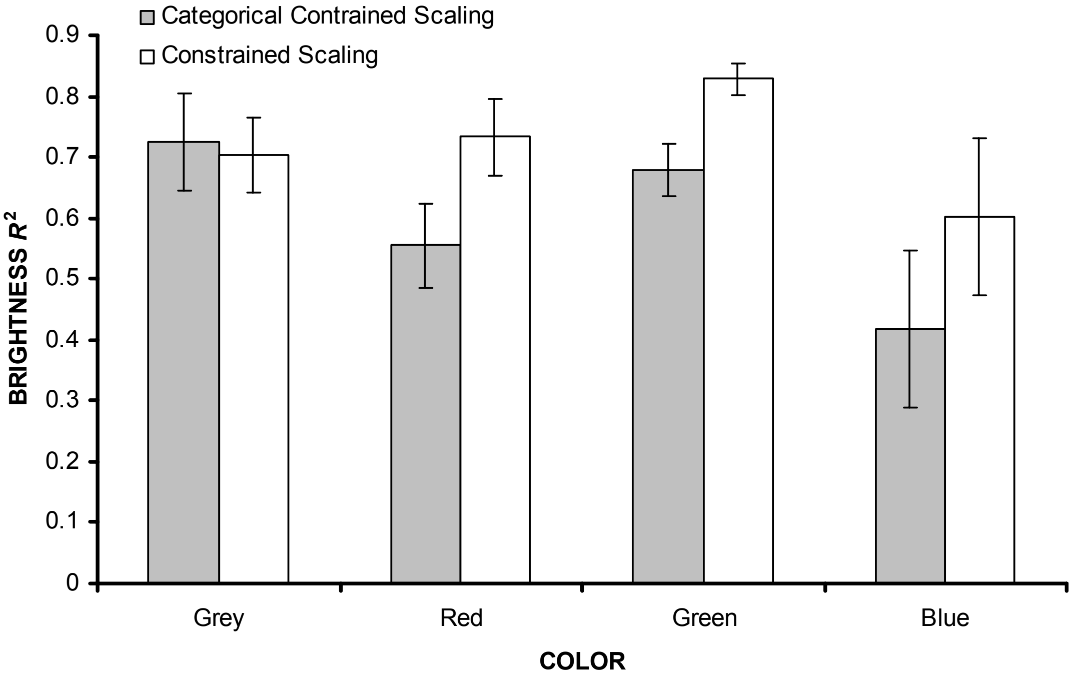
NOTE: The error bars represent the 95% confidence intervals.

Figure 44. Comparison of brightness exponents between Experiment 7 (constrained scaling) and Experiment 8 (categorical constrained scaling).



NOTE: The error bars represent the 95% confidence intervals.

Figure 45. Comparison of brightness intercepts between Experiment 7 (constrained scaling) and Experiment 8 (categorical constrained scaling).



NOTE: The error bars represent the 95% confidence intervals.

Figure 46. Comparison of goodness of fit coefficients between Experiment 7 (constrained scaling) and Experiment 8 (categorical constrained scaling).

lower exponent values than did conventional constrained scaling for red stimuli [$t(8) = -2.759$, $p < 0.05$], green stimuli [$t(8) = -2.405$, $p < 0.05$], and blue stimuli [$t(8) = -2.478$, $p < 0.05$]. There was no significantly difference scaling variance between constrained scaling methods for grayscale stimuli [$F(1,4) = 2.790$, $p = 0.172$], red stimuli [$F(1,4) = 0.659$, $p = 0.348$], green stimuli [$F(1,4) = 0.541$, $p = 0.283$], or blue stimuli [$F(1,4) = 1.956$, $p = 0.266$]. This minimal difference in exponent variance was confirmed by the similarity of the coefficient of variation between the two methods. The *SD*/*M* was 1.522 times higher for grayscale stimuli in conventional constrained scaling and 2.766 times higher for blue stimuli. It was nearly identical for red stimuli (1.063 times higher) and green stimuli (1.024 times lower). The highest-to-lowest exponent ratios were nearly identical across methods (1.156 and 1.005 times higher for grayscale and red stimuli, respectively, and 1.065 times lower for green stimuli), with the exception of the blue stimuli (1.918 times higher).

Intercept Values

Figure 45 shows the relationship between scaling intercepts for grayscale, red, green, and blue stimuli. Conventional constrained scaling resulted in a significantly lower intercept value for red stimuli [$t(8) = 2.161$, $p < 0.05$] and a marginally lower intercept value for green stimuli [$t(8) = 1.573$, $p = 0.077$] and blue stimuli [$t(8) = 1.679$, $p = 0.66$]. It resulted in a marginally higher intercept value for grayscale stimuli [$t(8) = -1.804$, $p = 0.054$]. There were no significant differences across methods in terms of variance [$F(1,4) = 1.499$, $p = 0.352$ for grayscale stimuli; $F(1,4) = 1.122$, $p = 0.457$ for red stimuli; $F(1,4) = 1.189$, $p = 0.436$ for green stimuli; $F(1,4) = 2.376$, $p = 0.211$ for blue stimuli]. The comparable levels of variance between both constrained scaling methods were confirmed by negligible differences in the coefficients of variation and the highest-to-lowest intercept ratios. For grayscale stimuli, conventional constrained scaling resulted in an *SD*/*M* for intercept values that was an average of 1.289 times higher than categorical constrained scaling.[43] For red stimuli, conventional constrained scaling resulted an SD/M for intercept values that was an average of 1.060 times lower than categorical scaling. For green stimuli, it was 1.004 times lower, and for blue stimuli, it was 1.093 times higher. Similarly, the *H*:*L* ratio was 1.023 times lower for grayscale stimuli using conventional constrained scaling, 1.010 times lower for red stimuli, 1.020 times higher for green stimuli, and 1.093 times lower for blue stimuli.

Goodness of Fit Coefficients

Figure 46 displays the goodness of fit coefficients for the color stimuli across constrained scaling methods. There was no significant difference in terms of average R^2 values across methods for grayscale stimuli [$t(8) = 0.426$, $p = 0.341$]. However, the average R^2 value was significantly higher using conventional constrained scaling for red stimuli [$t(8) = -3.705$, $p < 0.005$], green stimuli [$t(8) = - 5.797$, $p < 0.001$], and blue stimuli [$t(8) = -1.982$, $p < 0.05$]. As in Experiment 7, no *F*-statistic was calculated for the R^2 values.

[43] Slightly lower variability would be expected in the categorical scaling for the grayscale training stimuli, since the stimulus set was presented more often than in conventional constrained scaling.

General Discussion

Although there were some significant and marginally significant differences in terms of scaling exponents, intercepts, and goodness of fit coefficients, there was no significant difference between conventional and categorical constrained scaling in terms of variance. These findings suggest that participants, on average, were able to learn the ordinal scale and subsequently use it to scale interval stimuli with the same reliability as if they had been trained on an interval scale. However, the resulting scale values, especially the exponent values, differed markedly from those values resulting from conventional constrained scaling. With the exception of the grayscale training stimuli, color scaling exponents tended to be smaller when using categorical constrained scaling. The reason for this difference is not clear. One possibility is that participants tend to memorize categorical scales[44] and simply use the learned values as rote anchor points, even when they do not fully apply. To explore this hypothesis, a method will be explored in the next experiment to prevent rote memorization of the categorical training scale.

[44] In their absolute identification experiments, Shiffrin and Nosofsky (1994) noted a memory limitation in terms of labeling unidimensional scales around 7 ± 2 items, similar to Miller's (1956) short-term memory capacity limit. The present experiment complements these earlier findings by demonstrating the learnability of a categorical scale and its applicability for scaling novel continuous-type scales.

EXPERIMENT 9

Introduction

Experiment 8 featured an implementation of constrained scaling, in which an ordinal scale was used for training and subsequently applied to ratio scale stimuli. The danger of training with a categorical type scale like an ordinal scale is that it is possible for participants to memorize the feedback values. It is presumed that rote memorization of anchor values is not the same cognitive process as learning the range of a scale. Compared to conventional constrained scaling, participants in the categorical constrained scaling condition had significantly lower exponent values and regression line goodness of fit coefficients. These findings combine to confirm that participants applied the learned grayscale differently between categorical and conventional constrained scaling. Yet, the reason for this remains unclear.

When participants memorize a limited set of stimulus-response combinations and are later presented with stimuli that fall outside those anchor points, they must either round to the nearest memorized stimulus-response pair or extrapolate between two bounding stimulus-response pairs. Anecdotal evidence from discussions with participants during debriefing suggests that participants liberally rounded their response rather than inputting the exact value they had memorized for the grayscale stimuli. This was a byproduct of the time required by participants to fine tune the slider scale to exact values. As a timesaver, they approximated the nearest scale value within a few whole numbers of the desired response value. Because of response rounding and the resultant response scale noisiness, it was not possible to determine whether participants memorized and categorized novel stimuli into the existing learned scaling categories or extrapolated between two memorized anchors.

In order to determine if rote memorization played a role on constrained scaling in Experiment 8, the present experiment attempted to prevent rote memorization. It replicated the same categorical training scale as in Experiment 8 but added ±5% random noise to the feedback provided to participants during training. It was hypothesized that the inclusion of random noise would not generally diminish learning of the grayscale stimuli but would increase scaling performance for novel brightness stimuli.[45] Specifically, it was hypothesized that the scaling exponent, intercept, and goodness of fit coefficient would be comparable for grayscale stimuli but that color measures would more closely reflect conventional constrained scaling in the noisy categorical constrained scaling condition.

Method

Participants

As in previous brightness scaling experiments, five volunteers with self-reported normal color vision were enlisted as participants for the experiment. Previous participants for brightness experiments were precluded from volunteering. The volunteers were paid $5.00 for their participation in the experiment.

[45] The feedback provides an anchoring range of scale values for each grayscale stimulus intensity value, and this anchoring range should be readily learnable given the high number of training iterations provided in the experiment.

Apparatus and Stimulus Materials

The experimental control software was identical to that used in Experiment 8. The brightness stimuli consisted only of grayscale and red stimuli. The green stimuli were eliminated from the present experiment, because the results for green stimuli were virtually identical to the results for red stimuli and, thus, represented a redundant stimulus set. The blue stimuli were eliminated, because the results for blue stimuli systematically differed from the results for the other colors and the grayscale stimuli.[46]

Design and Procedure

The design and procedure of the present experiment replicated Experiment 8 with two exceptions. First, the five grayscale stimuli were presented with feedback that contained ±5% random variability. Second, only grayscale and red stimuli were tested, resulting in a shorter experiment than previously. The design and procedure of the present experiment are summarized in Figure 47.

Results and Discussion

No reaction time data were recorded for this experiment. The results were analyzed as in Experiment 8 and are summarized in Table 16 and Figures 48 – 49. The results for the exponent values, intercept values, and goodness of fit coefficients are discussed next, followed by a summary section comparing the results from the present experiment with the results from Experiment 8.

[46] As noted earlier, the blue phosphor of the CRT was the faintest and required considerable mixing of the red and green phosphors to produce the full brightness range used in the experiments. This mixing of colors may have caused a shift to achromatic perception of the blue stimuli.

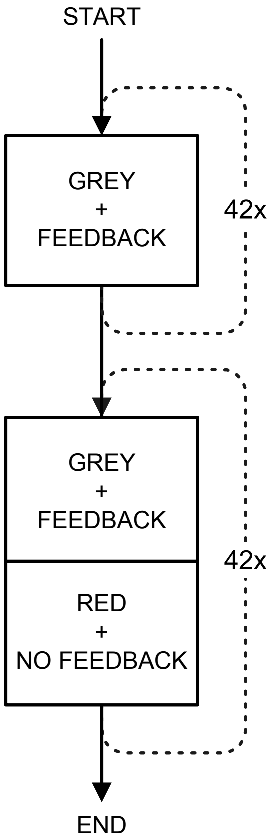
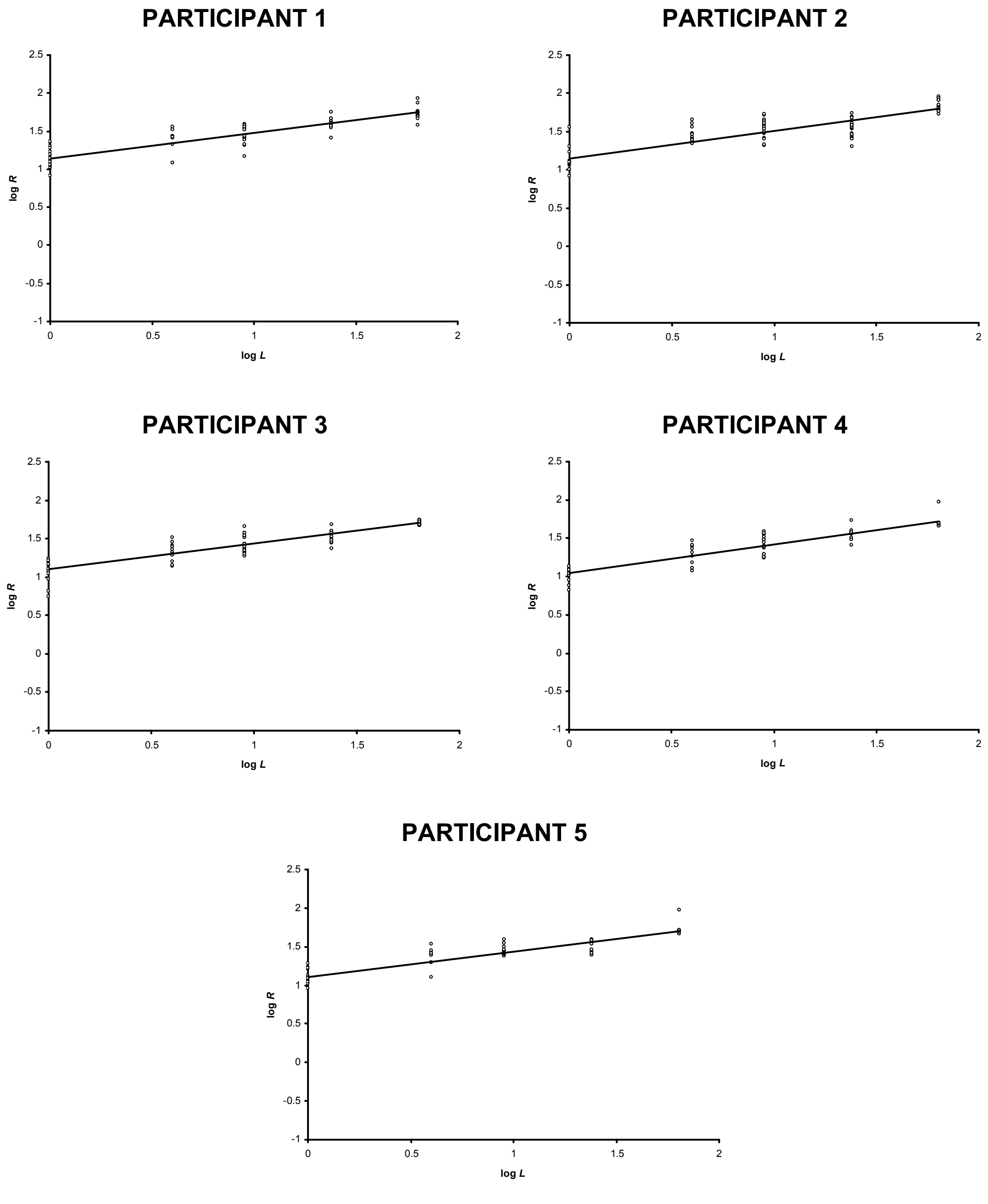


Figure 47. Schematic flow of the design for noisy categorical constrained scaling in Experiment 9.

Table 16. Summary of participants' brightness scaling in Experiment 9.

	Grayscale Brightness + Feedback			**Red Brightness + No Feedback**			**Ratio of Grayscale:Red Brightness**		
P	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2	**Exponent**	**Intercept**	R^2
1	0.336	1.138	0.818	0.361	1.199	0.696	0.930	0.949	1.176
2	0.361	1.149	0.754	0.405	1.201	0.567	0.891	0.957	1.329
3	0.333	1.098	0.815	0.297	1.115	0.814	1.123	0.985	1.001
4	0.372	1.048	0.850	0.203	1.432	0.784	1.827	0.732	1.084
5	0.327	1.109	0.867	0.330	1.275	0.756	0.992	0.870	1.147
M	0.346	1.108	0.821	0.319	1.244	0.723	1.153	0.899	1.147
SD	0.019	0.040	0.043	0.076	0.119	0.098	0.387	0.102	0.122



PARTICIPANT 1

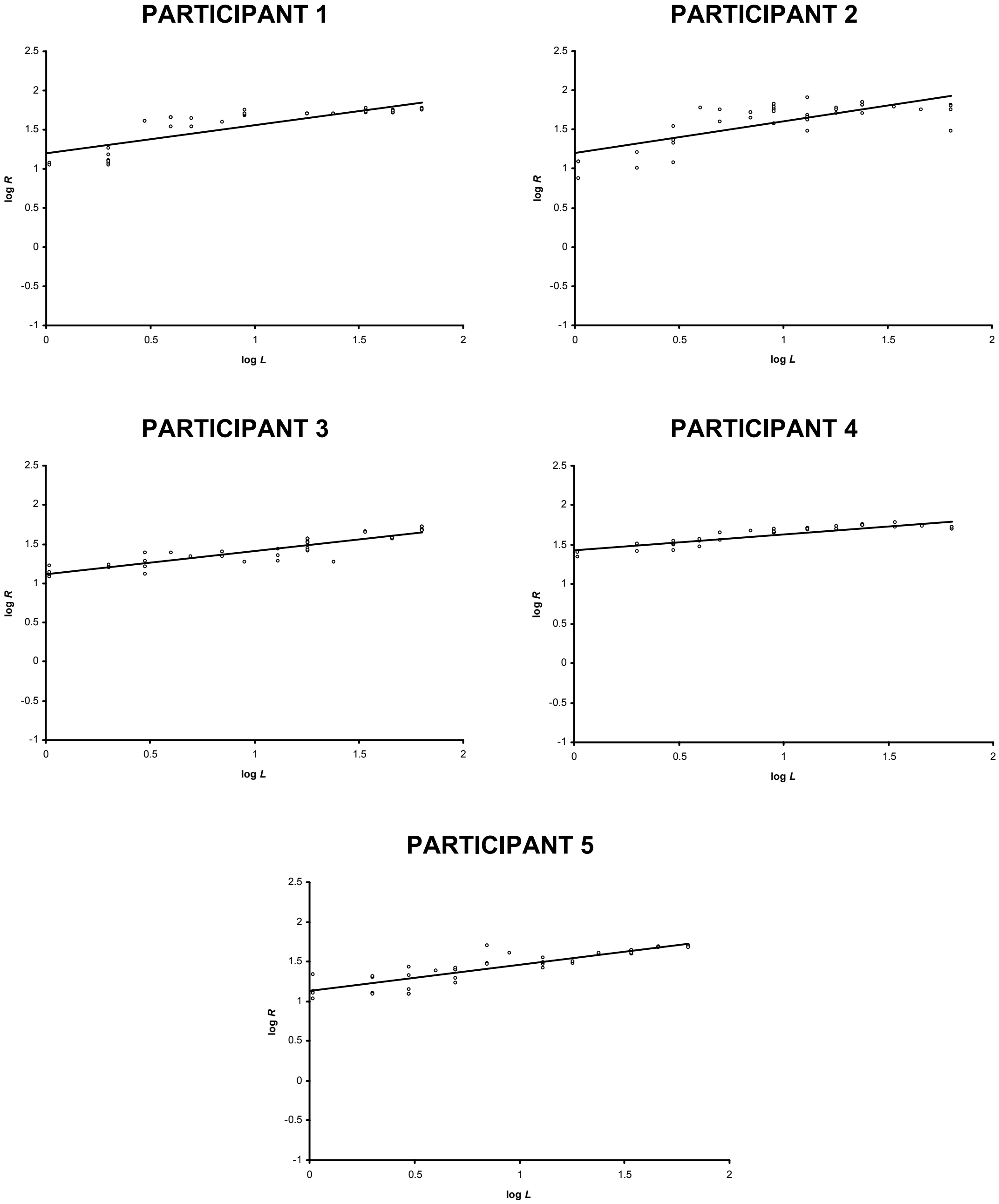
PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

*Figure 48. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for grayscale squares with feedback in Experiment 9.*



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

*Figure 49. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for red squares without feedback in Experiment 9.*

Exponent Values

The average scaling exponent for grayscale stimuli with feedback was 0.346 with *SD*/*M* = 0.056 and *H*:*L* = 1.136:1. The average scaling exponent for red stimuli without feedback was 0.319 with *SD*/*M* = 0.238 and *H*:*L* = 1.991:1. The average ratio of grayscale to red scaling exponents was 1.153.

Intercept Values

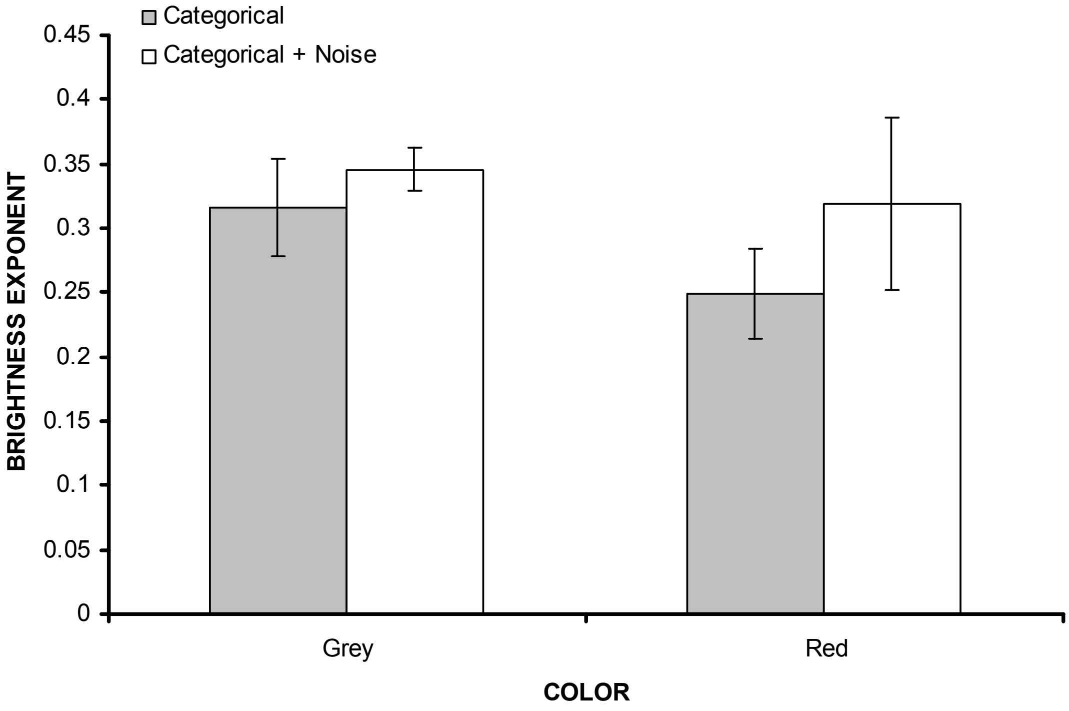
For grayscale stimuli with feedback, the average scaling intercept was 1.108 with *SD*/*M* = 0.036 and *H*:*L* = 1.096:1. For red stimuli without feedback, the average scaling intercept was 1.244 with *SD*/*M* = 0.096 and *H*:*L* = 1.284:1. The average ratio of grayscale to red scaling intercepts was 0.899.

Goodness of Fit Coefficients

The average goodness of fit coefficient for scaling of the grayscale stimuli was 0.821 with *SD*/*M* = 0.053 and *H*:*L* = 1.150:1. For the red stimuli, R^2 averaged 0.723 with *SD*/*M* = 0.135 and *H*:*L* = 1.435:1. The average ratio of grayscale to red R^2 values was 1.147.

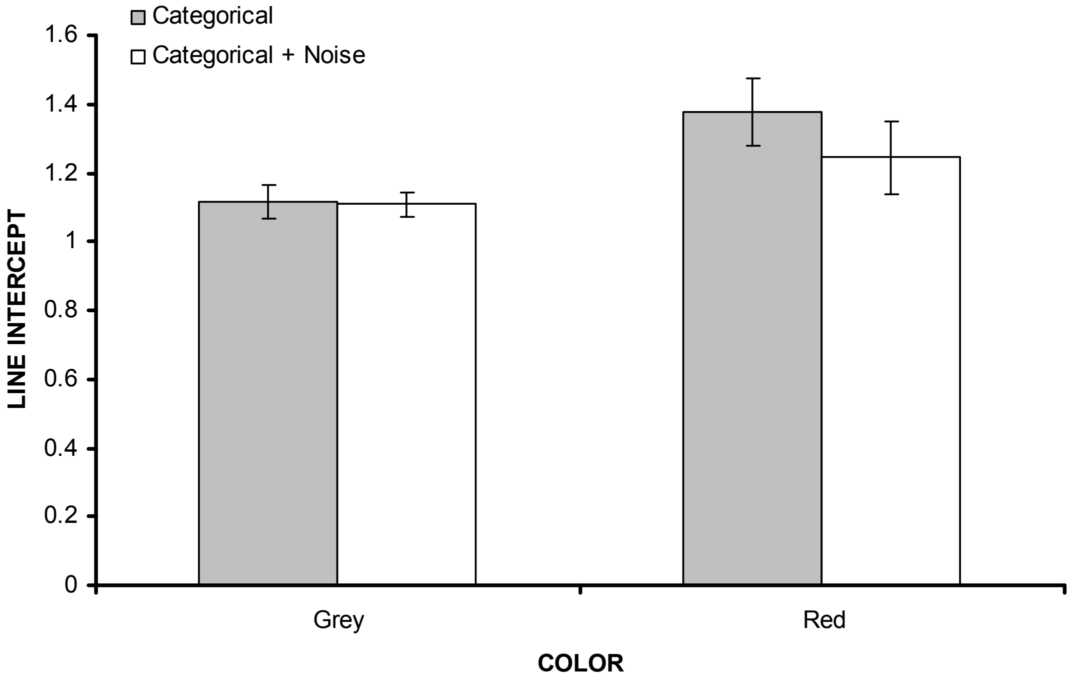
Comparison of Experiments 8 and 9

Categorical constrained scaling (Experiment 8) is contrasted with noisy categorical constrained scaling (the present experiment) in Figures 50 – 52. Note that since the present experiment only included grayscale and red stimuli, only the grayscale and red stimuli are included from the Experiment 8 dataset. The figures illustrate that the addition of noise to the training trial feedback resulted in higher exponent, similar intercept, and higher R^2 values for the training and testing stimuli.



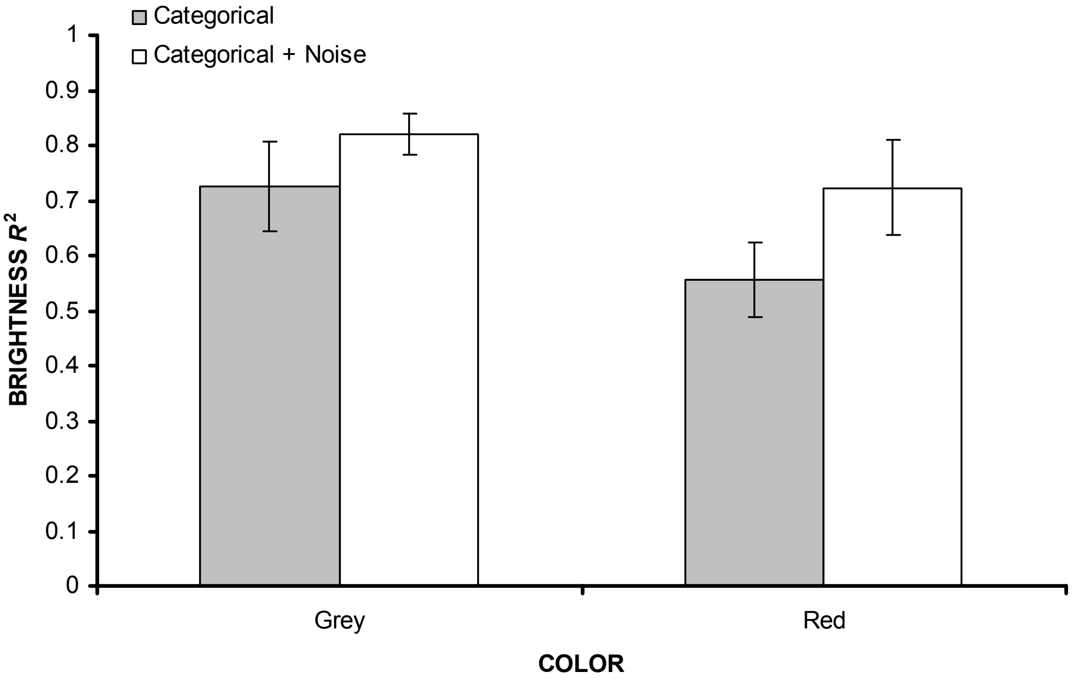
NOTE: The error bars represent the 95% confidence intervals.

Figure 50. Comparison of exponents between Experiment 8 (categorical constrained scaling) and Experiment 9 (noisy categorical constrained scaling).



NOTE: The error bars represent the 95% confidence intervals.

Figure 51. Comparison of line intercepts between Experiment 8 (categorical constrained scaling) and Experiment 9 (noisy categorical constrained scaling).



NOTE: The error bars represent the 95% confidence intervals.

Figure 52. Comparison of goodness of fit coefficients between Experiment 8 (categorical constrained scaling) and Experiment 9 (noisy categorical constrained scaling).

Exponent Values

Figure 50 contrasts the scaling exponent values for the categorical and noisy categorical constrained scaling experiments. For the grayscale stimuli, the noisy categorical condition resulted in a marginally greater exponent value [$t(8) = 1.409$, $p = 0.098$]. Categorical and noisy categorical constrained scaling results were approximately equally far apart from the training exponent of 0.33. For the red stimuli, the noisy categorical condition resulted in a significantly greater exponent value [$t(8) = 1.835$, $p = 0.052$). The red stimulus scaling exponent for noisy categorical constrained scaling was nearly identical to that produced by conventional categorical scaling, 0.319 versus 0.326, respectively.

Noisy categorical constrained scaling had marginally less variance than categorical constrained scaling for grayscale stimuli [$F(1,4) = 0.201$, $p = 0.074$]. In terms of the SD/M for the grayscale scaling exponents, the noisy categorical method was 2.445 times lower, whereas, in terms of $H{:}L$, it was 1.269 times lower. However, categorical constrained scaling exhibited slightly less variance than noisy categorical constrained scaling for red stimuli [$F(1,4) = 3.733$, $p = 0.115$]. The categorical SD/M for the red scaling exponent was 1.507 times lower than the noisy categorical SD/M, and the categorical $H{:}L$ was 1.388 times lower.

Intercept Values

Figure 51 illustrates the relationship between scaling intercepts for categorical and noisy categorical constrained scaling results. There was no significant difference in average grayscale intercepts between the two methods [$t(8) = -0.306$, $p = 0.384$]. The noisy categorical method produced a significantly lower red intercept [$t(8) = -1.792$, $p = 0.0554$], which was nearly identical to the intercept produced by conventional constrained scaling, respectively 1.244 vs. 1.226.

There were no significant differences in terms of intercept variance between categorical and noisy categorical constrained scaling for grayscale and red stimuli [$F(1,4) = 0.493$, $p = 0.255$, and $F(1,4) = 1.112$, $p = 0.460$, respectively]. For the grayscale stimuli, the *SD*/*M* was 1.412 times lower for the noisy categorical method, but the *H*:*L* ratio was nearly identical at 1.038 times lower. For the red stimuli, the *SD*/*M* was 1.166 times lower for the categorical method, with the *H*:*L* ratio nearly identical (1.021 times greater).

Goodness of Fit Coefficients

Figure 52 illustrates the relationship of the goodness of fit coefficients for categorical and noisy categorical constrained scaling. For both the grayscale and the red stimuli, the noisy categorical method produced significantly higher R^2 values [$t(8) = 2.119$, $p < 0.05$, for grayscale stimuli, and $t(8) = 3.009$, $p < 0.01$, for red stimuli]. As in previous experiments, the variance of the R^2 values was not analyzed.

General Discussion

The incorporation of noise into the feedback provided in the training trials significantly improved the results for categorical constrained scaling. When noise was added, the categorical constrained scaling results closely resembled the results obtained through conventional constrained scaling, including better matched exponent and intercept values as well as better goodness of fit by the scaling data to the regression line.

Noisy categorical constrained scaling provides a useful refinement to the repertoire of constrained scaling methodology in that it potentially offers a simplified training regimen from which participants can learn a constrained scale. It further provides a technique for training on an ordinal scale for subsequent testing on a ratio scale—an approach that may hold application when working within stimulus domains for which an interval training scale may be unfeasible. The use of noisy categorical training also effectively eliminates the scaling artifacts of rote memorization, which in Experiment 8 may have contributed to deflated testing exponent values, inflated testing intercept values, and low goodness of fit coefficients compared to conventional constrained scaling.

EXPERIMENT 10

Introduction

An important issue involving constrained scaling is the amount of inculcation that is necessary to maintain learning. The benefit of training on a stimulus scale is clearly illustrated with the constrained scaling method. Training calibrates individual scale use, thereby significantly reducing scaling variability. However, guidelines for the optimal level of training are anecdotal, and no systematic study of the effects of reduced or increased training exists. Moreover, the necessity of continued training at regular intervals is not clearly documented.

Training may actually be counterproductive when too much is applied. This would especially be likely given the single-session experiments that are typical for constrained scaling. Increased amounts of training beget increased testing times, which may push the bounds of the individual participant's attention span. Task vigilance has been found to decrease sharply after 30 minutes (Mackworth, 1948). The four-block (i.e., training, testing, training, and testing) experiments common in West et al. (2000) and replicated here in Experiments 1 and 2 required up to 60 minutes for completion. The use of a block of initial and intermediate training trials, coupled with a training trial interspersed between every testing trial, contributed greatly to this experimental duration. Part of this problem was resolved by shortening the number of blocks. In West et al. and here in Experiments 3 onward, only an initial training block and a subsequent testing block were performed. For simpler experiments, this shortened the average experimental duration to around 30 minutes. More complex experiments, such as Experiments 6 – 8, which involved scaling multiple colors, required longer experimental duration.[47]

The effects of experimental duration on performance are illustrated through a review of Experiment 1. Recall that Experiment 1 consisted of an initial training block of 50 trials of 1000 Hz tones with feedback. This training block was followed by 100 interspersed trials of 1000 Hz tones with feedback and 65 Hz tones without feedback. This procedure was duplicated with another 50 training trials and 100 interspersed testing trials. Table 17 shows the participants' absolute error rates[48] for blocks of learning trials throughout the duration of Experiment 1. In the table, the training and testing trials are subdivided into blocks of 25 trials. Note that the table only includes results for the 1000 Hz tones with feedback, and not for the 65 Hz tones without feedback.

Overall, there is a slight decrease in the error rate from the first block (M = 5.690, SD = 1.710) to the second block (M = 4.077, SD = 1.582). A paired sample two-tailed t-test revealed that the decrease in error rate between the first and second training blocks was not significant, $t(4) = 1.584$, $p = 0.189$. Over the course of the experiment, participants' error rates for learning trials subsequently increased. While the average error rate is 4.883 for the first training blocks, the average error rate increased to 6.867 for the testing blocks, to 6.956 for the second set of training blocks, to 8.506 for the final

[47] It may be tempting to attribute the unusual performance at scaling blue colors in Experiments 6 – 8 to the deleterious performance effects of the long experimental duration. However, such deleterious effects would affect the brightness scaling of all colors, not just blue colors. The placement of blue last in the table does not, of course, imply the presence of scaling chronology or order effects in the experiment.

[48] The absolute error rate is the absolute difference between the feedback value and the scale value selected by the participant.

		P1	P2	P3	P4	P5	*M* ± *SD*	
Training	Block 1	5.771	2.760	6.207	7.105	6.605	5.690 ± 1.710	4.883 ± 1.770
	Block 2	5.883	3.800	3.369	5.373	1.959	4.077 ± 1.582	
Testing	Block 3	5.196	3.828	10.089	6.033	3.388	5.707 ± 2.668	6.867 ± 2.774
	Block 4	6.485	4.786	8.885	5.045	7.164	6.473 ± 1.672	
	Block 5	7.201	7.846	12.434	3.782	5.648	7.382 ± 3.230	
	Block 6	8.147	5.276	13.956	5.662	6.496	7.907 ± 3.557	
Training	Block 7	3.112	6.329	13.380	6.462	6.551	7.167 ± 3.762	6.956 ± 2.981
	Block 8	4.901	4.896	10.682	7.163	6.087	6.746 ± 2.394	
Testing	Block 9	8.079	4.247	11.152	7.996	6.308	7.556 ± 2.545	8.506 ± 3.542
	Block 10	12.498	10.295	13.377	4.681	8.276	9.825 ± 3.494	
	Block 11	6.329	7.025	14.171	6.001	6.067	7.919 ± 3.519	
	Block 12	13.039	3.803	14.887	7.066	4.813	8.722 ± 4.971	

Table 17. Mean error rates for participants per blocks of 25 learning trials of 1000 Hz tones in Experiment 1.

testing blocks. There was a significant effect for the error rate across trials, $F(3) = 3.358$, $p = 0.055$, suggesting that, on average, the error rate increased over time. The mean error rate of the initial training blocks compared to the initial testing blocks and second training blocks was not significant, $t(4) = -1.570$, $p = 0.10$, and $t(4) = -1.394$, $p = 0.12$, respectively. There was a significant increase in the error rate between first training blocks and the final testing blocks, $t(4) = -2.575$, $p = 0.03$. Overall, these results suggest that increased training did not improve error rates after the first training blocks. In fact, over time, participants became nominally worse at assigning magnitude values to reflect the loudness intensities they heard.

The field of human reliability analysis offers useful insights into the interaction of training and attention. Human reliability analysis combines the influences of several performance shaping factors to determine the probability of human error (Gertman & Blackman, 1994). One popular method of human reliability analysis, SPAR-H (Gertman, Blackman, Marble, Byers, Haney, & Smith, 2004) specifically identifies *Experience and Training* and *Fitness for Duty* (of which fatigue is a major component) as performance shaping factors. In the SPAR-H method, given cognitively engaging tasks such as scaling, training would on average be expected to decrease the participants' probability of error by one-half. However, any benefit of training is overshadowed by the deleterious effect of fatigue. A moderate amount of fatigue may degrade human performance significantly, resulting in a fivefold increase in the probability of human error. Composite effects in SPAR-H are calculating by multiplying the effects of individual performance shaping factors (i.e., $0.5 \times 5 = 2.5$, in the present example). Thus, even with training, fatigue during a lengthy experimental session could result in a two-and-a-half times increase in the likelihood of making a scaling error.[49]

Since constrained scaling also works within shorter experimental sessions in which fatigue is not expected to play a significant role, the question remains regarding the optimal level of training to administer so that participants can master the training scale. In the present experiment, I investigated the effect of the training reinforcement trials within the context of color scaling. The initial block of 41 grayscale training stimuli was maintained as in previous brightness scaling experiments. However, instead of a training reinforcement of a brightness stimulus with feedback interspersed between every testing stimulus without feedback, the ratio was decreased to a training reinforcement stimulus interspersed between every other testing stimulus.

Method

Participants

Five participants with self-reported normal color vision were enlisted to participate in the experiment. Participants who had previously participated in a color brightness scaling experiment were excluded. The participants were paid $5 for taking part in the experiment.

[49] The human reliability analysis example using the SPAR-H method is intended for illustrative purposes only. The SPAR-H method was developed for the US Nuclear Regulatory Commission to model human performance in nuclear power plants. The actual error rate modifiers may be quite different in a non safety critical environment. Work is underway to determine the generalizability of the approach to other domains (Boring, Gertman, & Marble, 2004).

Apparatus and Stimulus Materials

The experimental control software was based on the software used in Experiment 9, with the exception that the reinforcement ratio was decreased during the testing trials. The stimulus materials consisted of grayscale and red colored squares at 14 levels of luminous intensity.

Design and Procedure

Figure 53 illustrates the design and procedure of the present experiment. The design is similar to the design of Experiment 9 with two exceptions. First, unlike the categorical constrained scaling experiments (Experiments 8 and 9), feedback was provided at all 14 luminous intensity levels for grayscale stimuli. This use of conventional constrained scaling was similar to Experiment 7. Second, unlike any previous experiments, during the testing block, grayscale stimuli with feedback were only presented once for every two red stimuli without feedback. As a consequence, the testing block was only repeated 21 times instead of 42 times, since each testing iteration resulted in two ratings of red brightness stimuli.

Results and Discussion

The results were analyzed as in previous brightness constrained scaling experiments. The results are summarized in Table 18 and Figures 54 – 55 and are discussed below. Later in this chapter, a comparison between the results from Experiment 9 and the present Experiment is provided.

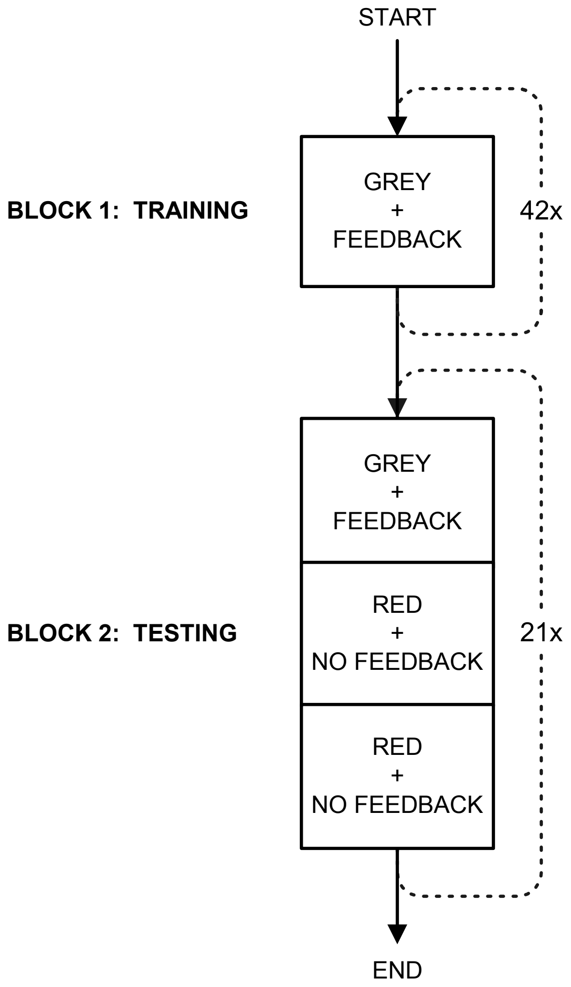


Figure 53. *Schematic flow of the reduced feedback ratio design in Experiment 10.*

Table 18. Summary of participants' brightness scaling in Experiment 10.

	Grayscale Brightness + Feedback			**Red Brightness + No Feedback**			**Ratio of Grayscale:Red Brightness**		
P	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***
1	0.437	0.980	0.804	0.455	1.031	0.866	0.960	0.951	0.928
2	0.285	1.169	0.615	0.242	1.451	0.720	1.178	0.806	0.854
3	0.431	1.001	0.775	0.403	1.162	0.650	1.069	0.861	1.192
4	0.403	1.028	0.401	0.153	1.662	0.196	2.634	0.619	2.046
5	0.339	1.130	0.665	0.411	1.078	0.777	0.825	1.048	0.856
M	0.379	1.061	0.652	0.333	1.277	0.642	1.333	0.857	1.175
SD	0.065	0.083	0.160	0.129	0.270	0.261	0.739	0.162	0.506

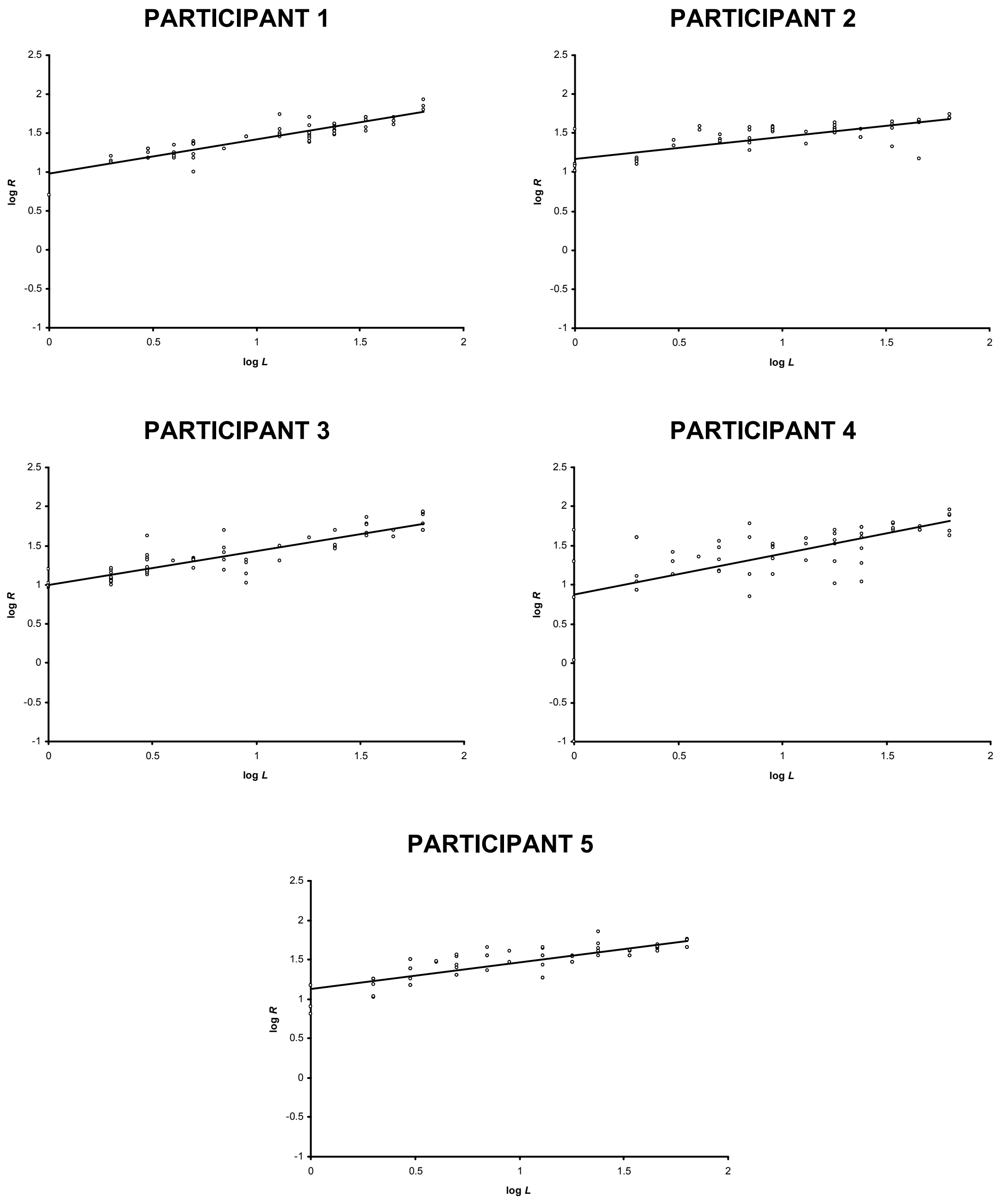


*Figure 54. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for grayscale squares with feedback in Experiment 10.*

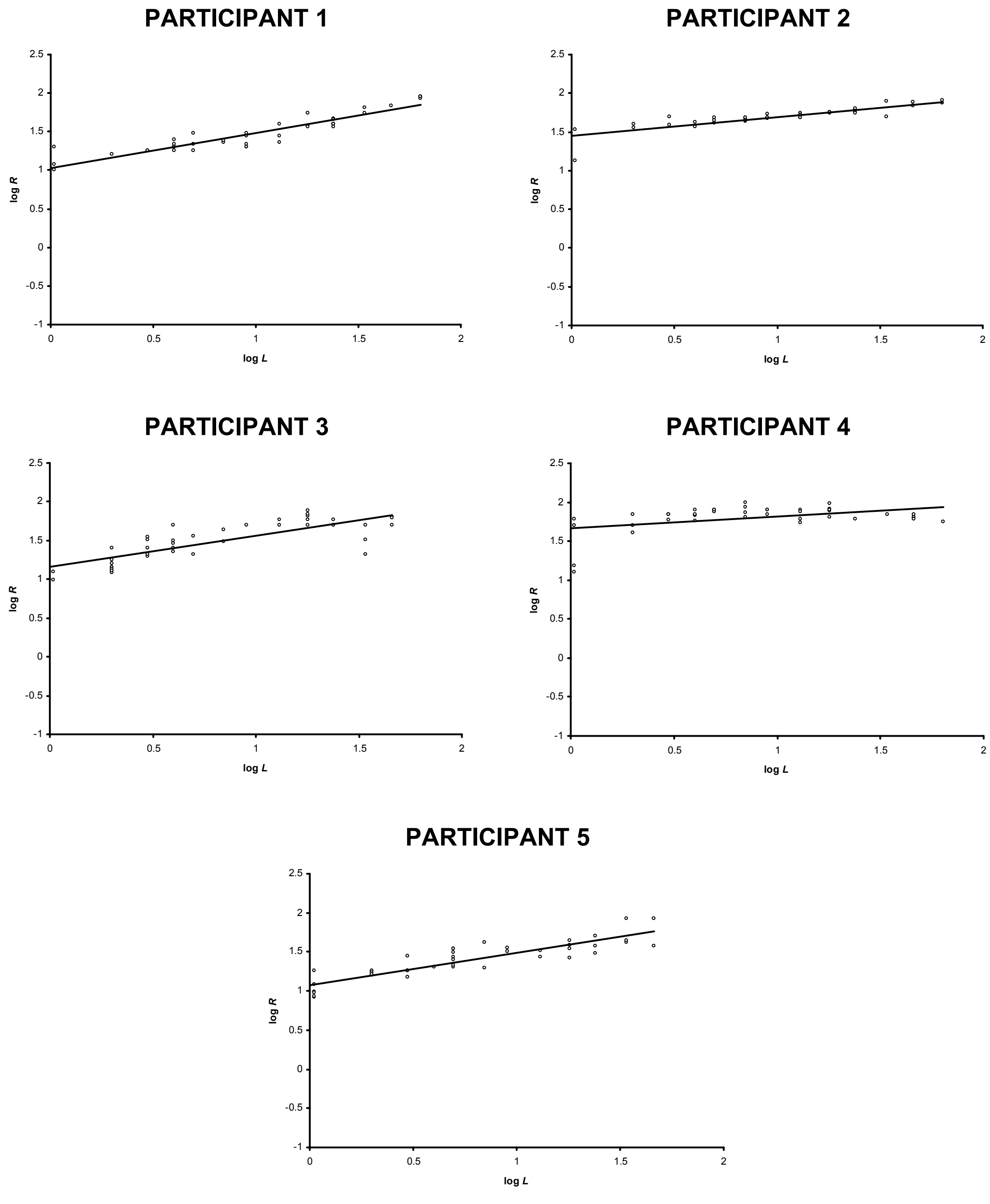


*Figure 55. Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for red squares without feedback in Experiment 10.*

Exponent Values

The average scaling exponent value for grayscale stimuli with feedback was 0.379 with *SD*/*M* = 0.172 and *H*:*L* = 1.533:1. For red stimuli without feedback, the average scaling exponent value was 0.333 with *SD*/*M* = 0.387 and *H*:*L* = 2.974:1. The average exponent ratio of grayscale to red stimuli was 1.333.

Intercept Values

The average scaling intercept for grayscale stimuli was 1.061 with *SD*/*M* = 0.078 and *H*:*L* = 1.193:1. For red stimuli, the average scaling intercept was 1.277 with *SD*/*M* = 0.211 and *H*:*L* = 1.612:1. The average intercept ratio of grayscale to red stimuli was 0.857.

Goodness of Fit Coefficients

The average goodness of fit coefficient for the regression line of the scaling data was 0.652 for grayscale stimuli, with *SD*/*M* = 0.245 and *H*:*L* = 2.005:1. For red stimuli, the average R^2 equaled 0.642 with *SD*/*M* = 0.407 and *H*:*L* = 4.418:1. The average ratio of grayscale to red stimulus R^2 values was 1.753.

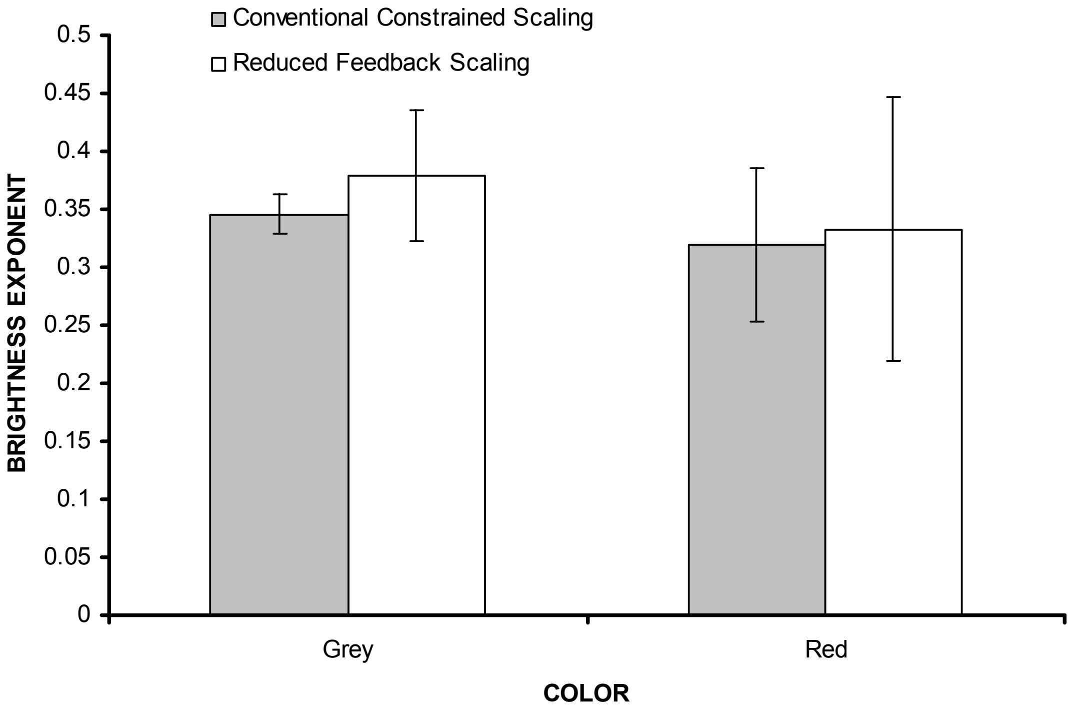
Comparison of Experiments 9 and 10

Although Experiment 9 represented noisy categorical constrained scaling, its results were closely aligned with the results from conventional constrained scaling. Thus, to compare the results of the present experiment's reduced feedback ratio with conventional constrained scaling results, Experiment 9 serves as a valid surrogate. In contrast, comparing the present experiment to a true conventional constrained scaling experiment, such as Experiment 7, would be fraught with potential confounds due to the extended duration of Experiment 7. Experiment 7 was a considerably longer experiment because of its use of grayscale, red, green, and blue stimuli. Experiment 9 and the present experiment are comparable designs that only utilize grayscale and red stimuli, thereby minimizing participant fatigue as a contributing factor to the results. By comparing the present experiment to Experiment 9, it was possible to determine the effects solely of the feedback reinforcement rate, without the need to account for potential performance differences due to the duration of the experiments.

Experiment 9 and the present experiment were compared as in previous experiments. The average exponent, intercept, and R^2 values and their variance were contrasted between the two experiments to determine if constrained scaling with a reduced feedback reinforcement rate was as effective as conventional constrained scaling. The comparison results are summarized in Figures 56 – 58.

Exponent Values

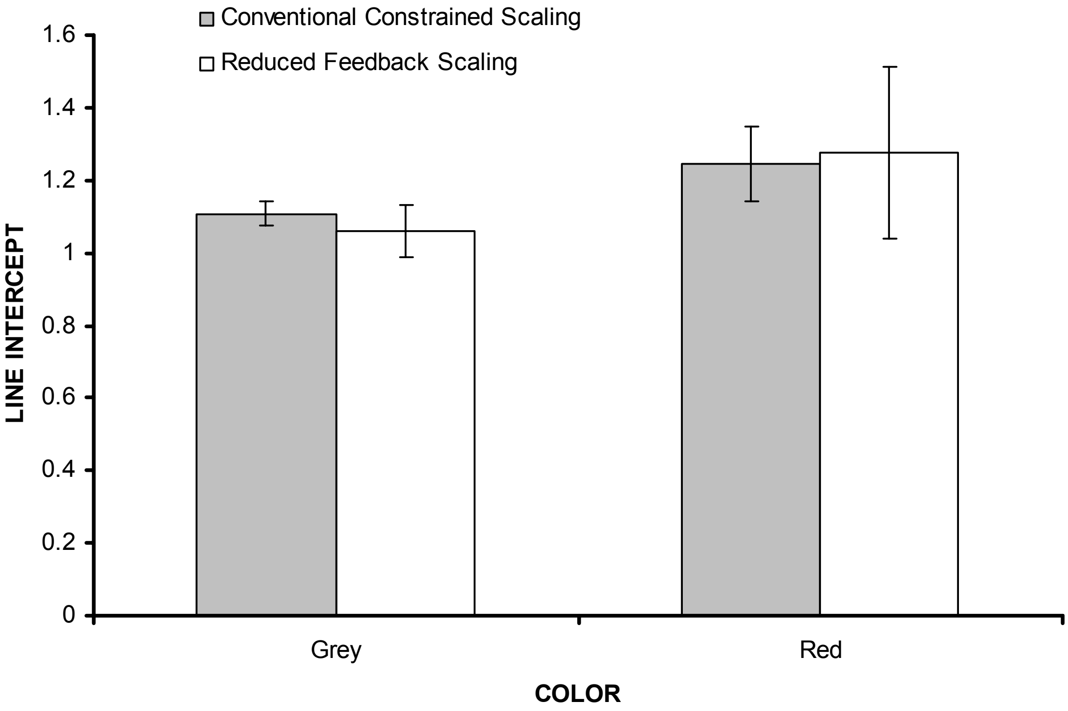
Figure 56 contrasts the brightness scaling exponent values for conventional constrained scaling and constrained scaling with a reduced feedback reinforcement rate. There was no significant difference between Experiment 9 and the current experiment for grayscale stimulus [$t(8) = 1.092, p = 0.153$] or red stimulus [$t(8) = 0.203, p = 0.422$] exponents. The present experiment produced significantly higher exponent variance for grayscale stimuli [$F(1,4) = 11.403, p < 0.05$] but not for red stimuli [$F(1,4) = 2.878, p = 0.165$]. For grayscale stimuli, the *SD*/*M* was 3.064 times greater and the *H*:*L* ratio was 1.349 times greater for the present experiment than for Experiment 9. For red stimuli, the



NOTE: Conventional constrained scaling is represented by noisy categorical constrained scaling from Experiment 9.

NOTE: The error bars represent the 95% confidence intervals.

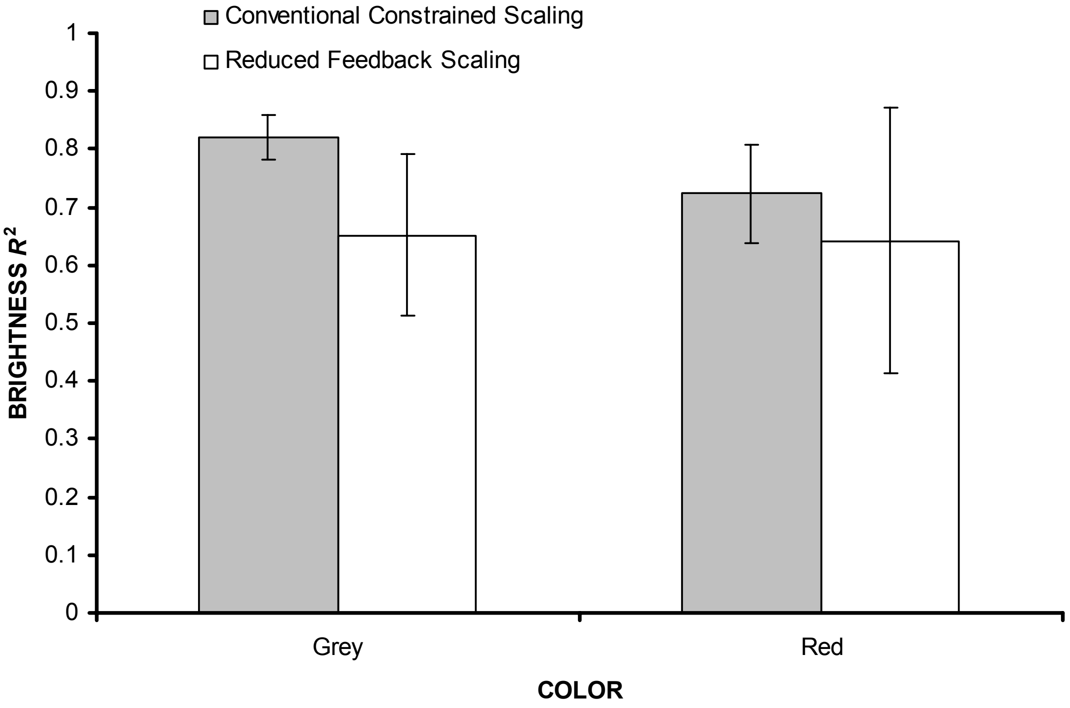
Figure 56. Comparison of exponents between Experiment 9 (representing conventional constrained scaling) and Experiment 10 (constrained scaling with reduced feedback).



NOTE: Conventional constrained scaling is represented by noisy categorical constrained scaling from Experiment 9.

NOTE: The error bars represent the 95% confidence intervals.

Figure 57. Comparison of line intercepts between Experiment 9 (representing conventional constrained scaling) and Experiment 10 (constrained scaling with reduced feedback).



NOTE: Conventional constrained scaling is represented by noisy categorical constrained scaling from Experiment 9.

NOTE: The error bars represent the 95% confidence intervals.

Figure 58. Comparison of goodness of fit coefficients between Experiment 9 (representing conventional constrained scaling) and Experiment 10 (constrained scaling with reduced feedback).

SD/*M* was 1.626 times greater and the *H*:*L* ratio was 1.494 times greater for the present experiment.

Intercept Values

Figure 57 illustrates the relationship between brightness scaling intercepts for conventional constrained scaling versus constrained scaling with a reduced feedback reinforcement rate. The average line intercepts did not differ significantly for grayscale [$t(8) = -1.138$, $p = 0.144$] or red [$t(8) = 0.246$, $p = 0.406$] stimuli. The variance was marginally greater in the present experiment for both grayscale [$F(1,4) = 4.425$, $p = 0.089$] and red [$F(1,4) = 5.145$, $p = 0.071$] stimuli. For the grayscale scaling intercepts, the *SD*/*M* in the present experiment was 2.193 times greater and the *H*:*L* ratio was 1.088 times greater than in Experiment 9. For the red intercepts, the *SD*/*M* was 2.209 times greater and the *H*:*L* ratio was 1.256 times greater.

Goodness of Fit Coefficients

Figure 58 illustrates the relationship between R^2 values for conventional constrained scaling and constrained scaling with a reduced feedback reinforcement rate. The average R^2 values did not differ significantly for grayscale [$t(8) = -0.654$, $p = 0.266$] or red [$t(8) = -0.654$, $p = 0.266$] stimuli. As in previous experiments, the variance of the goodness of fit coefficients was not analyzed.

General Discussion

Reducing the rate at which feedback was provided during testing sessions did not affect the average exponent or intercept values for scaling brightness. However, it increased, in some cases quite significantly, participants' scaling variability. Since the original tenet of constrained scaling was its effectiveness at reducing scaling variability, the benefits of constrained scaling are largely lost when the feedback reinforcement rate is decreased during testing trials.[50] Based on the findings in this experiment, it is not a profitable strategy to increase scaling efficiency by decreasing interspersed training trials. Previous experiments have demonstrated that participants learn a scale with fewer training trials than originally estimated. The present experiment demonstrates that it is necessary for participants to be reinforced in their scale learning. The key to optimizing constrained scaling as a method resides in the shortening of the overall number of trials, not in the reduction of reinforcement training trials.

[50] Another advantage of constrained scaling might be exponent and intercept consistency. Experiments 6 and 7 revealed notable differences in these values between magnitude estimation and constrained scaling. Those differences were not present in the comparison between conventional constrained scaling and constrained scaling with a reduced feedback reinforcement rate.

EXPERIMENT 11

Brief Introduction to Experiments 11 – 13

Experiment 11 deals with perceptual scaling, whereas Experiments 12 and 13 extend constrained scaling to assess subjective perceptions. Because Experiments 11 – 13 are methodological complements to each other, I first provide a brief overview of them together in this chapter. I subsequently dedicate the remainder of this chapter to a discussion of Experiment 11, and provide separate chapters for Experiments 12 and 13.

Experiments 11 – 13 share a technique known as methodological triangulation (Denzin, 1970). Triangulation simply refers to the formal process of examining a phenomenon from more than one perspective in order to get a complete, or nearly complete, understanding of that phenomenon. Triangulation is commonly used in social scientific research as a between-method rationale for combining quantitative and qualitative data (Blaikie, 1993). A traditional between-method triangulation study would involve using dissimilar research methods to measure the same underlying phenomenon. The specific intent of the multiple methods is to ensure that most or all relevant dimensions and nuances of a phenomenon are recorded for subsequent analysis. In contrast, Denzin defined within-method triangulation as the use of variations of a single empirical method to measure an underlying phenomenon.

Here I have extended Denzin's (1970) categorization to incorporate yet another important type of methodological triangulation. In Experiments 11 – 13, I have used constrained scaling to establish the relationship between two types of phenomena. These experiments are special cases of triangulation in which a single empirical method is used to measure more than one type of phenomenon. I have called this approach *multiphenomenal within-method triangulation*. Multiphenomenal within-method triangulation means that I have used variations within a single method of constrained scaling to look at the relationship between two different types of scaling phenomena. Table 19 recasts Denzin's original methodological triangulation to encompass more than a single measured phenomenon. His existing taxonomy comprises what I refer to as *uniphenomenal within-method* and *between-method triangulation*, while the new taxonomy adds *multiphenomenal within-method* and *between-method triangulation.*

Introduction to Experiment 11

Experiment 11 used multiphenomenal within-method triangulation to determine the basic learnability of two different perceptual constrained scales. This was a test to see if constrained scaling as a single method would map two different scaling phenomena. The goal was to determine if training on two different brightness scaling exponent values translated into two likewise scaled brightness scaling exponents for untrained stimuli.

This was a two-part experiment. Both parts of the experiment were similar to the method adopted in Experiments 8 – 10, in that brightness scale training was provided using grayscale stimuli with feedback and testing was performed using red stimuli without feedback. In one part of the experiment, participants were trained to respond to grayscale stimuli according to a brightness scale with an exponent of 0.20:

$$R = 20L^{0.20} . \qquad (28)$$

Table 19. Types of methodological triangulation.

	Same Research Method	**Different Research Methods**
Single Phenomenon	uniphenomenal within-method triangulation	uniphenomenal between-method triangulation
Multiple Phenomena	multiphenomenal within-method triangulation	multiphenomenal between-method triangulation

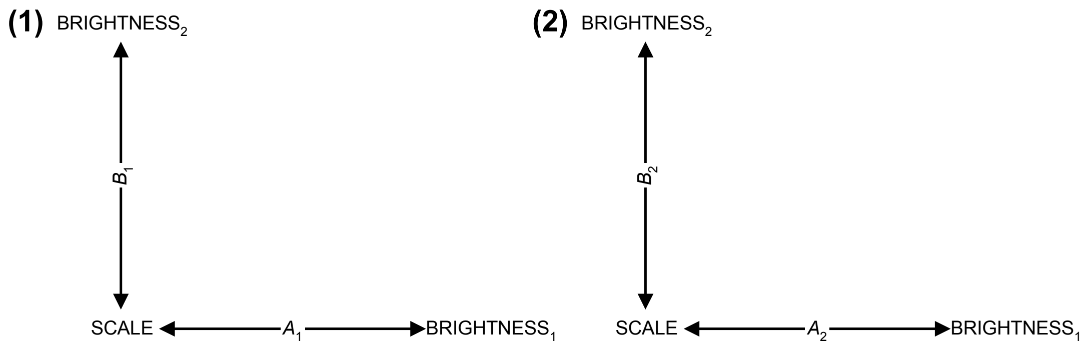


Figure 59. Representation of the two-stage triangulation in Experiment 11.

In the second part of the experiment, following a one-week interstice, the same participants were trained to respond to the identical grayscale stimuli according to a brightness scale with an exponent of 0.46:[51]

$$R = 7L^{0.46}. \tag{29}$$

These two exponent values were equidistant from 0.33, the natural brightness scaling exponent (Stevens, 1975) used in previous experiments.[52]

Figure 59 depicts the two stages of Experiment 11. The variables labeled A_1 and A_2 represent the relationship between the brightness of the grayscale squares (BRIGHTNESS_1) and the magnitude scale. This relationship is equivalent to the slope of the regression line obtained for each participant when asked to estimate the brightness of grayscale stimuli after initial training. Perfect learning would result in A_1 equal to 0.46, equivalent to the exponent value for training trials in Equation 29. Likewise, the expected value of A_2 would be equal to or near the training trial exponent of 0.20 in Equation 28.[53] The variables labeled B_1 and B_2 represent the relationship between the brightness of the red squares (BRIGHTNESS_2) and the magnitude scale for the different exponent values.

[51] The order of the exponent training was counterbalanced across participants to prevent possible order or practice effects.

[52] Note that the intercept values were varied in order to keep the upper end of the feedback scale constant.

[53] The assignment of Equation 29 to A_1 and Equation 28 to A_2 is arbitrary. Remapping the equations to different A variables would not change the conclusions made in this section.

Since brightness is not perceived entirely equivalently for grayscale and red stimuli, the *B* values are expected to differ from the *A* values. Owing to the low interparticipant variability afforded by constrained scaling, it is expected that one participant's *B* values will be very similar to another participant's *B* values. It is, however, not expected that B_1 and B_2 are equivalent. Since B_1 reflects the training influences of A_1 and since B_2 reflects the training influences of A_2, B_1 and B_2 should differ considerably.

The key to triangulation in Experiment 11 is the relationship of the use of the scale with an exponent of 0.20 to the use of the scale with an exponent of 0.46. This experiment is designed to determine if the constrained scale is applied in the same manner for the novel red stimuli, even when the training scale exponents are different. The relationship under review can be expressed more formally as:

$$\frac{A_1}{B_1} = \frac{A_2}{B_2}, \tag{30}$$

which is to say that the ratio of A_1 to B_1 is expected to be equivalent to the ratio of A_2 to B_2. In other words, the scale used for Part 1 of the experiment is expected to be applied in the same manner as the scale used in Part 2. If this conclusion is *not* true, then it brings into question issues surrounding the participants' ability to generalize a learned number scale to perceived magnitudes. An inherent assumption in constrained scaling is that people can learn a naturalistic scale, which they can in turn use to estimate their perceptual magnitude. The present experiment tests the assumption that constrained scaling works equivalently for different learned scales. If participants cannot use the scale with an exponent of 0.20 with the same degree of effectiveness as the scale with an exponent of 0.46, or vice versa, this suggests that some scales are inherently more learnable than others.

In the present experiment, the scaling ratio was expected to differ between the two conditions. I hypothesized that A_1 would roughly be equivalent to 0.46 and that A_2 would roughly be equivalent to 0.20. Substituting these values into Equation 30 produces the following:

$$\frac{0.46}{B_1} = \frac{0.20}{B_2}, \qquad (31)$$

which simplifies to:

$$B_1 = 2.3B_2\,. \qquad (32)$$

Thus, I hypothesized that the scale exponent for the red stimuli would be more than double for the learned scales with exponents of 0.46 than for the learned scales with exponents of 0.20.

Method

Participants

Five volunteers with self-reported normal color vision were enlisted to participate in the experiment. The volunteers were drawn from participants who had previously taken part in brightness scaling experiments, although not in the previous month. The participants were compensated $5.00 for each half session in which they participated.

Apparatus and Stimulus Materials

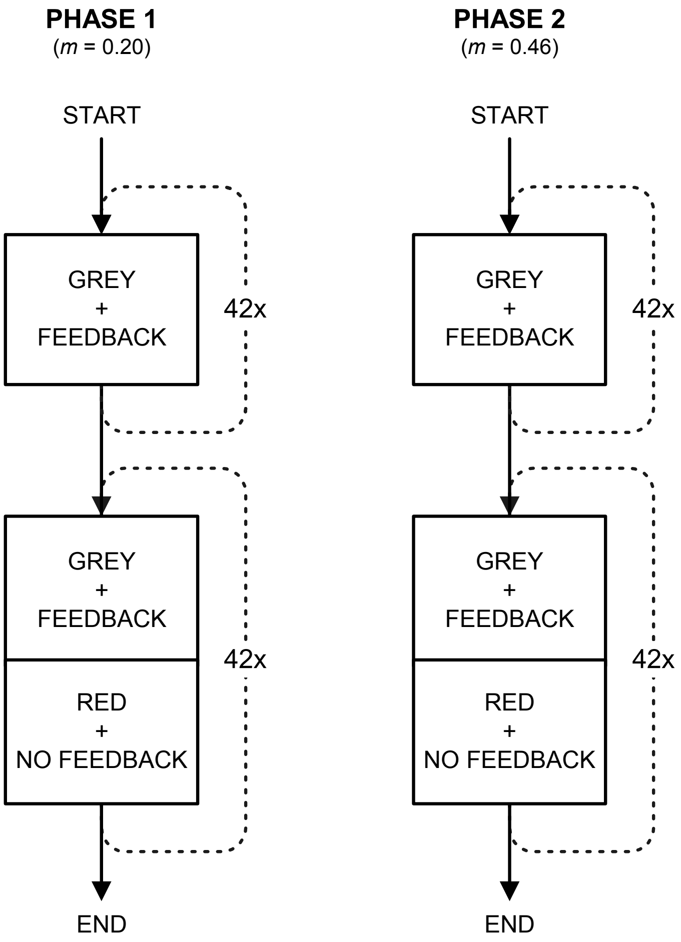
The experimental control software used in previous experiments was modified to pair participants with the appropriate 0.20 and 0.46 training exponents in successive sessions. Participants were counterbalanced, so that odd-numbered participants began with the 0.20 training exponent, and even-numbered participants began with the 0.46 training exponent. In the second phase of the experiment, odd-numbered participants were paired with the 0.46 training exponent, while even-numbered participants were paired with the 0.20 training exponent. As in Experiment 10, the stimulus materials consisted of grayscale and red brightness stimuli at 14 luminous intensity levels.

Design and Procedure

The experiment consisted of two phases, as depicted in Figure 60. As discussed above, for one part of the experiment, participants were trained to scale the brightness of grayscale squares with a feedback exponent of 0.20. Participants were subsequently tested on the brightness of red squares. For the other part of the experiment, participants were trained to scale the brightness of red squares with a feedback exponent of 0.46 and were subsequently tested on the brightness of red squares. The order of the two phases was alternated between participants. One week separated the two test phases.

Results and Discussion

The data were analyzed as in previous experiments, with separate results for the exponent, intercept, and R^2 values. These results were calculated for each phase of the experiment and then compared to determine the scaling relationship between the two sets of scaling exponents. The results are summarized in Tables 20 – 21 and in Figures 61 – 64, and are described in detail below.



NOTE: Participants may begin in Phase 1 or 2. Following a one-week interstice, participants complete the other phase.

Figure 60. Schematic flow of the two phases in Experiment 11.

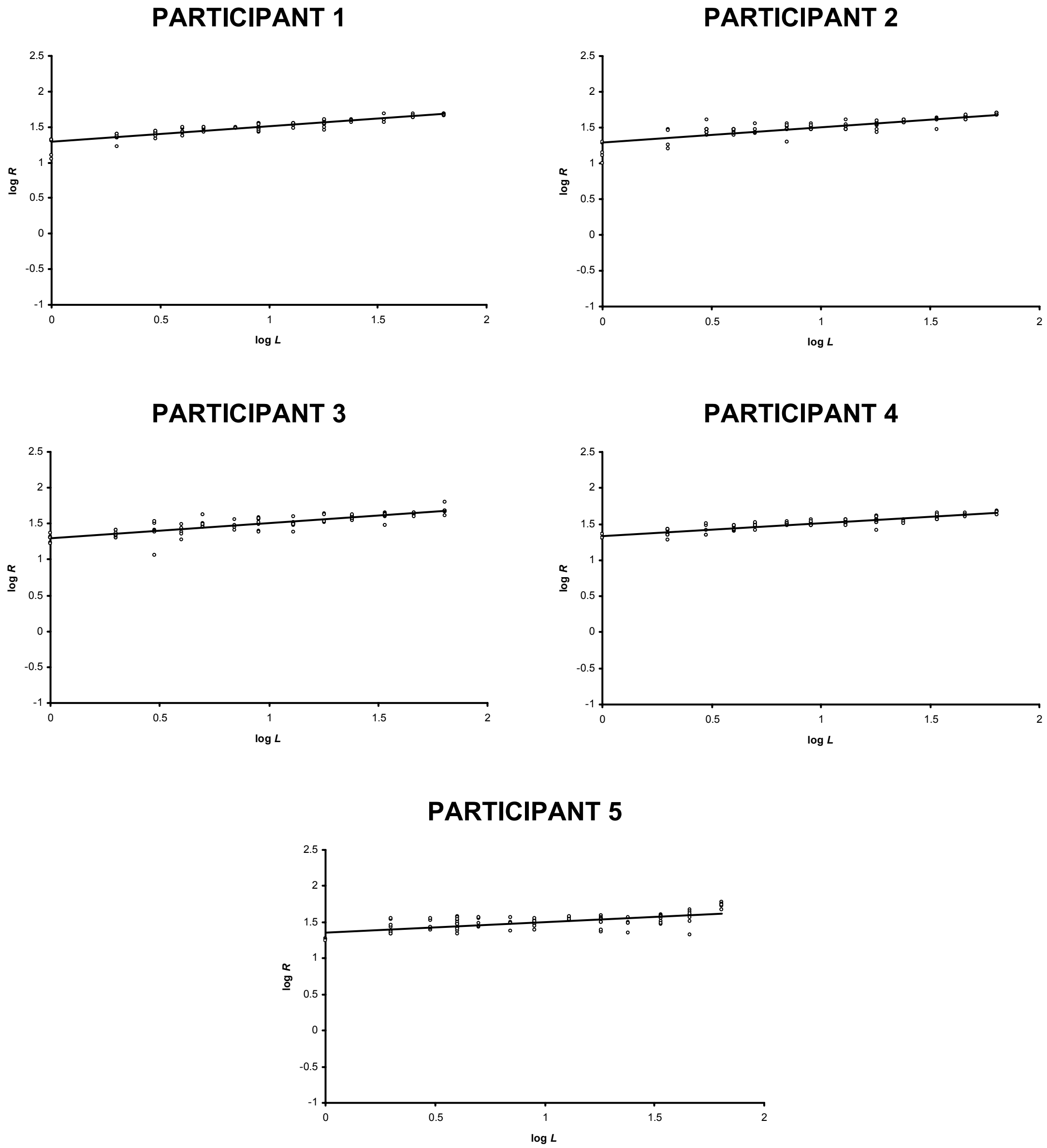
Table 20. Summary of participants' brightness scaling for 0.20 exponent training in Experiment 11.

	Grayscale Brightness + Feedback			**Red Brightness + No Feedback**			**Ratio of Grayscale:Red Brightness**		
P	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***	**Exponent**	**Intercept**	***R^2***
1	0.216	1.290	0.845	0.225	1.261	0.862	0.960	1.023	0.980
2	0.212	1.294	0.719	0.292	1.380	0.893	0.726	0.938	0.805
3	0.210	1.294	0.728	0.192	1.320	0.688	1.094	0.980	1.058
4	0.180	1.328	0.836	0.195	1.334	0.928	0.923	0.996	0.901
5	0.143	1.369	0.457	0.239	1.489	0.826	0.598	0.919	0.553
M	0.192	1.315	0.717	0.229	1.357	0.839	0.860	0.972	0.860
SD	0.031	0.034	0.157	0.041	0.085	0.093	0.197	0.042	0.195

Table 21. Summary of participants' brightness scaling for 0.46 exponent training in Experiment 11.

	Grayscale Brightness + Feedback			**Red Brightness + No Feedback**			**Ratio of Grayscale:Red Brightness**		
P	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**
1	0.389	0.964	0.834	0.446	0.869	0.869	0.872	1.109	0.960
2	0.407	0.960	0.759	0.492	1.035	0.855	0.827	0.928	0.888
3	0.423	0.921	0.785	0.371	1.017	0.777	1.140	0.906	1.010
4	0.358	0.973	0.805	0.441	0.933	0.913	0.812	1.043	0.882
5	0.277	1.082	0.627	0.428	1.161	0.889	0.647	0.932	0.705
M	0.371	0.980	0.762	0.436	1.003	0.861	0.860	0.983	0.889
SD	0.058	0.060	0.080	0.043	0.111	0.052	0.178	0.089	0.116

GRAYSCALE STIMULI
TRAINING EXPONENT = 0.20



PARTICIPANT 1

PARTICIPANT 2

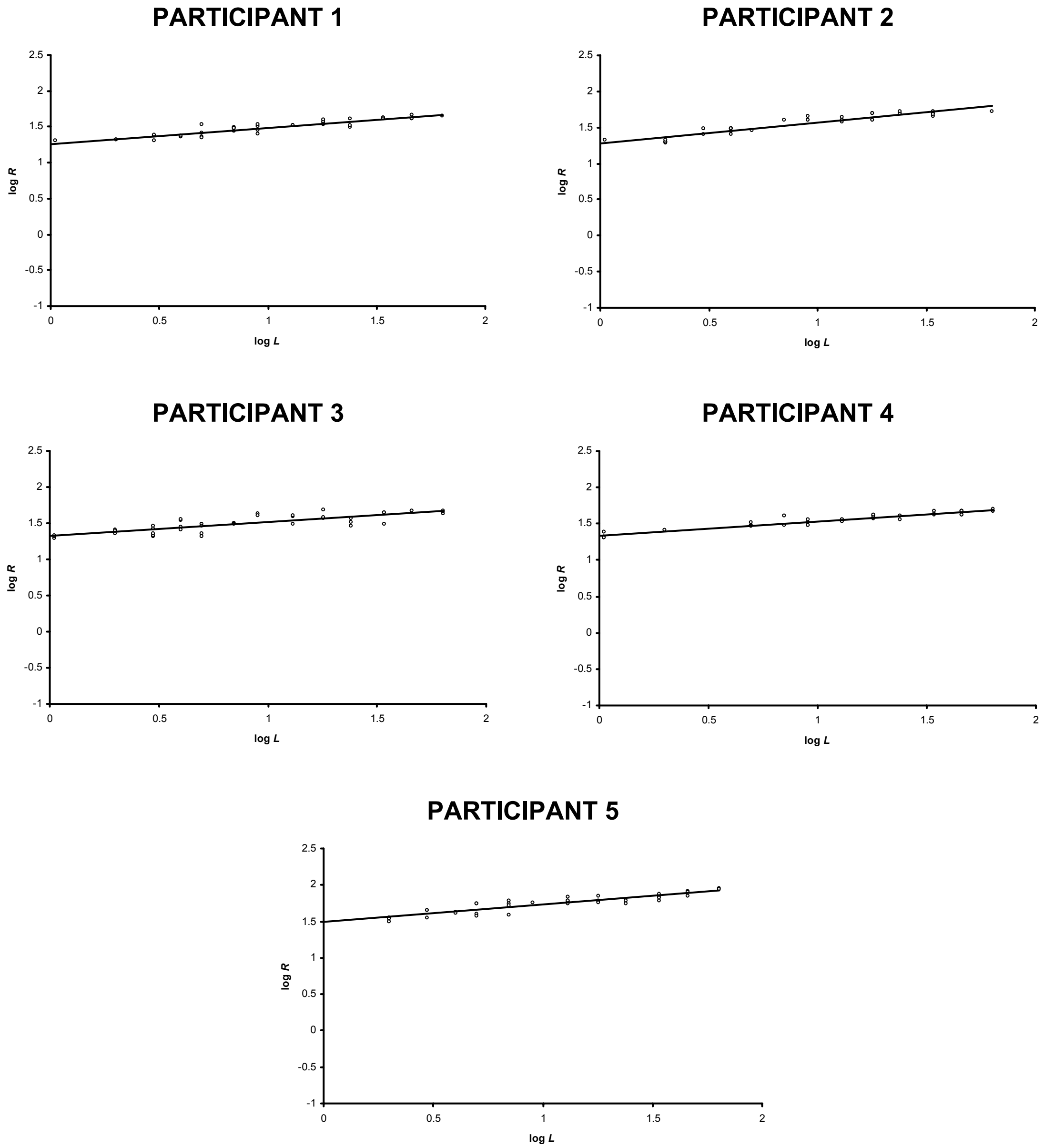
PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 61. *Logarithmic scatterplot and regression line for brightness in cd/m*^{2} *(*L*) and participant response (*R*) for grayscale squares with feedback for a training exponent value of 0.20 in Experiment 11.*

RED STIMULI
TRAINING EXPONENT = 0.20


PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

Figure 62. *Logarithmic scatterplot and regression line for brightness in cd/m*^{2} *(*L*) and participant response (*R*) for red squares without feedback for a training exponent value of 0.20 in Experiment 11.*

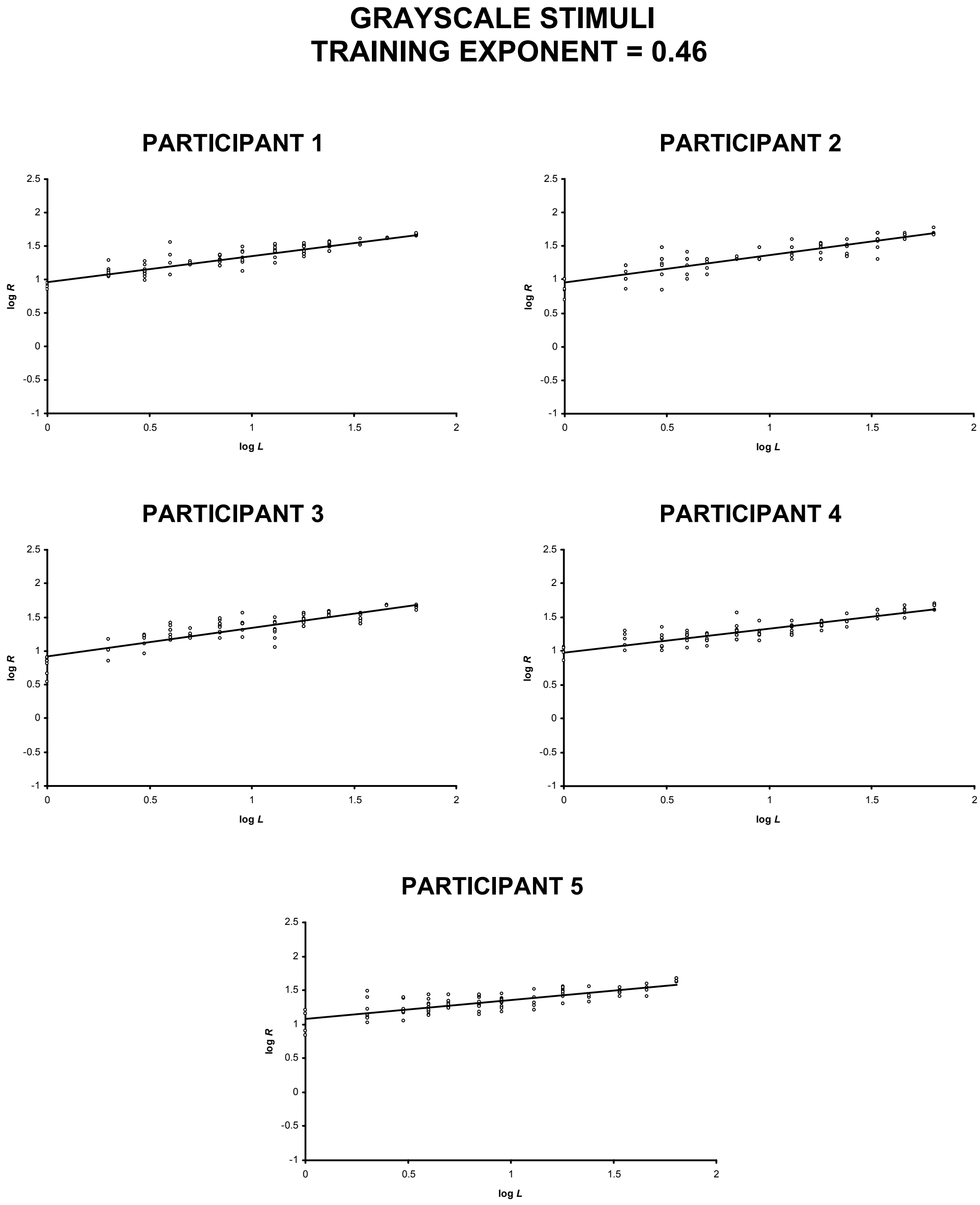


Figure 63. Logarithmic scatterplot and regression line for brightness in cd/m^2 *(*L*) and participant response (*R*) for grayscale squares with feedback for a training exponent value of 0.46 in Experiment 11.*

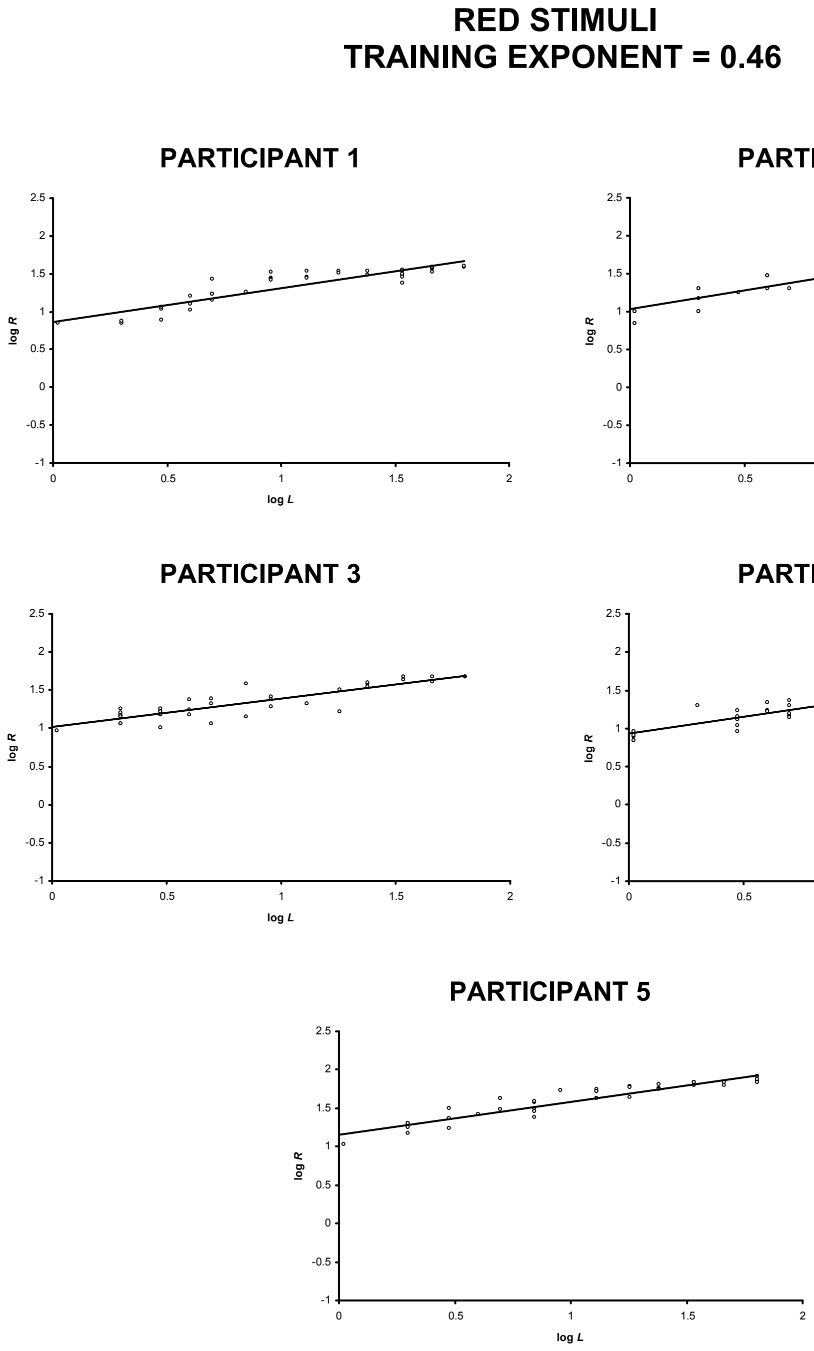


Figure 64. *Logarithmic scatterplot and regression line for brightness in cd/m^2 (*L*) and participant response (*R*) for red squares without feedback for a training exponent value of 0.46 in Experiment 11.*

Exponent Values

For the condition in which the training exponent equaled 0.20, the average scaling exponent for grayscale stimuli with feedback was 0.192 with *SD*/*M* = 0.161 and *H*:*L* = 1.510:1. For the red stimuli without feedback in the same condition, the average scaling exponent was 0.229 with *SD*/*M* = 0.178 and *H*:*L* = 1.521:1. The average ratio of grayscale to red stimulus exponents was 0.860. For the condition in which the training exponent equaled 0.46, the average scaling exponent for grayscale stimuli was 0.371 with *SD*/*M* = 0.156 and *H*:*L* = 1.527:1. For the red stimuli in the same condition, the average scaling exponent was 0.436 with *SD*/*M* = 0.100 and *H*:*L* = 1.326:1. The average ratio of grayscale to red stimulus exponents was 0.860. The average exponent values cannot be compared to the exponent values from previous brightness scaling experiments, because the training exponents do not match. However, the variability in terms of *SD*/*M* and *H*:*L* was comparable to the levels found in the previous brightness scaling experiments.

Intercept Values

For the condition in which the training exponent equaled 0.20, the average scaling intercept for grayscale stimuli was 1.315 with *SD*/*M* = 0.026 and *H*:*L* = 1.061:1. For the red stimuli, the average scaling intercept was 1.357 with *SD*/*M* = 0.063 and *H*:*L* = 1.181:1. The average ratio of grayscale to red stimulus intercepts was 0.971. For the condition in which the training exponent equaled 0.46, the average scaling intercept for grayscale stimuli was 0.980 with *SD*/*M* = 0.062 and *H*:*L* = 1.175:1. The average ratio of grayscale to red stimulus intercepts was 0.983. The scaling variability was low for both conditions, indicative of the efficacy of constrained scaling when used with different training modalities. As expected, there was a tendency for the intercept to increase the lower the exponent value was. This finding was indicative of the different intercept training values, and both conditions provided good matches between the training intercepts and the participants' resultant intercept scaling values.

Goodness of Fit Coefficients

For the condition in which the training exponent equaled 0.20, the average goodness of fit coefficient was 0.717 for grayscale stimuli, with *SD*/*M* = 0.157 and *H*:*L* = 1.849:1. For red stimuli, the average R^2 was 0.839 with *SD*/*M* = 0.110 and *H*:*L* = 1.181:1. The average ratio of grayscale to red stimulus goodness of fit coefficients was 0.860. For the condition in which the training exponent equaled 0.46, the average R^2 for grayscale stimuli was 0.762 with *SD*/*M* = 0.105 and *H*:*L* = 1.330:1. For red stimuli, the average R^2 was 0.861 with *SD*/*M* = 0.060 and *H*:*L* = 1.175:1. For this condition, the average ratio of grayscale to red stimulus goodness of fit coefficients was 0.889. The average R^2 values were well within the range expected for constrained scaling, suggesting a successful implementation of the constrained scaling method in the two experimental conditions.

General Discussion

Recall that Equation 30 hypothesized that $\frac{A_1}{B_1} = \frac{A_2}{B_2}$, where A_1 is the exponent value for grayscale stimulus scaling for training with an exponent equal to 0.46, B_1 is the related exponent value for red stimulus scaling, A_2 is the exponent value for grayscale stimulus scaling for training with an exponent equal to 0.20, and B_2 is the related exponent value for red stimulus scaling. Substituting the actual values into Equation 30 results in the following equation:

$$\frac{0.371}{0.192} = \frac{0.436}{0.229}, \tag{33}$$

which simplifies as:

$$1.929 \approx 1.906. \tag{34}$$

Similarly, substituting the appropriate intercept values into Equation 30 yields the following equation:

$$\frac{0.980}{1.003} = \frac{1.315}{1.357}, \tag{35}$$

which simplifies as:

$$0.977 \approx 0.969. \tag{36}$$

As the ratios of the exponent and intercept values[54] clearly demonstrate, the training scale was applied in the same manner to the testing stimuli across the two conditions. This finding lends evidence to the notion that different learned scales are used in the same manner to estimate the magnitude of novel stimuli.

[54] The ratios were not calculated for goodness of fit coefficients, since the goodness of fit coefficients are indirect measures of the scale and are not informative to the direct scale comparisons in the present experiment.

EXPERIMENT 12

Introduction

The previous experiments in this dissertation share a common theme—to use constrained scaling to match a magnitude scale to a perceptual stimulus. In previous experiments, it was straightforward to show the relationship of a measurable physical stimulus to a perception of magnitude. When scaling psychophysical stimuli, there is usually a continuous stimulus magnitude and a continuous response magnitude. Such stimulus-response pairs lend themselves to the types of analysis heretofore described. The data are suited to graphing on a two-dimensional Cartesian plane, with the stimulus values serving as abscissae and the response values serving as the ordinates. Using log-log coordinates, the stimulus-response pairs typically form a linear relationship, making it possible to analyze the aggregate data in terms of a parsimonious regression analysis.

In the previous experiments, constrained scaling worked because people perceive physical stimuli in a highly similar way. For example, the way person *X* neurophysiologically perceives loudness is identical to the way person *Y* perceives it. There may be slight perceptual variations due to differences in hearing sensitivity, shape of the ear canal, or other factors. But, the process of loudness perception is largely invariant across humans. Similarly, there is a high degree of neurophysiological invariance in the way a person perceives the brightness of objects. Factors such as light or dark adaptation, age-related diminished visual acuity, color blindness, and other factors may contribute to differences between individual humans, but the mechanism by which humans encode and perceive brightness is largely biologically determined.

Constrained scaling operates in such a way that person *X*'s perception is scaled to the same scale used by person *Y*. The variability found in traditional magnitude estimation is therefore seen as an artifact of conventional scaling methods and not as a reflection of true individual differences.

Psychometric scaling of subjective experience is a much different phenomenon than psychophysical scaling. To begin, there often exists no underlying physical stimulus continuum. Consider a number of subjective factors that might be scaled: happiness, sadness, satisfaction, interest value, and so forth. What underlying physical stimulus could be used to trigger subjective responses suitable for scaling? One would not typically use loudness as an instigator of affective states such as happiness or sadness; nor would one find much merit in scaling cognitive factors such as the satisfaction or interest value elicited by pure tones of varying amplitudes. Subjective experience, whether it is primarily affective or cognitive in nature, is the product of complex phenomena that are not readily manipulated with the psychometrician's tool chest of experimental methods. Affective and cognitive responses are caused by the interplay of multiple dimensions of physical stimuli as well as internal processes caused by natural affinities and learned responses. While it is possible to develop a continuum of physical stimuli to elicit a subjective response, such a task would be fraught with a daunting variety of confounds and complications.

A second consideration in the scaling of subjective experience concerns the very nature of that experience. I have just discussed the difficulties of finding a continuum of physical stimuli that could elicit subjective experience. The next question becomes:

what subjective experience? Subjective experience, by its very nature, is a highly individualistic phenomenon. Suppose a researcher developed a set of stimuli that could elicit different levels of happiness in an individual. There is no reason to assume that one person's level of happiness would equal another person's level of happiness for that set of stimuli.[55] Whereas human biology dictates a largely homogenous set of perceptual experiences through our perceptual sense organs, affect and cognition are not bound by the same constraints as perception. There are definitely common neurophysiological underpinnings for affect and cognition, but the links that bind experience with particular affective or cognitive responses are not hardwired. The same phenomenon may elicit very different subjective responses across individuals.

Therein lays the main difference between Experiments 12 and 13 and the preceding experiments. Experiments 12 and 13 center on the topic of the psychometric scaling of subjective experience. The subjective experience in question is the amount of happiness a certain amount of money elicits. In this case, I have employed money as an underlying stimulus dimension. Money is an emotionally laden stimulus that conveniently falls along a continuum.

The groundwork for Experiments 12 and 13 is in a preliminary study by West and Ward (1998). There, constrained scaling was utilized as a method to determine individual differences in the subjective value of money. Whereas previous research on constrained scaling had focused exclusively on the scaling of perceptual phenomena,

[55] It can be argued that one persistent driving force (and perhaps folly) of our society is the assumption that everyone's level of happiness should be equal.

West and Ward's study addressed the scaling of affective and cognitive factors. The study was set up identically to the cross-modal perceptual studies in West et al. (2000), with the exception that participants were trained on a loudness scale and subsequently applied that scale to the perceived utility of various amounts of money.

Perceptual studies using constrained scaling demonstrate a consistent and significant decrease in interparticipant variability when using constrained scaling compared to traditional psychophysical methods such as magnitude estimation (West, 1996; West & Ward, 1994; West et al., 2000). The results for the scaling of subjective experience reveal the opposite effect. In West and Ward (1998), interparticipant variability actually increased when using constrained scaling to assess the subjective value of money. The authors noted that this result was expected. Since constrained scaling offers a truer scale expression than other methods of a person's experience of magnitude, one would expect:

1. Decreased interparticipant variability for psychophysical domains, in which perception is a largely hardwired translation of physical stimulus information into a scale;
2. Increased interparticipant variability for subjective domains to reflect the true individual differences in subjective experience.

The latter point is the crucial assumption when applying constrained scaling to subjective phenomena. Where there are real differences between individuals, an accurate measurement scale will capture those differences, while minimizing measurement differences where no real interparticipant differences exist.

This duplicitous definition presents something of an ironic twist for the scaling community. A calibrated scale must, on the one hand, virtually eliminate all variability when measuring objective psychophysical processes, but, on the other hand, it must also highlight variability when measuring subjective psychometric processes. The calibrated scale must serve seemingly opposite ends.

Consistent with the findings in West and Ward (1998), it is my hypothesis that constrained scaling will exhibit greater interparticipant variability than conventional scaling for subjective measures. I conducted two experiments to test this hypothesis. In the present experiment, participants were displayed a sum of money and asked to scale on a 100-point scale how happy that sum of money would make them if they were to win it.[56] This experiment provided a simple baseline of human performance on the task using magnitude estimation. In Experiment 13, the magnitude estimation task is contrasted with a constrained scaling task for the same stimulus set. Experiment 13 borrows the triangulation method from Experiment 11 to determine if participants could consistently apply learned perceptual scales cross-modally to a psychometric arena.

The constrained scaling task in Experiment 13 will be discussed in greater depth in the next chapter. The present chapter continues with a discussion of the method and results for the magnitude estimation experiment.

[56] *Happiness* is used as a convenient label that might more appropriately be called the subjective utility of money.

Method

Participants

Six participants were enlisted for the experiment. Since the stimuli did not involve perceptual stimuli, no screening was necessary to ensure normal hearing or color vision. To prevent carryover effects, the participants were recruited from people who had not previously participated in scaling experiments. The participants were paid $5.00 for volunteering to take part in the experiment.

Apparatus and Stimulus Materials

The experimental control software used in previous experiments was modified to display a monetary sum instead of a colored square. The sum was kept on the screen until the participant selected a happiness level on the 100-point scale. The stimuli consisted of monetary sums ranging from $50.12 to $1,000,000.00, calculated according to the following equation:

$$M = 10^{x/10}, \tag{37}$$

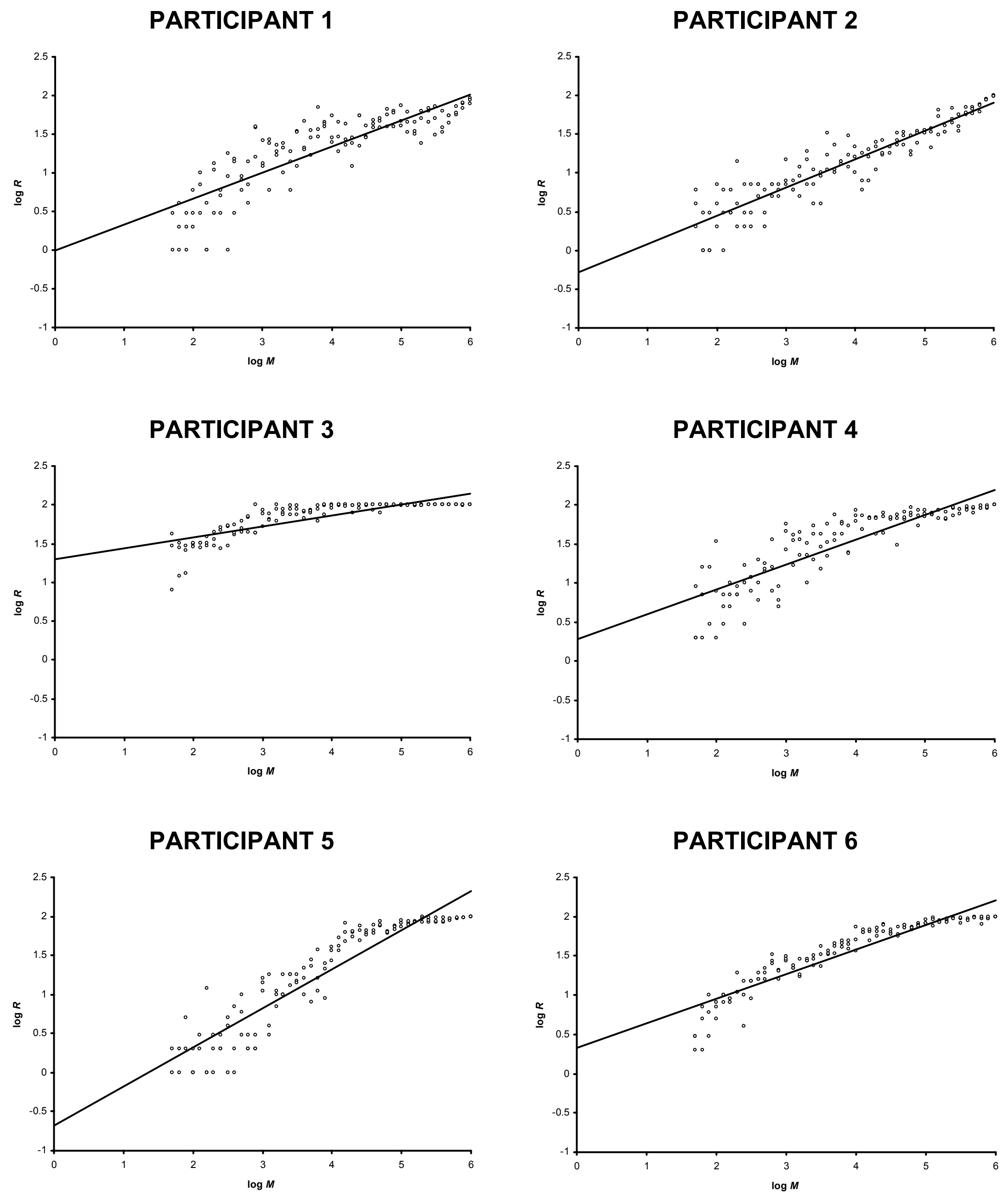
where M is the monetary sum and where x ranges from 17 to 60 in whole-number increments.

Design and Procedure

There were three rounds of the experiment. In each round, the complete set of monetary sums was presented in random order, to which participants responded using the 100-point scale with a rating of their level of happiness if they had won the displayed amount of money.

Table 22. Summary of participants' happiness scaling using magnitude estimation in Experiment 11.

P	**Exponent**	**Intercept**	***R*²**
1	0.337	-0.007	0.713
2	0.364	-0.282	0.854
3	0.141	1.300	0.653
4	0.320	0.278	0.750
5	0.502	-0.685	0.861
6	0.320	0.278	0.750
M	0.331	0.147	0.764
SD	0.116	0.673	0.081



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

PARTICIPANT 6

*Figure 65. Logarithmic scatterplot and regression line for money (*M*) and participant response (*R*) using magnitude estimation in Experiment 12.*

Results

The results were analyzed as in previous experiments. The scaling exponent and intercept values as well as the goodness of fit coefficient were calculated. The results are summarized in Table 22 and Figure 65.

Exponent Value

The average exponent value for the happiness elicited by money was 0.331 with *SD*/*M* = 0.349 and *H*:*L* = 3.560:1. This level of variability was considerably higher than the variability found in constrained scaling experiments, and was comparable to the levels of variability found in the magnitude estimation of color brightness in Experiment 6 and in magnitude estimation experiments in West et al. (2000).

Intercept Value

The average intercept value for happiness elicited by money was 0.147 with *SD*/*M* = 0.458 and *H*:*L* = -1.898. Note that the negative highest-to-lowest ratio was the result of some participants having an intercept in logarithmic space that was below 0. By definition, the logarithm of any number between 0 and 1 produces a negative number. A negative highest-to-lowest ratio must be interpreted as a potentially infinitely large number, indicative of a great deal of variability in intercept scoring by participants. Although participants were within the normal variability range for scaling exponents in magnitude estimation,[57] the intercepts exhibited much higher variability, which is likely a reflection of individual differences in the scaling of subjective happiness.

Goodness of Fit Coefficient

The average goodness of fit for the regression line of money-to-happiness scaling was 0.764 with *SD*/*M* = 0.106 and *H*:*L* = 1.319:1. There was a strong goodness of fit to the data, although a visual inspection of Figure 65 indicates that the goodness of fit might be even better for nonlinear curve fitting. Several of the participants exhibited sharp initial rises followed by a ceiling effect. There was a sharp ascent in the response values at the lower end of the monetary scale for several participants (see, especially, participants 1, 3, and 6 in Figure 65), and there was a flattening or tapering off of response values at the upper end of the monetary scale for several participants (see participants 3 – 6 in Figure 65). Such curvilinear data are indicative of the complex subjectivity behind happiness that is unlikely to be the product of a single factor such a monetary sum. For many participants, there is the point at which a little money would make them a little happier, more money would make them quite happy, and a large amount of money would have diminishing returns in terms of increasing their happiness.

[57] To understand that a negative number is a large number for the highest-to-lowest ratios, consider a hypothetical case in which the highest non-logarithmic value in a list was 5. If the lowest value were 3, *H*:*L* would equal log 5/log 3 or 0.699/0.301, which equals 2.322:1. If the lowest value were 1.1, *H*:*L* would equal log 5/log 1.1 or 0.699/0.041, which equals 17.049. If the lowest value were 0.9, one would expect the *H*:*L* ratio to be even greater. But, when translating to logarithmic space, the *H*:*L* becomes log 5/log 0.9 or 0.699/-0.046, which equals -15.196. This value, although negative, indicates a greater disparity between the highest and lowest values than was the case with the other *H*:*L* ratios.

Anecdotal evidence from discussion with Participant 4 indicated that he translated the money in his head into tangible things that he could buy. Thus, the lower range of money would buy certain material niceties (e.g., a computer or a vacation). As the range increased, the material items became larger and more appealing (e.g., a car or a house). At some point, the monetary scale exceeded the material items that the participant would like to possess, thus resulting in a flattening of the scale. The money represented the tangible material items plus "money in the bank," the latter being a perk that did not facilitate happiness in the participant.

General Discussion

The present experiment demonstrated the magnitude estimation of the relationship between the amount of money won and the level of subjective happiness afforded by that money. Participant results for exponent and R^2 values did not differ from the expected results for magnitude estimation in the perceptual domain of color brightness scaling. Nonetheless, there was an especially high degree of variability in the intercept scaling values. Moreover, a closer examination of the scaling indicated that it did not always follow the standard Power Law form. The next experiment pits these results against psychometric scaling using constrained scaling to see how training on a perceptual scale affects responses on a subjective scale.

EXPERIMENT 13

Introduction

The purpose of the present experiment is to determine if a perceptual training scale can be applied to novel subjective stimuli. A subjective scale cannot be used for training, because subjective matters are, by definition, individual. It would not, for example, make sense to train participants on a happiness scale, because there is no basis to assume that two people's subjective mappings of happiness are equivalent. A reasonable compromise solution is to train individuals on a perceptual scale in order to provide participants with a mapping from internal magnitude states to an external scale. Once participants have learned this mapping, it is assumed that this mapping will also allow them to translate their subjective mental states to an external scale.

One possible consequence of this constrained scaling extension, as noted in Experiment 12, is that the increased fidelity afforded by perceptual scale training may actually increase subjective scaling sensitivity to individual differences. To determine the ability of constrained scaling to detect individual differences, the variability from the present experiment is compared to the variability levels found for happiness scaling using magnitude estimation in Experiment 12. An increase in variability compared to magnitude estimation would be strong evidence that constrained scaling had increased the participants' ability to map subjective mental magnitudes to the numeric scale in a way reflective of individual differences.[58]

[58] Another possibility is simply that the perceptual scale confounded participants' ability to map their internal magnitude states to the numeric scale. Any scaling confusion would be expected to interfere with both the subjective and the perceptual scaling results.

The present experiment features the subjective counterpart to the perceptual triangulation in Experiment 11. Participants were trained on two sets of perceptual exponents for grayscale stimuli. Then, they applied the learned scale to measure their happiness in response to the monetary stimuli from Experiment 12. Figure 66 depicts the two training and testing phases in the present experiment. As in Experiment 11, the *A* values represent the exponents of the learned scale for the brightness of grayscale squares. Here, the *B* values represent the exponents of the scale used to scale the happiness elicited by certain amounts of money. The two *A* and *B* values represent separate training on brightness scales with exponent values of 0.20 (A_2) and 0.46 (A_1).

Like Experiment 11, the present experiment is based on multiphenomenal within-method triangulation. The use of the multiphenomenal within-method triangulation provided the opportunity to determine if the learned perceptual scales are applied in a consistent manner. Triangulation serves as a litmus test of the cross-modal perceptual-to-subjective scaling method. The goal in using triangulation is to determine that constrained scaling was being applied to scale subjective experience. Specifically, it was hypothesized that the ratio of a learned brightness scale to the scaling of the perceived utility of money holds constant for both brightness training scales. In Experiment 11, I found strong agreement in the *B* values between participants as a reflection of the trained *A* values. In the present experiment, I expected to find little agreement between participants on the individual *B* values, reflecting individual differences in the subjective utility of money. However, if the constrained scale revealed true differences, I expected

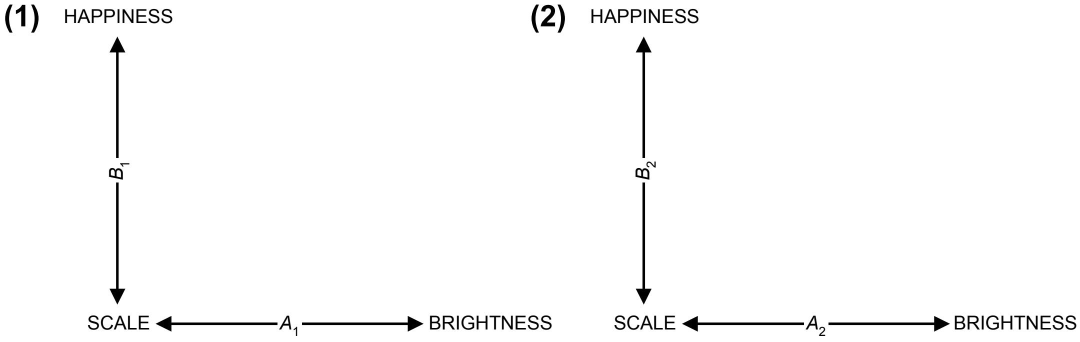


Figure 66. Representation of the two-stage subjective magnitude triangulation in Experiment 13.

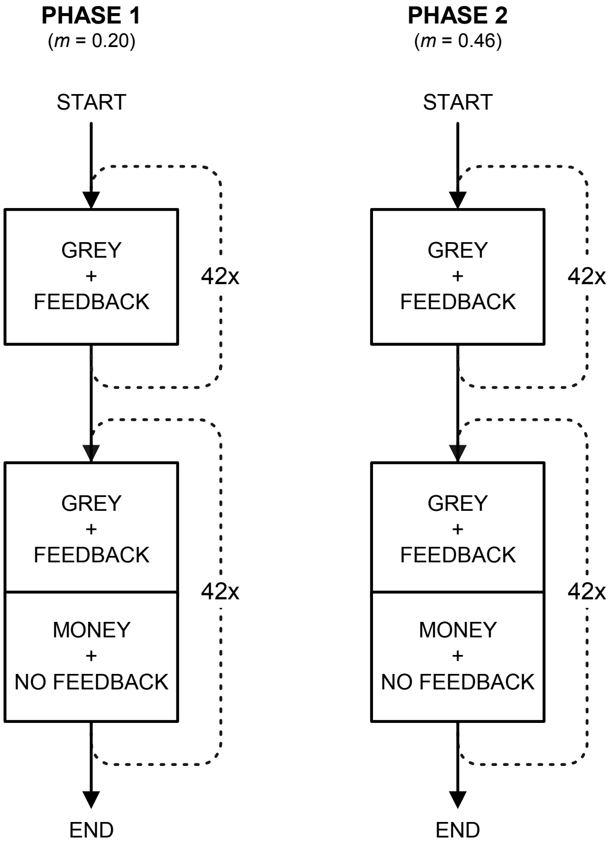


Figure 67. Schematic flow of the two phases in Experiment 13.

those differences to be evident even when a different constrained scale is used. Thus, I expected that the B_1 and B_2 values would be consistent within each participant but not between different participants. As in Experiment 11, I hypothesized that the ratio of A_1 to B_1 would be the same as the ratio of A_2 to B_2.

Method

Participants

Five participants with self-reported normal color vision were enlisted to participate in the experiment. The participants were the same participants who volunteered for the perceptual triangulate study in Experiment 11. The participants were paid $5 for each of the two experimental phases.

Apparatus and Stimulus Materials

The experimental control software from Experiments 11 and 12 was combined to provide training on the grayscale stimuli and testing on the happiness elicited by monetary sums. The stimulus materials were identical to the grayscale training stimuli in Experiment 11 and the monetary testing stimuli in Experiment 12.

Design and Procedure

The design and procedure were a combination of Experiments 11 and 12. The design of the experiment is outlined in Figure 67. As in Experiment 11, there were two experimental phases, separated by a week. During one phase, participants were trained to scale grayscale stimuli according to an exponent equal to 0.20. In the other phase, participants were trained to scale grayscale stimuli according to an exponent equal to 0.46. The order of the experimental phases was alternated between participants. In both experimental conditions, the participants were instructed to use the learned scale to estimate how happy the displayed amounts of money would make them. During the testing phase, the monetary stimuli without feedback were alternated with the grayscale stimuli with feedback.

Results and Discussion

The results were analyzed as in Experiments 11 and 12 and are summarized in Tables 23 – 24 and Figures 68 – 71. Exponent values were obtained by regressing the magnitude values against brightness and the happiness values to money for each participant. The coefficient of variation and the highest-to-lowest exponent ratio were calculated and compared to the equivalent measures obtained from the baseline study in Experiment 12. It was predicted that these measures of variance would actually increase for constrained scaling vs. magnitude estimation to reflect constrained scaling's greater sensitivity to true individual differences across participants.

Exponent Values

In the experimental condition in which the training exponent equaled 0.20, the average scaling exponent value for the grayscale stimuli was 0.175 with *SD*/*M* = 0.158 and *H*:*L* = 1.524:1. For the monetary stimuli, the average scaling exponent was 0.195 with *SD*/*M* = 0.505 and *H*:*L* = 3.284:1. The average ratio of the grayscale to the monetary exponents was 1.139. In the experimental condition in which the training exponent equaled 0.46, the average scaling exponent value for the grayscale stimuli was 0.397 with *SD*/*M* = 0.076 and *H*:*L* = 1.183:1. The average exponent for the monetary stimuli was 0.269 with *SD*/*M* = 0.570 and *H*:*L* = 4.589:1. The average ratio of grayscale

Table 23. Summary of participants' brightness and happiness scaling for 0.20 exponent training in Experiment 13.

	Grayscale Brightness + Feedback			**Happiness + No Feedback**			**Ratio of Grayscale:Happiness**		
P	**Exponent**	**Intercept**	***R*²**	**Exponent**	**Intercept**	***R*²**	**Exponent**	**Intercept**	***R*²**
1	0.188	1.321	0.904	0.289	0.216	0.917	0.651	6.116	0.986
2	0.188	1.325	0.788	0.265	0.640	0.785	0.709	2.070	1.004
3	0.182	1.343	0.759	0.246	0.982	0.718	0.740	1.368	1.057
4	0.192	1.307	0.798	0.088	1.118	0.908	2.182	1.169	0.879
5	0.126	1.397	0.494	0.089	1.391	0.575	1.416	1.004	0.859
M	0.175	1.339	0.749	0.195	0.869	0.781	1.139	2.345	0.957
SD	0.028	0.035	0.153	0.099	0.454	0.142	0.661	2.146	0.085

Table 24. Summary of participants' brightness and happiness scaling for 0.46 exponent training in Experiment 13.

	Grayscale Brightness + Feedback			**Happiness + No Feedback**			**Ratio of Grayscale:Happiness**		
P	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**
1	0.420	0.907	0.901	0.299	0.176	0.763	1.405	5.153	1.181
2	0.369	1.017	0.811	0.514	-1.189	0.802	0.718	-0.855	1.011
3	0.427	0.932	0.850	0.235	0.309	0.928	1.817	3.016	0.916
4	0.409	0.915	0.838	0.184	0.523	0.926	2.223	1.750	0.905
5	0.361	0.991	0.720	0.112	1.144	0.498	3.223	0.866	1.446
M	0.397	0.952	0.824	0.269	0.193	0.783	1.877	1.986	1.092
SD	0.030	0.049	0.067	0.153	0.857	0.176	0.936	2.262	0.227

GRAYSCALE STIMULI
TRAINING EXPONENT = 0.20

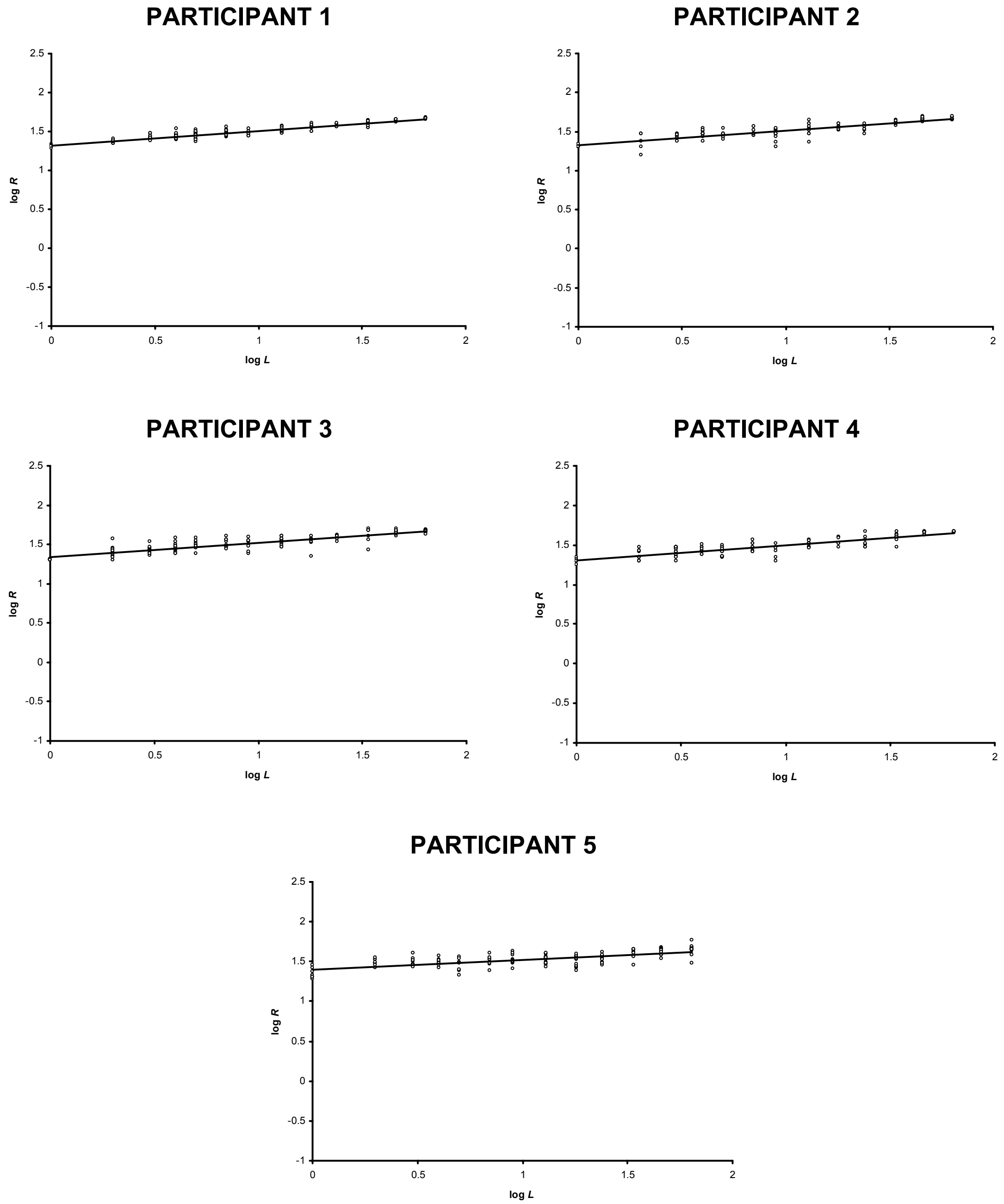


Figure 68. Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for grayscale squares with feedback for a training exponent value of 0.20 in Experiment 13.

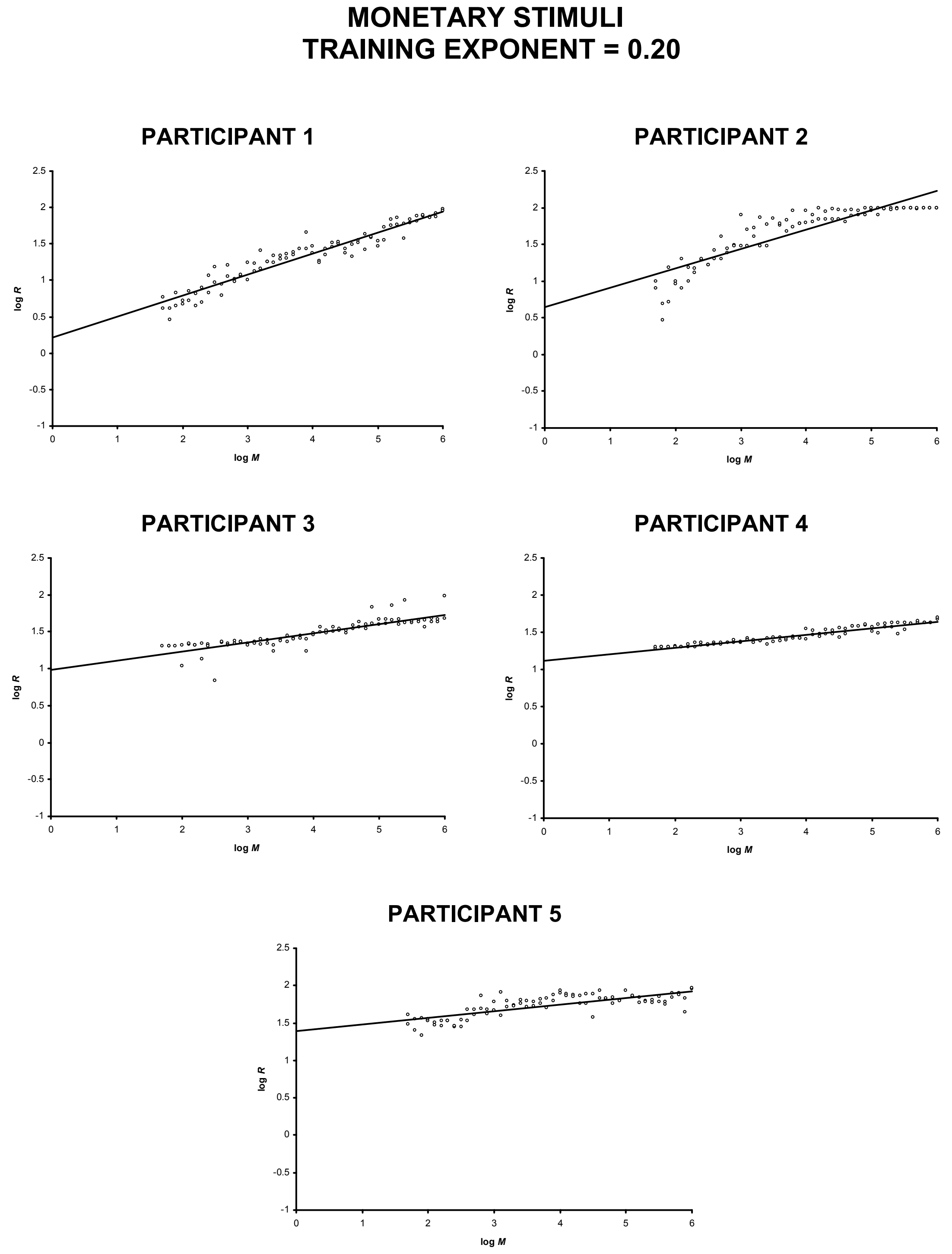


*Figure 69. Logarithmic scatterplot and regression line for happiness (*R*) in response to monetary sum (*M*) for a training exponent value of 0.20 in Experiment 13.*

GRAYSCALE STIMULI
TRAINING EXPONENT = 0.46

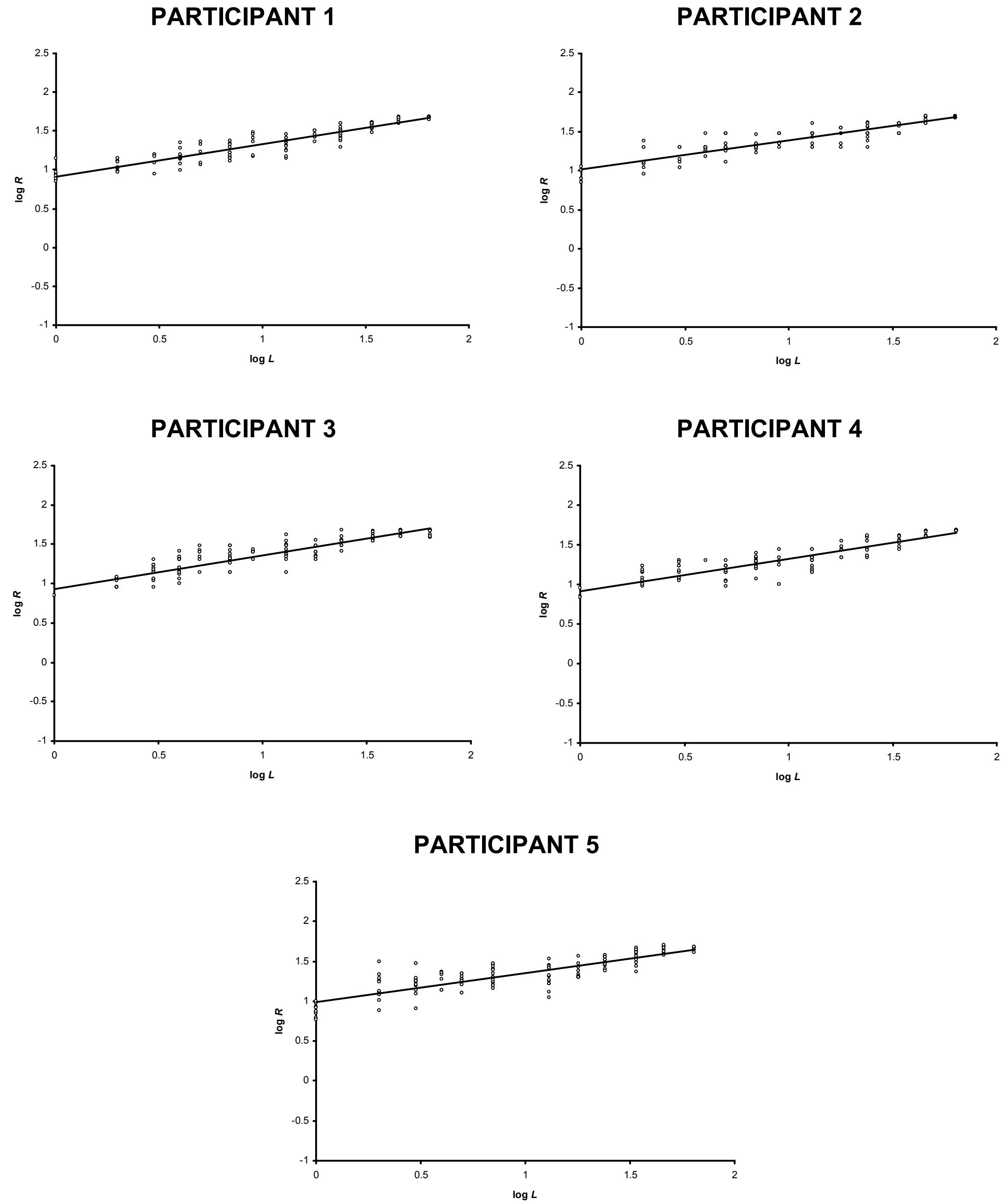
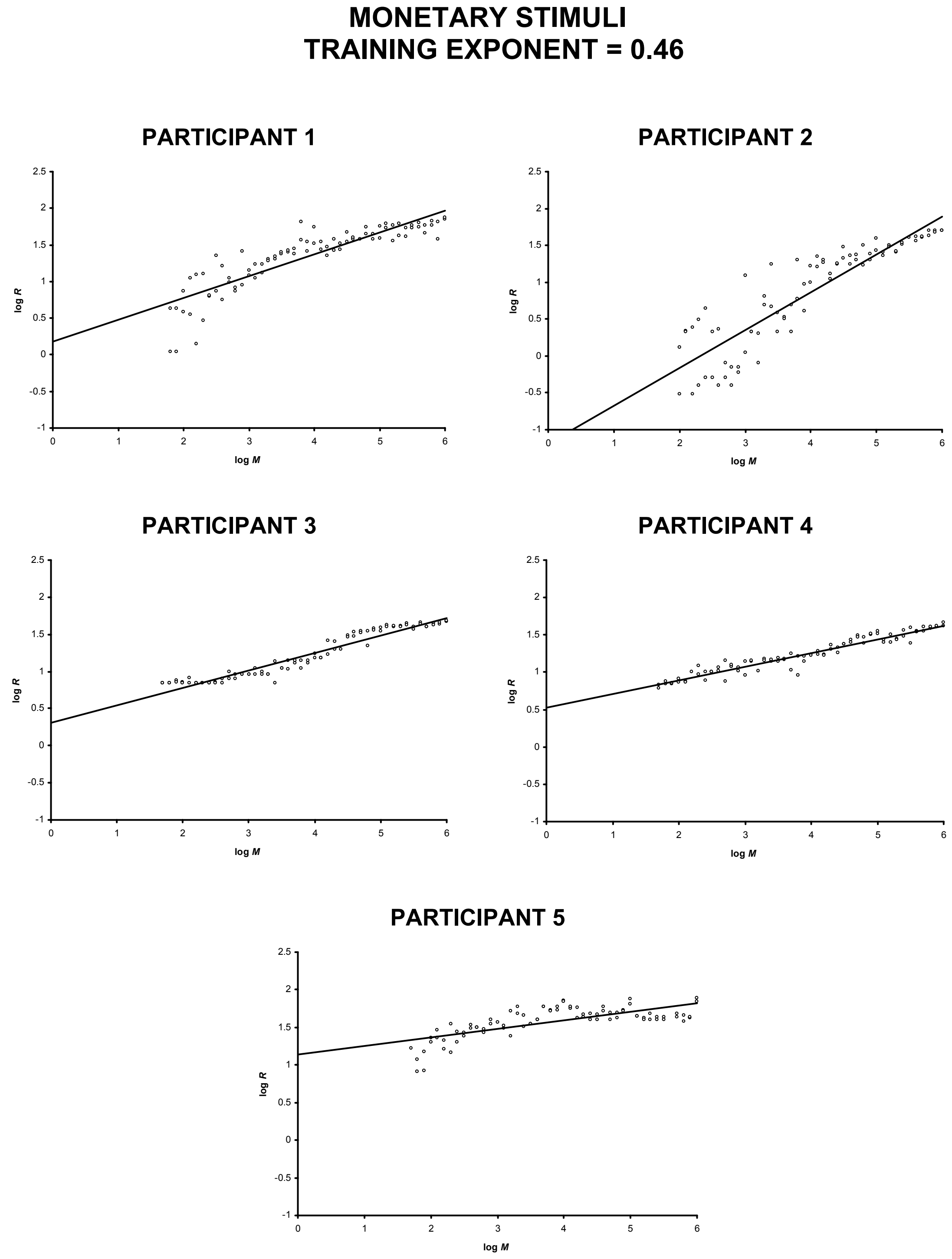


Figure 70. Logarithmic scatterplot and regression line for brightness in cd/m^2 (L) and participant response (R) for grayscale squares with feedback for a training exponent value of 0.46 in Experiment 13.



MONETARY STIMULI
TRAINING EXPONENT = 0.46

PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

*Figure 71. Logarithmic scatterplot and regression line for happiness (*R*) in response to monetary sum (*M*) for a training exponent value of 0.46 in Experiment 13.*

to the monetary exponents was 1.877. The average grayscale exponents and their variability levels were comparable to or better than the values obtained in Experiment 11. The variability for the monetary-to-happiness scale was greater than the variability found with magnitude estimation in Experiment 12.

Intercept Values

In the experimental condition in which the training exponent equaled 0.20, the average scaling intercept for the grayscale stimuli was 1.334 with *SD*/*M* = 0.026 and *H*:*L* = 1.069:1. For the monetary stimuli, the average scaling intercept was 0.869 with *SD*/*M* = 0.523 and *H*:*L* = 6.440:1. The average ratio of grayscale to monetary intercepts was 2.345. In the experimental condition in which the training exponent equaled 0.46, the average scaling intercept for grayscale stimuli was 0.952 with *SD*/*M* = 0.051 and *H*:*L* = 1.121:1. For the monetary stimuli, the average scaling intercept was 0.193 with *SD*/*M* = 4.449 and *H*:*L* = -0.962.[59] The average ratio of grayscale to monetary intercepts was 1.986. In both conditions, the average intercept values and variability levels for the grayscale stimuli were comparable to those obtained in Experiment 11. Likewise, the intercept variability in the present experiment was comparable to that obtained through magnitude estimation in Experiment 12.

Goodness of Fit Coefficients

For the experimental condition in which the training exponent equaled 0.20, the average scaling R^2 value for grayscale stimuli was 0.749 with *SD*/*M* = 0.204 and *H*:*L* =

[59] Recall that a negative highest-to-lowest ratio is indicative of a large ratio.

1.830:1. For the monetary stimuli, the average scaling R^2 value was 0.781 with *SD*/*M* = 0.142 and *H*:*L* = 1.595:1. The average ratio of grayscale to monetary goodness of fit coefficients was 0.957. In the experimental condition in which the training exponent equaled 0.46, the average R^2 value for the grayscale stimuli was 0.824 with *SD*/*M* = 0.081 and *H*:*L* = 1.251:1. For the monetary stimuli, the average R^2 value was 0.783 with *SD*/*M* = 0.224 and *H*:*L* = 1.863:1. The average ratio of grayscale to monetary goodness of fit was 1.092. Across both conditions, the grayscale goodness of fit was actually slightly better than in Experiment 11. The monetary-happiness goodness of fit was nearly identical to the fit found through magnitude estimation in Experiment 12.

General Discussion

As predicted, the variability of the happiness scaling actually increased in the constrained scaling condition. Since participants exhibited comparable mastery of the grayscale training scale as in previous experiments, it is not assumed that the increased variability is a byproduct of a failure to master the perceptual-subjective cross-modal scaling. Instead, it is assumed that the increased variability was a product of the constrained scalers' improved scaling fidelity, making the scale more sensitive to true individual differences.

Referring back to Equation 30 in Experiment 11, it was found that the training to different scale exponents (*A*) carried over to scaling novel stimuli (*B*), such that $\frac{A_1}{B_1} = \frac{A_2}{B_2}$. The question remains whether this constant relationship also holds when *A* and *B* represent different modalities, namely perceptual and subjective, respectively.

Substituting the appropriate exponent values from the present experiment into Equation 30, the following equation is produced:

$$\frac{0.397}{0.175} = \frac{0.269}{0.195}, \tag{38}$$

which simplifies as:

$$2.269 \neq 1.379. \tag{39}$$

The scaling relationship is not constant, suggesting that constrained scaling using a perceptual training scale did not consistently map to the subjective domain.[60] In terms of the intercept values, substituting the appropriate values into Equation 30 yields the following equation:

$$\frac{0.952}{1.339} = \frac{0.193}{0.869}, \tag{40}$$

which simplifies as:

$$0.711 \neq 0.222. \tag{41}$$

Again, the scaling relationship is not constant. The trained perceptual scale did not map to the subjective scale in a consistent manner for exponents or intercepts across the two conditions.

Two sets of findings must be reconciled. First, constrained scaling resulted in greater variability between participants than did magnitude estimation for the subjective domain of scaling happiness. Second, individual participants' mapping of the perceptual

[60] I have assumed that subjective happiness is relatively constant over time. If subjective happiness changes considerably over time, a two-phase longitudinal experiment like the present experiment is not a well suited method for testing how well different perceptual scales are applied to subjective domains.

scale to the subjective domain did not remain constant over a one-week interval. The utility of constrained scaling for cross-modal perceptual-subjective scaling must be punctuated with a question mark. On the one hand, preliminary findings suggest that this methodological extension may hold the key to increasing sensitivity to individual differences. On the other hand, individual differences appear in places where they would not necessarily be expected, in the mapping of the learned scale to the testing stimuli.

This experiment fails to resolve this quandary, leaving open the ultimate verdict on the use of constrained scaling for subjective domains. However, some resolution may be found by looking at the exponent and intercept values of individual participants. The relationship between training and testing exponents does, in fact, hold constant for two of the five participants. If constrained scaling worked in both cases, then the ratio of the exponents for the two learned scales (i.e., brightness) divided by the ratio of the exponents for the two unlearned scales (i.e., happiness) should equal 1. For participants 2 and 4, these ratios were equal to 0.987 and 0.982, respectively. These ratios suggest that the application of the two learned scales held largely constant across the two testing phases, although the intercept ratios failed to hold constant. Participants 1, 3, and 5 also showed remarkable consistency, but in a different way. For them, the ratio of the exponents for the unlearned happiness scales was approximately equal to 1 (1.035, 0.955, and 1.258, respectively). This indicates that they simply ignored the training in the second round of testing and used the same scale they had learned in the first round of testing. , in contrast, featured considerably larger variability. Participants 2 and 4 shared a common order for testing phases, having first been trained to scale with an exponent equal to 0.46 and then an exponent of 0.20 in the second phase. Participants 1, 3, and 5 received the opposite order. It remains unclear why there would be a potential order effect for participants involved in subjective scaling but not in the psychophysical scaling in Experiment 11. To help resolve unanswered questions regarding cross-modal scaling in a subjective domain, the next chapter further explores constrained psychometric scaling.

EXPERIMENT 14[61]

Introduction

The present experiment builds on Experiments 12 and 13 by further determining the utility of a constrained scale as a subjective measure. Experiments 12 and 13 sought to establish the validity of using constrained scaling as a method to determine true individual differences. In the present experiment, I investigate constrained scaling vs. magnitude estimation for the domain of subjective affect in computing.

In the field of human-computer interaction (HCI), computer software is evaluated in terms of its usability (Dillon, 1983; Dumas & Redish, 1999; Lindgaard, 1994; Nielsen, 1993; Rubin, 1994). Usability, in turn, is defined according to several subcomponents. Nielsen, for example, proposed that usability should be defined according to interface learnability, efficiency of use, interface memorability, user errors, and user satisfaction. More recently, the International Standards Organization (1998) has proposed three definitional components of usability, which are effectiveness, efficiency, and user satisfaction. Effectiveness of software is typically measured in terms of the user's success rate at completing tasks, whereas the efficiency of software is generally measured by the user's time to complete those tasks. The third factor, user satisfaction, represents an affective component in the use of software. User satisfaction is commonly measured with Likert-style satisfaction scales (Dumas, 2001) such as the Software Usability

[61] Portions of Experiment 14 were first presented as papers at the Annual Meeting of the Human Factors and Ergonomics Society (Boring, 2003) and the Association for Computing Machinery's CHI Conference on Human Factors in Computing Systems (Boring & Fernandes, 2004).

Measurement Inventory (Kirakowski, 1996; Kirakowski & Corbett, 1993) or the System Usability Scale (Brooke, 1996).

Frøkær, Hertzum, and Hornbæk (2000) found the three components of usability specified by the International Standards Organization (1998) to be orthogonal to one another, suggesting that usability should be assessed according to all three dimensions simultaneously. Nonetheless, there is still debate about the best way to assess user satisfaction (Dudek & Lindgaard, 2004; Lindgaard & Dudek, 2003). In particular, while user satisfaction continues to be the primary affective metric in usability studies, new research has suggested that aesthetic factors may be an important component of usable software (Karvonen, 2000; Norman, 2004). There is a significant relationship between perceived usability and aesthetic appeal (Kurosu & Kashimura, 1995; Tractinsky, 1997; Traktinsky, Katz, & Ikar, 2000), suggesting that user satisfaction may be related as much to aesthetic factors as to ease of use. Due to the current interest in the role of aesthetics in computer software as well as the lack of well-established aesthetics scales for computer software, the present experiment investigated the most effective way to scale visual aesthetic appeal on computers.

As part of a larger series of studies, Fernandes (2003) conducted a study to evaluate the subjective aesthetic visual appeal of 100 Web pages presented in random order. Using an unmarked line scale with the equivalent of 100 scale units, 22 participants rated screen shots of each Web page. In order to gather initial subjective impressions by the participants, the Web pages were displayed for 500 ms, a time appropriate to gauge the mere exposure effect, the aesthetic first impression of the Web pages (Veryzer, 1999). The participants completed two evaluations of each Web page to ensure intraparticipant reliability. Fernandes standardized the scores and rank ordered the Web pages, subsequently using the 25 lowest rated and 25 highest rated Web pages for a follow-up study.

The goal of the present experiment is to utilize constrained scaling in a domain where there were documented individual differences. Fernandes' (2003) data set provides a validated data set in which there are Web pages with a reasonable amount of disagreement in terms of visual appeal among participants. The raw data set was obtained from Fernandes and reanalyzed for variability. The log of each participant's subjective visual appeal rating for each Web page was regressed against the log of the average subjective visual appeal rating for the same Web page:

$$\log R = m \log W + \log a \,, \tag{42}$$

where *W* represents the average visual appeal rating for a particular Web page and *R* represents the specific participant's rating of the same Web page. Transforming Equation 42 from a logarithmic scale to a standard scale produces the familiar Power Law form:

$$R = aW^{m} \,. \tag{43}$$

The exponent value, *m*, represents the degree to which the participants' individual ratings match the average ratings. Averaging *m* across all 22 participants in Fernandes revealed that the average match between the individual ratings and the collective ratings had a slope equal to 0.382, with a standard deviation of 0.401. The coefficient of variation was 0.953, and the highest-to-lowest exponent ratio was 4.277:1.

Note that the interparticipant variability was considerably higher in Fernandes (2003) than in the magnitude estimation, cross-modality matching, and constrained scaling experiments reviewed in West et al. (2000). The average coefficient of variation for the ten reviewed magnitude estimation studies in West et al. was 0.327, and the average highest-to-lowest ratio was 2.991:1. Similarly, the average coefficient of variation for the four reviewed cross-modality matching studies was 0.348, while the highest-to-lowest ratio was 3.003:1. The levels for the constrained scaling experiments were significantly lower. For the subjective visual appeal data, the coefficient of variation was approximately three times greater than the coefficients of variation obtained from conventional magnitude estimation or cross-modality matching studies. The highest-to-lowest ratio is inconclusive. As West et al. noted, the highest-to-lowest ratio is susceptible to outliers, meaning that the large number of participants in Fernandes' study would predict a large highest-to-lowest ratio.

The present experiment compared the results from Fernandes (2003) against a new set of visual appeal ratings obtained through constrained scaling. It was possible that the participants in the study by Fernandes perceived the same Web pages with equal levels of affect but differed in the way they used the scale across the range of possible scores. This conclusion would suggest that the high variability found in aesthetics was an artifact of scaling more than the product of differences in aesthetic preference. If this were the case, it would be expected that constrained scaling would result in lower variability than was found in the study by Fernandes. By calibrating individuals to a common scale, constrained scaling should minimize differences in scale usage that increase interparticipant variability. Alternately, it was possible that the participants in the study by Fernandes actually perceived the same Web pages with different levels of affect. If this were the case, it would be expected that constrained scaling would result in higher variability. By calibrating individual participants to a common scale, constrained scaling would prove more sensitive to true individual differences. If subjective aesthetic visual appeal for Web pages differs across individuals, a constrained scale should prove more sensitive than a magnitude estimation scale to those differences.

Method

As in Experiment 13, the present experiment uses training on grayscale brightness to achieve scale mastery. Whereas in Experiment 13 participants used the learned scale to rate the degree of happiness that would be obtained by different degrees of money, in the present experiment participants were then instructed to use the learned scale to rate the subjective visual appeal of 100 Web pages.

Participants

Because the results in the present experiment would be compared to the results from a study with a larger number of participants, the decision was made to incorporate more than the customary five participants per study. Eight volunteers with self-reported normal color vision were enlisted to participate in the experiment. Volunteers who had previously participated in constrained scaling brightness experiments were eligible to participate, and the final participants were drawn from a pool of existing and new volunteers. The volunteers were paid $10 for their participation.

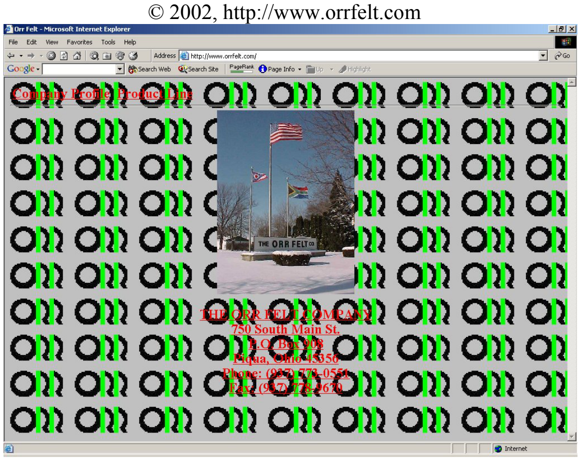
Apparatus and Stimulus Materials

The experimental control software used in the previous experiments was modified to allow the display of screenshots from Web pages. The training stimuli consisted of grayscale squares at the 14 luminous intensity values described in Table C-1 in Appendix C. The training exponent was set at 0.33. The testing stimuli consisted of the screenshots of 100 Web pages of varying levels of visual appeal that were used in Fernandes' earlier study (2003). As in Fernandes, the screenshots from the Web pages were displayed on the screen for 500 ms to create a mere exposure effect suitable for gauging the participants' initial impression of the Web pages. Precise timing was achieved by incorporating a Visual Basic algorithm developed by Bedell (2000).

Sample screenshots of the Web pages in the present experiment are presented in Table 25. The Web pages that were presented were the intellectual property and, in many cases, the copyright of their respective owners. The Web pages were presented for educational research purposes only and were in compliance with U.S. "Fair Use" and Canadian "Fair Dealing" clauses of copyright statutes [Title 17 of *United States Code* §107 (2000) and *Canada Copyright Act*, R.S., c. C-42, s. 29 (1985)]. To minimize unnecessary reproduction of copyrighted materials, only the sample screenshots in Table 25 are included in this document. Note that the original Web pages were presented in color. Web page color selection was a likely contributor to subjective visual appeal. Also note that no effort has been made in this document to clarify the reason for particular ratings, nor should ratings be construed as a critique or endorsement of particular Web pages in terms of their ultimate utility or usability. Since the time the

Table 25. Sample Web pages of varying levels of aesthetic visual appeal used in Fernandes (2003) and in Experiment 14.

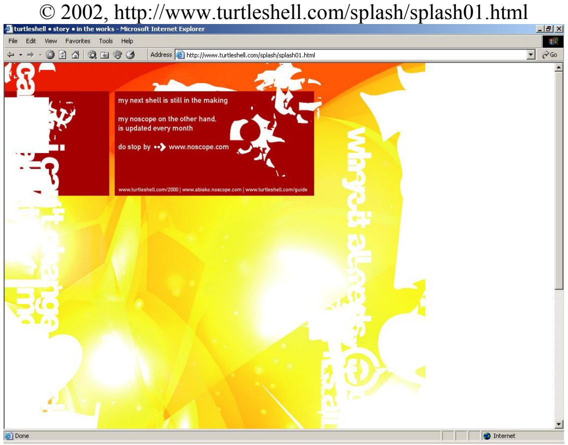
© 2002, http://www.orrfelt.com


Orr Felt - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.orrfelt.com/
Go
Google
Search Web
Search Site
PageRank
Page Info
Up
Highlight
THE ORR FELT CO
Internet

Company Profile Product Line

THE ORR FELT COMPANY
750 South Main St.
P.O. Box 908
Piqua, Ohio 45356
Phone: (937) 773-0551
Fax: (937) 778-9670

© 2002, http://www.turtleshell.com/splash/splash01.html


turtleshell • story • in the works - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.turtleshell.com/splash/splash01.html
Go
Done
Internet

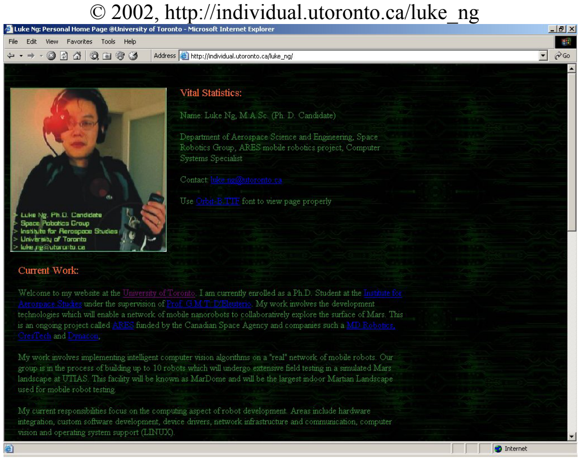
my next shell is still in the making

my noscope on the other hand,
is updated every month

do stop by ••> www.noscope.com

www.turtleshell.com/2000 | www.abisko.noscope.com | www.turtleshell.com/guide

© 2002, http://individual.utoronto.ca/luke_ng


Luke Ng: Personal Home Page @University of Toronto - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://individual.utoronto.ca/luke_ng/
Go
> Luke Ng, Ph.D. Candidate
> Space Robotics Group
> Institute for Aerospace Studies
> University of Toronto
> luke.ng@utoronto.ca
Internet

Vital Statistics:

Name: Luke Ng, M.A.Sc. (Ph. D. Candidate)

Department of Aerospace Science and Engineering, Space Robotics Group, ARES mobile robotics project, Computer Systems Specialist

Contact: luke.ng@utoronto.ca

Use Orbit-B TTF font to view page properly

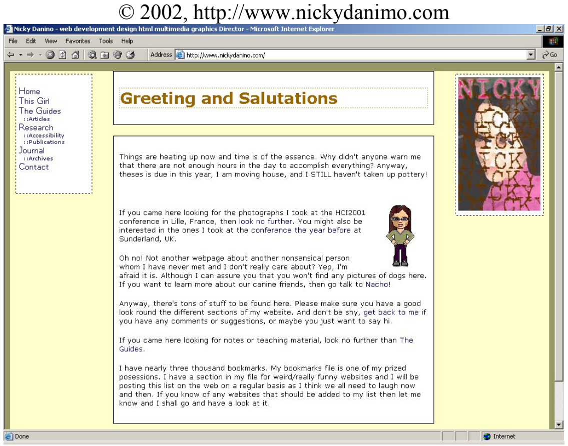
Current Work:

Welcome to my website at the University of Toronto. I am currently enrolled as a Ph.D. Student at the Institute for Aerospace Studies under the supervision of Prof. G.M.T. D'Eleuterio. My work involves the development technologies which will enable a network of mobile nanorobots to collaboratively explore the surface of Mars. This is an ongoing project called ARES funded by the Canadian Space Agency and companies such a MD Robotics, CresTech and Dynacon.

My work involves implementing intelligent computer vision algorithms on a "real" network of mobile robots. Our group is in the process of building up to 10 robots which will undergo extensive field testing in a simulated Mars landscape at UTIAS. This facility will be known as MarDome and will be the largest indoor Martian Landscape used for mobile robot testing.

My current responsibilities focus on the computing aspect of robot development. Areas include hardware integration, custom software development, device drivers, network infrastructure and communication, computer vision and operating system support (LINUX).

© 2002, http://www.nickydanimo.com


Nicky Danino - web development design html multimedia graphics Director - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.nickydanino.com/
Go
Home
This Girl
The Guides
::Articles
Research
::Accessibility
::Publications
Journal
::Archives
Contact
NICKY
Done
Internet

Greeting and Salutations

Things are heating up now and time is of the essence. Why didn't anyone warn me that there are not enough hours in the day to accomplish everything? Anyway, theses is due in this year, I am moving house, and I STILL haven't taken up pottery!

If you came here looking for the photographs I took at the HCI2001 conference in Lille, France, then look no further. You might also be interested in the ones I took at the conference the year before at Sunderland, UK.

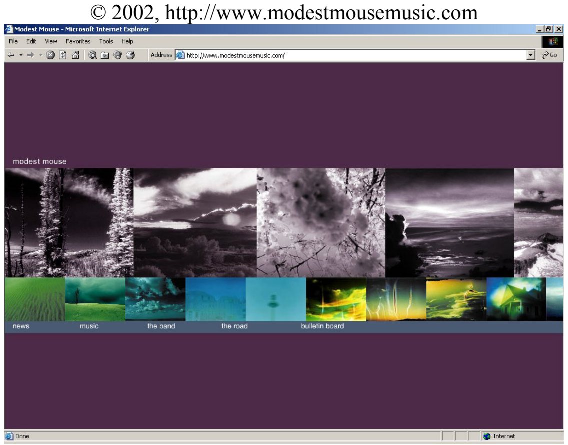
Oh no! Not another webpage about another nonsensical person whom I have never met and I don't really care about? Yep, I'm afraid it is. Although I can assure you that you won't find any pictures of dogs here. If you want to learn more about our canine friends, then go talk to Nacho!

Anyway, there's tons of stuff to be found here. Please make sure you have a good look round the different sections of my website. And don't be shy, get back to me if you have any comments or suggestions, or maybe you just want to say hi.

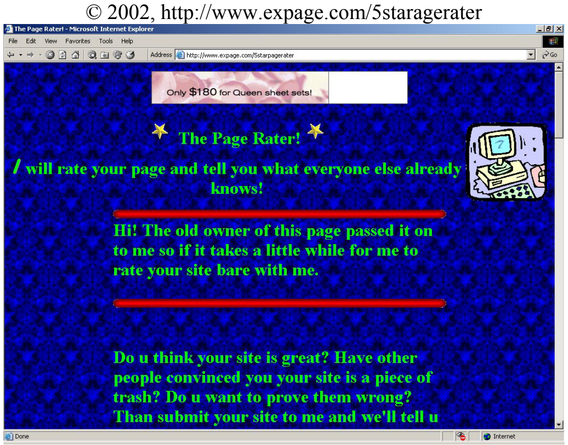
If you came here looking for notes or teaching material, look no further than The Guides.

I have nearly three thousand bookmarks. My bookmarks file is one of my prized posessions. I have a section in my file for weird/really funny websites and I will be posting this list on the web on a regular basis as I think we all need to laugh now and then. If you know of any websites that should be added to my list then let me know and I shall go and have a look at it.

© 2002, http://www.modestmousemusic.com


Modest Mouse - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.modestmousemusic.com/
Go
modest mouse
news
music
the band
the road
bulletin board
Done
Internet

© 2002, http://www.expage.com/5staragerater


The Page Rater! - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.expage.com/5starpagerater
Go
Only $180 for Queen sheet sets!
Done
Internet

The Page Rater!

I will rate your page and tell you what everyone else already knows!

Hi! The old owner of this page passed it on to me so if it takes a little while for me to rate your site bare with me.

Do u think your site is great? Have other people convinced you your site is a piece of trash? Do u want to prove them wrong? Than submit your site to me and we'll tell u

Web pages were captured as screenshots during summer 2002, many Web pages presented have undergone considerable revision or have ceased to be available online.

Design and Procedure

The experiment was an extension of the basic design and procedure used in Experiment 13. Participants were trained on grayscale brightness stimuli over 50 initial iterations.[62] The participants then used this learned scale to evaluate the aesthetic visual appeal of the 100 Web pages, interspersed with training trials to reinforce scale learning. The experimental design is summarized in Figure 72.

Note that the participants in Fernandes (2003) used a line scale to estimate the magnitude of visual appeal. The use of a line scale is a variation of absolute magnitude estimation, first suggested by Stevens and Galanter (1957). The length of the line serves as a gauge of the magnitude that is being assessed. Zwislocki (1983) compared values obtained using both line scales and conventional magnitude estimation. He found slight differences between the line scale and magnitude estimation values, which he attributed to a failure on the part of the magnitude estimation participants to translate their magnitude perceptions to a numerical scale correctly. The differences in values between the two scales did not, however, occur systematically, and those differences were mostly eliminated when averaging the results across multiple trials and multiple participants. The advantages of using a line scale over a numerical scale are minor and do not represent a significant deviation over conventional magnitude estimation techniques. It

[62] Experiment 14 incorporated nine more training trials than in previous brightness constrained scaling experiments. The training trials were randomly sampled.

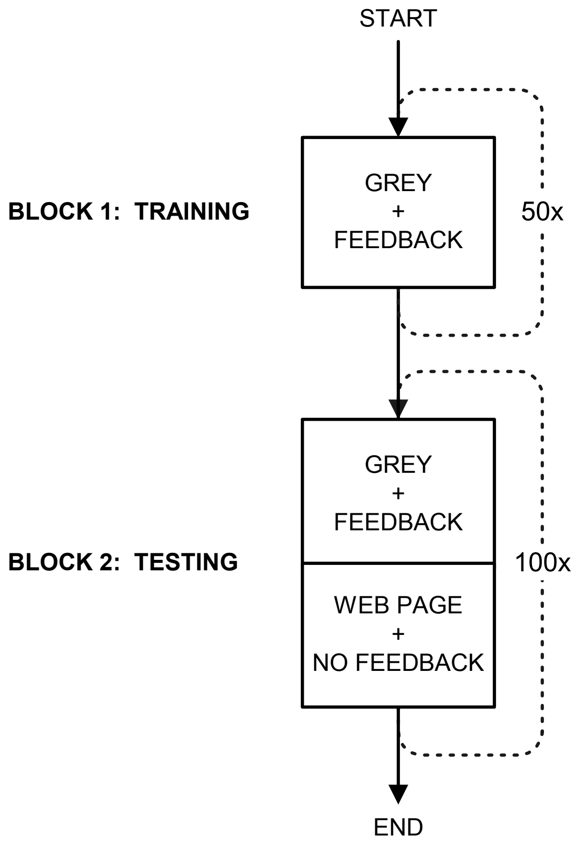


Figure 72. Schematic flow of design in Experiment 14.

should, however, be noted that the method presented in the current experiment compares constrained scaling to line length scaling—which may have produced slightly different results than conventional magnitude estimation. No attempt was made in the present experiment to adapt the constrained scaling slider to the line length scaling in Fernandes.

Results and Discussion

The results for the grayscale training stimuli were analyzed as in previous experiments. The scores for Web pages were logarithmically regressed against the average individual Web page ratings from Fernandes (2003). This use of average ratings afforded a basis for comparison to the results from Fernandes' magnitude estimation à la line-length scaling experiment. The use of the average ratings as the *x*-axis for scatterplots allows the unidimensional data from the visual appeal ratings to be treated similarly to the two-dimensional stimulus-response data from the perceptual constrained scaling experiments earlier in this document. The transformation of unidimensional data to the Cartesian coordinate system is controversial;[63] therefore, more established statistical comparisons appropriate for unidimensional data sets conclude the analyses. The results are summarized in Table 26 and Figures 73 – 74.

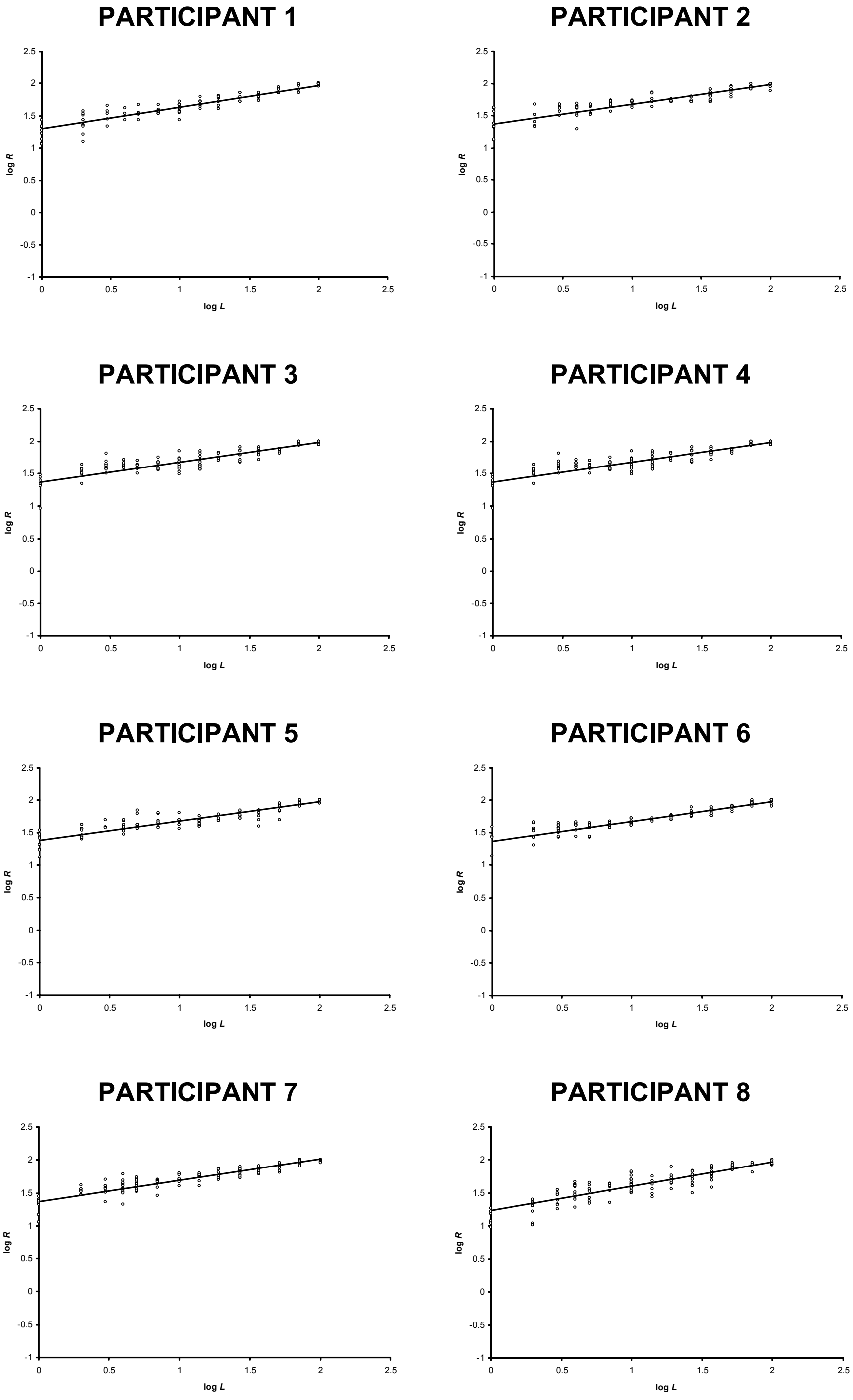
Results for Grayscale Training Stimuli

As in previous experiments, participants were trained to scale brightness according to stimulus values raised to an exponent of 0.333. Participants exhibited an

[63] At issue is the fact that average data are being used to construct an underlying "objective" scaling continuum for matching to individual results. With rare exception, this underlying scale has very little to do with individual scaling results, making it less an objective scale than an averaged scale for the convenience of analytic comparisons.

Table 26. Summary of participants' brightness and visual appeal scaling in Experiment 14.

	Grayscale Brightness + Feedback			**Web Visual Appeal + No Feedback**			**Ratio of Grayscale:Visual Appeal**		
P	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**	**Exponent**	**Intercept**	***R*^{2}**
1	0.335	1.303	0.916	0.992	0.069	0.488	0.338	18.884	1.877
2	0.307	1.370	0.874	1.627	-1.192	0.457	0.189	-1.149	1.912
3	0.307	1.373	0.869	0.719	0.501	0.218	0.427	2.741	3.986
4	0.326	1.348	0.857	3.619	-4.770	0.662	0.090	-0.283	1.295
5	0.297	1.379	0.890	1.600	-1.118	0.733	0.186	-1.233	1.214
6	0.303	1.373	0.909	1.042	0.009	0.630	0.291	152.556	1.443
7	0.324	1.365	0.909	1.702	-1.261	0.494	0.190	-1.082	1.840
8	0.367	1.233	0.814	0.269	1.290	0.023	1.364	0.956	35.391
M	0.321	1.343	0.880	1.446	-0.809	0.463	0.384	21.424	6.120
SD	0.023	0.051	0.034	1.009	1.840	0.238	0.410	53.413	11.860



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

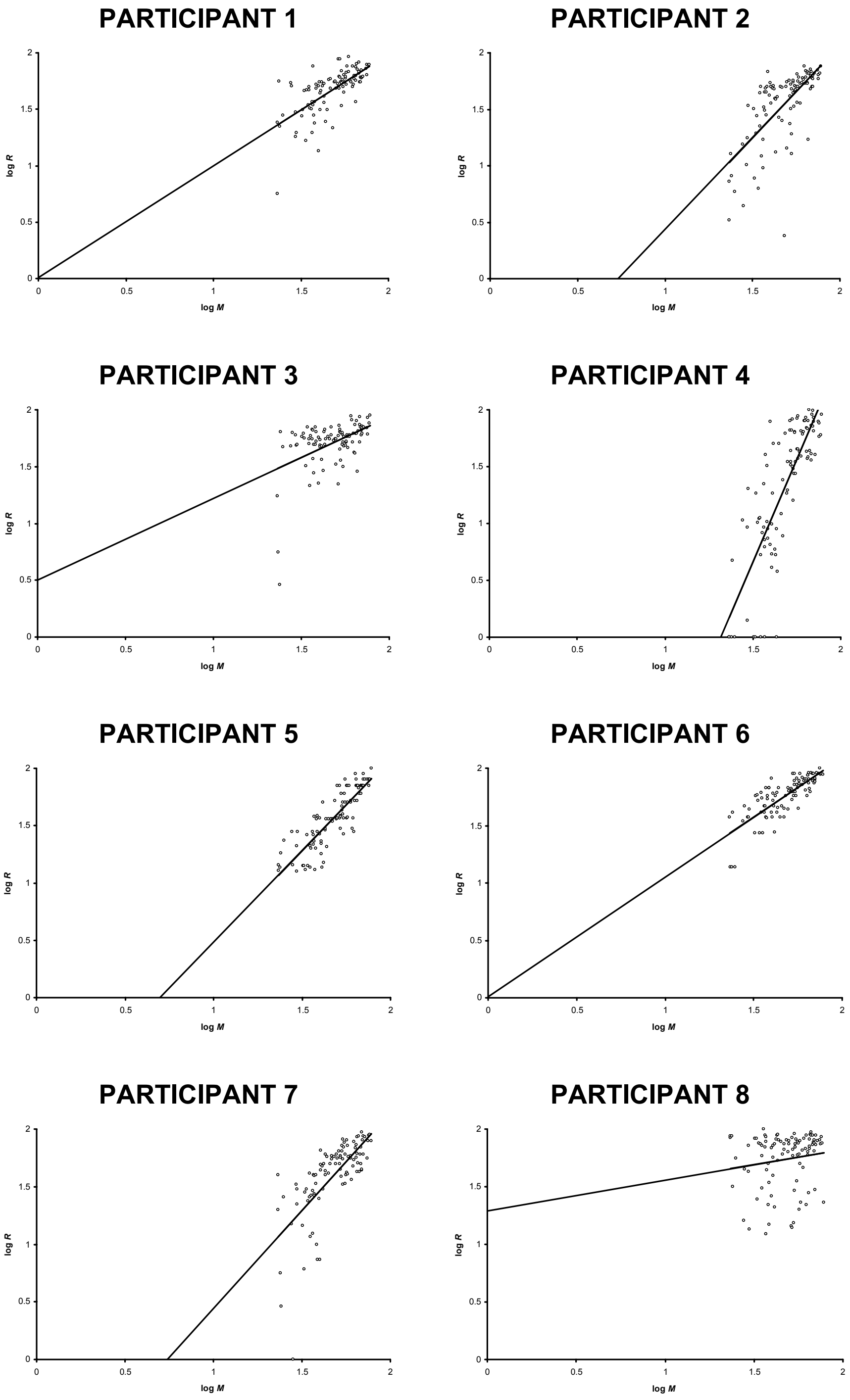
PARTICIPANT 5

PARTICIPANT 6

PARTICIPANT 7

PARTICIPANT 8

Figure 73. Logarithmic scatterplot and regression line for brightness in cd/m^2 *(*L*) and participant response (*R*) for grayscale squares with feedback in Experiment 14.*



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

PARTICIPANT 6

PARTICIPANT 7

PARTICIPANT 8

*Figure 74. Logarithmic scatterplot and regression line for participant aesthetic visual appeal of Web pages (*R*) against average visual appeal of Web pages (*W*) in Experiment 14.*

average learned exponent of 0.321 for the grayscale brightness squares, with *SD*/*M* = 0.071 and *H*:*L* = 1.236:1. The average learned intercept was 1.343 with *SD*/*M* = 0.038 and *H*:*L* = 1.118:1. The average R^2 value was 0.880 with *SD*/*M* = 0.039 and *H*:*L* = 1.125:1. The results were highly consistent with earlier constrained scaling findings, suggesting that participants were able to learn the training scale.

Results for Web Page Visual Appeal Ratings

Logarithmically regressing the individual visual appeal ratings against the collective ratings from Fernandes (2003) for each Web site revealed an exponent equal to 1.446 with *SD*/*M* = 0.698 and *H*:*L* = 13.453:1. The same participants who first learned the brightness scale exhibited a coefficient of variation that was over twice the average rate for magnitude estimation and cross-modality matching studies described in West et al. (2000) and a highest-to-lowest exponent ratio over ten times the rate for the same studies. Moreover, compared to the classical scaling employed in the Fernandes experiment (2003), the coefficient of variation was 1.6 times higher and the highest-to-lowest ratio was twice as high. The average exponent value across Fernandes' 22 participants was 1.518 with *SD*/*M* = 0.424 and *H*:*L* = 6.519:1. In the present experiment, the average intercept value was -0.809 with *SD*/*M* = -2.274 and *H*:*L* = -0.270:1.[64] These values were comparable to those in Fernandes, where the mean intercept value was -0.953 with *SD*/*M* = -1.282 and *H*:*L* = -0.268:1. For the R^2 values, the average value in the present experiment was 0.463 with *SD*/*M* = 0.513 and *H*:*L* = 31.870:1, whereas the

[64] Recall that negative values related to the intercept are indicative of higher values than positive numbers.

average R^2 value in Fernandes was 0.325 with $SD/M = 0.522$ and $H:L = 9.500:1$. While the average goodness of fit was slightly higher in the present experiment than in Fernandes' study, the variability in terms of the highest-to-lowest ratio was three and a half times greater. The ratio of grayscale to Web page visual appeal ratings is included in Table 26 for reference purposes but is not discussed here.[65]

Comparison between Constrained Scaling and Magnitude Estimation

Further analysis revealed that, on average, constrained scalers assigned significantly higher visual appeal values than did the scalers in Fernandes' (2003) experiment, $t(195) = 2.770$, $p < 0.005$ (See Figure 75). Constrained scalers also exhibited a significantly better fit to a linear regression line than did the classical scalers in Fernandes' experiment, $R^2 = 0.463$ vs. $R^2 = 0.325$, respectively, $t(28) = 1.700$, $p = 0.05$. On average, constrained scalers used a narrower range of scale values to rate the visual appeal of Web pages. The average ratio of highest-to-lowest ratings across participants for each Web page was 36.986:1 for constrained scalers and 74.703:1 for the classical scalers in Fernandes' experiment.[66] Constrained scalers tended to correlate higher on average with each other, $r = 0.391$, than did classical scalers, $r = 0.306$. This finding suggests that constrained scalers generally agreed on the direction of their ratings,

[65] Given the high variability of scaling the visual appeal of Web pages, it is not clear how informative the ratio between the grayscale training and the Web page testing is.

[66] This highest-to-lowest ratio compares is the average of the highest and lowest ratings given for each Web page across all participants in Fernandes (2003) and Experiment 14. This value must not be confused with the highest-to-lowest ratios used for the exponent, intercept, and R^2 values of the regression lines fitting the ratings across all Web pages.

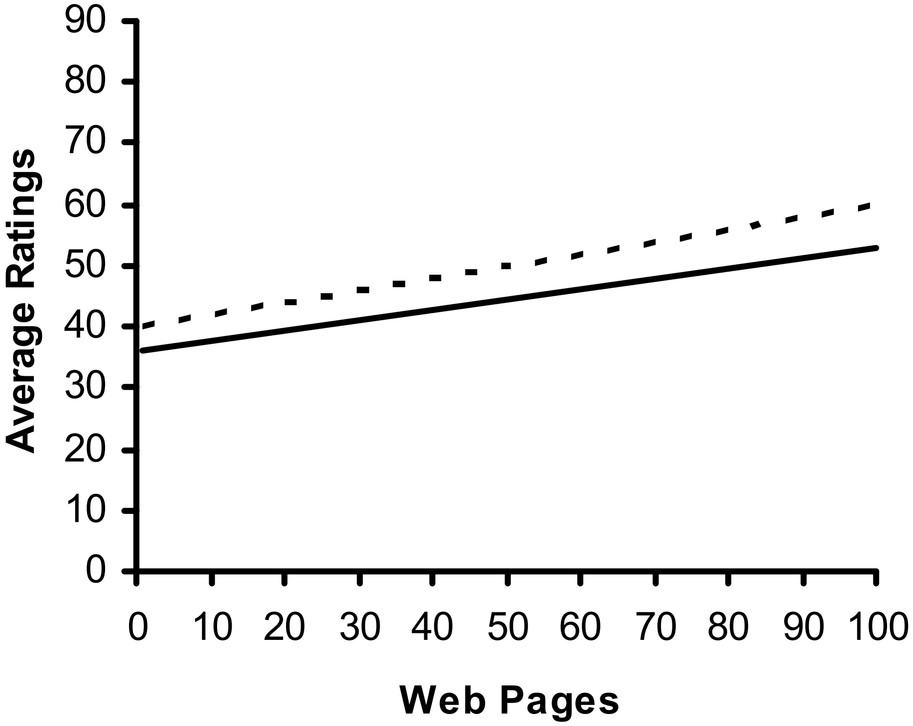


Figure 75. Average ratings for visual appeal of Web pages for classical scalers (solid line) and constrained scalers (dotted line).

although they may have used different actual scale values to represent the magnitude of their affect.

General Discussion

Did constrained scaling affect the scaling responses of participants for the visual appeal of Web pages? When trained to scale brightness, constrained scalers clearly exhibited a different scaling response pattern than did classical scalers. The results, however, did not follow the typical pattern for constrained scaling experiments, in which a significant reduction in scaling variability would be expected. The opposite effect was demonstrated. Constrained scalers actually exhibited much higher variability than did classical scalers.

Given that the constrained scalers demonstrated low variability for the grayscale training trials, it is assumed that the constrained scalers did, in fact, learn to match their mental magnitudes to the numerical scale. Further, since the constrained scalers demonstrated higher R^2 values for their Web page ratings than did classical scalers, it is assumed that the increased scaling variability in the constrained scaling condition was not due to poor application of the learned scale to the rating of Web pages. Instead, it can be concluded that the increased variability in the constrained scaling condition is a reflection of true individual differences in the participants' affective responses to different Web pages. By calibrating participants to map their mental magnitude to a numeric scale in a consistent manner, constrained scaling increases scaling sensitivity to true differences in subjective response.

EXPERIMENT 15[67]

To demonstrate the effectiveness of constrained scaling for another real-world application, I conducted an experiment in which participants judged the fluidity of motion for sample videos across different frame rates. Constrained scaling was used as a tool to facilitate selecting quality parameters for streaming video.

Streaming media is audio or video that is broadcast from a computer server over the Internet to a client media decoder, which typically consists of media player software on a personal computer. Presently, a user who wishes to watch streaming video or listen to streaming audio over the Internet faces few choices regarding the streaming quality. The speed of the user's Internet connection is the primary deciding factor for the streaming quality. A user with a slow Internet connection must content him or herself with relatively low quality streaming media, whereas a user with a fast Internet connection has the luxury of high quality video and audio.

Internet Protocol 6 (IPv6) introduces *quality of service* into the user's Internet experience (Hinden, 1996), whereby a required minimum level of bandwidth may be specified by the user. A quality of service contract reflects not only the speed with which the user connects to the Internet but also the data throughput across the Internet (West, Boring, Dillon, & Bos, 2001). In the context of streaming media, a quality of service contract guarantees that at no time will the speed of the connection between the streaming broadcaster and the user go below the level that the user has requested.

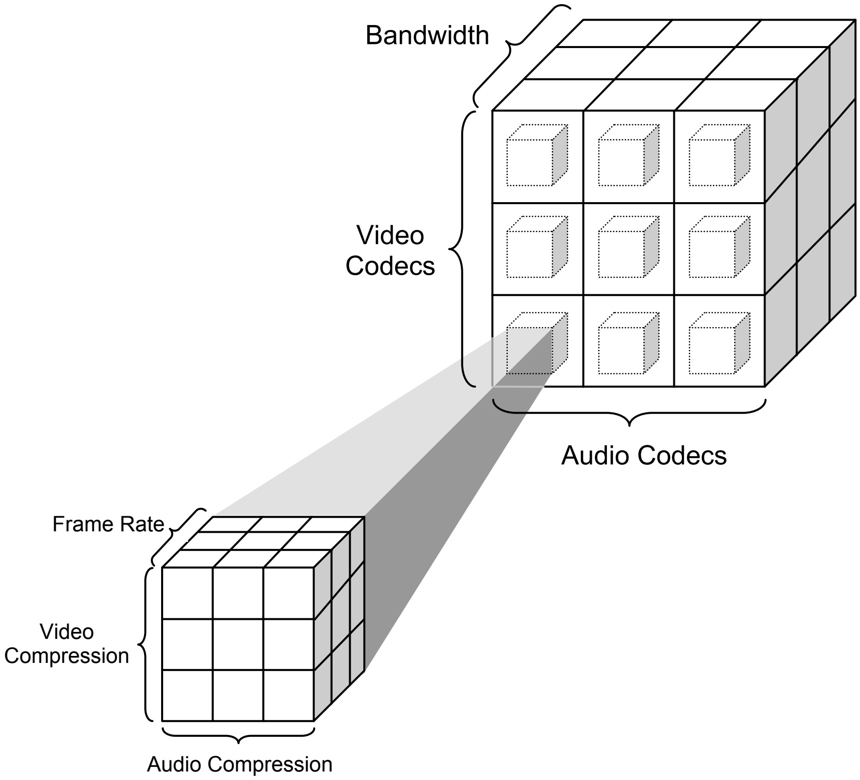
[67] An abridged version of this experiment first appeared in Boring, West, and Moore (2002) and West, Boring, and Moore (2002).

With the advent of quality of service, users must pay for Internet bandwidth.[68] Fortunately, bandwidth is not the only way in which users can maximize streaming quality. As the speed and, correspondingly, the user's cost of accessing streaming content increases, the user may be presented with a number of additional quality parameters. Among these parameters, a user may control the compression settings of the audio and video, the image size of the video, or even the frame rate at which the video is displayed (see Figure 76).

For example, consider a user who wishes to watch a talk show. To reflect the relatively static video images and primary importance of audio in a talk show, the user might select highly compressed video, resulting in poorer image quality. In exchange, the user may opt for minimally compressed, high quality audio (Apteker, Fisher, Kisimov, & Neishlos, 1995). The benefit to the user is that he or she pays primarily for the high-quality audio but not for the video. Such fine-tuning of video and audio allows the user to maximize the streaming quality on selective parameters while minimizing costs.

Providing selectable quality parameters affords the user greater control over his or her interaction with the streaming system. Simultaneously, it greatly increases the complexity of the user's interaction with the system. The complexity of this interaction is compounded when the quality parameter scale is not linear to human cognition. For

[68] The discussion of quality of service applies not only to Internet broadcasting but also broadcasting over wireless digital information transfer such as is used in mobile telephone services.



NOTE: Figure from West, Boring, Dillon, and Bos (2001) used by permission of the authors.

Figure 76. Representation of some of the factors affecting video streaming quality.

example, in a recent study (Bos, 2000), users were allowed to select between four compression settings for a streamed video. Users could select between 25%, 50%, 75%, and 100% quality settings allowed by the video compression algorithm. The problem is that the actual perceived quality did not map onto the users' expected quality for these settings. The users' perception of video quality jumped dramatically from 25% to 50%, but users could not readily differentiate the quality from 75% to 100%. Users expected the perceived quality to increase proportionate to the quality setting. In reality, the settings were a measure of mathematical compression that did not map onto user perception. There is a great disparity between a computer's video quality settings and users' perception of video quality. The goal of the present experiment was to present a method to help users determine quality settings for streaming video.

Numerous approaches exist for assessing user perception of video and audio quality, from traditional five-point subjective scales used by the International Telecommunication Union (1996; 2000a; 2000b) to physiological measures (Wilson & Sasse, 2000). Watson and Sasse (1998) point out that there are serious shortcomings with the scales used by the International Telecommunication Union, including a general inability of the scale labels to translate to different languages. Wilson (2001) argues that the main drawback of using subjective measures to assess video and audio quality is that subjective measures are cognitively mediated, meaning factors other than perception influence users' quality judgements. Wilson's approach to cognitive mediation is to abandon subjective judgments in favor of physiological indicators of stress such as galvanic skin response, heart rate, and blood volume pulse. She suggests that these psychophysiological indicators change reliably according to the level of stress caused by perceptually degraded video and media signals.[69]

There is merit to Wilson's (2001) argument against using subjective scaling measures. The cognitive factors that mediate scale usage have not been adequately controlled in studies of quality perception. This limitation is, however, adequately addressed by applying constrained scaling. Rather than discard subjective measures, in the present experiment I applied constrained scaling methodology to the domain of streaming media quality. It was hypothesized that constrained scaling would minimize the effect of mediating cognitive factors in quality judgements, resulting in lower interparticipant variability compared to conventional subjective scaling methods.

Method

The present study follows the constrained scaling framework outlined in previous experiments by first training participants to use a scale and then asking them to apply the learned scale to novel stimuli. The experiment implements training and testing within the same modality, namely video fluidity of motion. This simple design is implemented to demonstrate the utility of single modality constrained scaling for applied human factors research. Whereas Experiment 14 demonstrated how an achromatic training scale could be applied cross-modality to rate the subjective visual appeal of Web pages, the present

[69] A similar approach has been proposed by Galer and Page (1996) for usability in general. The authors cite the high variability and questionable validity of subjective measures in assessing computer interfaces, instead advocating the use of physiological measures.

experiment demonstrates how a frame rate training scale can be applied to within the same stimulus modality for scaling different video content types.

Participants

Ten participants[70] with self-reported normal vision served as volunteers for the experiment. The participants had not previously participated in a scaling experiment. They were paid $10 for volunteering to take part in the experiment.

Apparatus and Stimulus Materials

The experimental control software was modified to feature video playback. Videos were played back in a 320 x 240 pixel display area that was centered on the screen. The constrained scaling slider and the selection value display window were featured directly below the video display area. Since audio was not part of the scaling stimuli, the experimental control software was not programmed to play audio during the video presentations. Further details on the experimental apparatus are found in Appendix A.

Apteker et al. (1995) suggested that the video content type affects quality judgments. For example, low-action video content may only require a low frame rate to maintain the perception of fluid motion, whereas high-action video content may require a high frame rate to maintain an equivalent perception of fluid motion. In order to control

[70] Originally, it was my intention to conduct the experiment with the conventional five participants. However, upon analyzing the results, it became clear that the cognitive phenomenon at hand required more participants for proper documentation. An additional five participants were run to provide a larger sample through which to understand the results.

for video content type, three levels of action have been selected. A panel of three judges selected three video excerpts according to the level of action. A talking head comedy skit served as the low action video clip, a low impact exercise video was selected as the medium action video clip, and a video of a group of children running was the high action video clip.

The three video clips were captured digitally onto computer and manipulated using Adobe Premiere 5.1 software (Adobe Systems Incorporated, 1998). The National Television System Committee (NTSC) videos were captured using a Videum Winnov video capture card at a resolution of 320 x 240 pixels with 24-bit color resolution at 29.97 frames per second (fps). Each video was edited to be exactly 3 s long.

In order to verify that the videos represented three different action levels, the digitized videos were compressed using the MPEG 4 codec. MPEG 4 compression is highly sensitive to the degree of video change between successive frames of video (Puri & Eleftheriadis, 1998). A high level of image change between frames, as is characteristic of high action video, requires a high level of data to represent the change. Accordingly, a low level of image change between frames, as is characteristic of low action video, requires a low level of data to represent the change.   The MPEG 4 compression analysis confirmed the three action levels selected by the judges. The low action talking head comedy skit compressed to an average data rate of 200.15 kilobits per second (kbps); the medium action exercise video compressed to an average data rate of 355.65 kbps; the high action video of children running compressed to an average data rate of 577.26 kbps.

For optimal playback on the testing computer, the video source files were encoded using the Sorenson video codec (Sorenson Vision, 1997). The bitrate for data throughput was not constrained, allowing for the highest possible video fidelity by the compression algorithm. The three videos were each encoded at five different frame rates—2, 3, 5, 10, and 15 fps. These frame rate values corresponded closely to a loglinear stimulus scale.

Design and Procedure

Figure 77 outlines the experimental design for the present experiment. The participants judged slow and fast action videos across five frame rates. These participants were first trained on the medium action video. They received training on ten iterations of the five frame rates, by first making a fluidity judgment and then receiving feedback about the actual fluidity. The training video had a frame rate to fluidity slope of 1.00.[71] The feedback values were calculated according to the following equation:

$$R = 5F^{1.00}, \tag{44}$$

where R represents the response value and F represents the frame rate. This equation produced a response range from 10.0 to 75.0. The training scale was actually akin to an ordinal scale, with only five response values. To minimize the possibility of categorical scaling artifacts, the participants were instructed that the test videos would not necessarily have the same video fluidity levels as the training video. Following training, the participants received 100 total trials pairing training and either the low or high action

[71] At the time the experiment was conducted, the natural scaling exponent for frame rate was not known. Hence, an exponent value of 1.00 was adopted for training purposes.

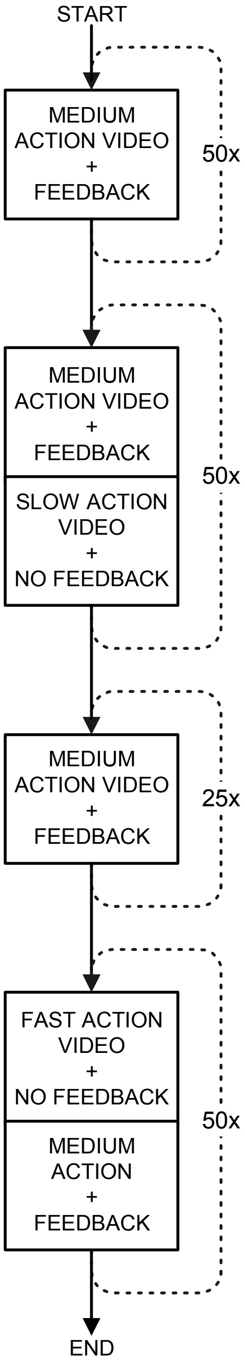


Figure 77. Schematic flow of design in Experiment 14.

testing videos, corresponding to ten iterations of the training and testing videos. A break followed, after which participants were presented with another block with 5 iterations of the training video. The experiment concluded with 100 total trials of the training video interspersed with either the low or high action testing videos, corresponding to the testing video that had not been presented in the earlier trials.

Results and Discussion

As with the previous studies, the coefficient of variation and the highest-to-lowest exponent ratios were calculated across the stimulus sets for scaling exponents, intercepts, and goodness of fit coefficients. The results are summarized in Tables 27 – 28 and Figures 78a – 79b. Additional post hoc analyses were conducted to explain findings. These analyses are presented at the close of this chapter.

Exponent Values

For the medium action training video, the average exponent value was 0.879 with *SD*/*M* = 0.088 and *H*:*L* = 1.348:1. This average exponent value was 0.121 orders of magnitude less than the training exponent. This represents a 12.1% underrating of the fluidity of video movement compared to the feedback values. However, it is important to note that this level of deviation from the training values is not unusual. For example, in Experiment 1, participants on average underrated loudness by 16.5%.[72] Importantly, the coefficient of variation and the highest-to-lowest exponent values were comparable to

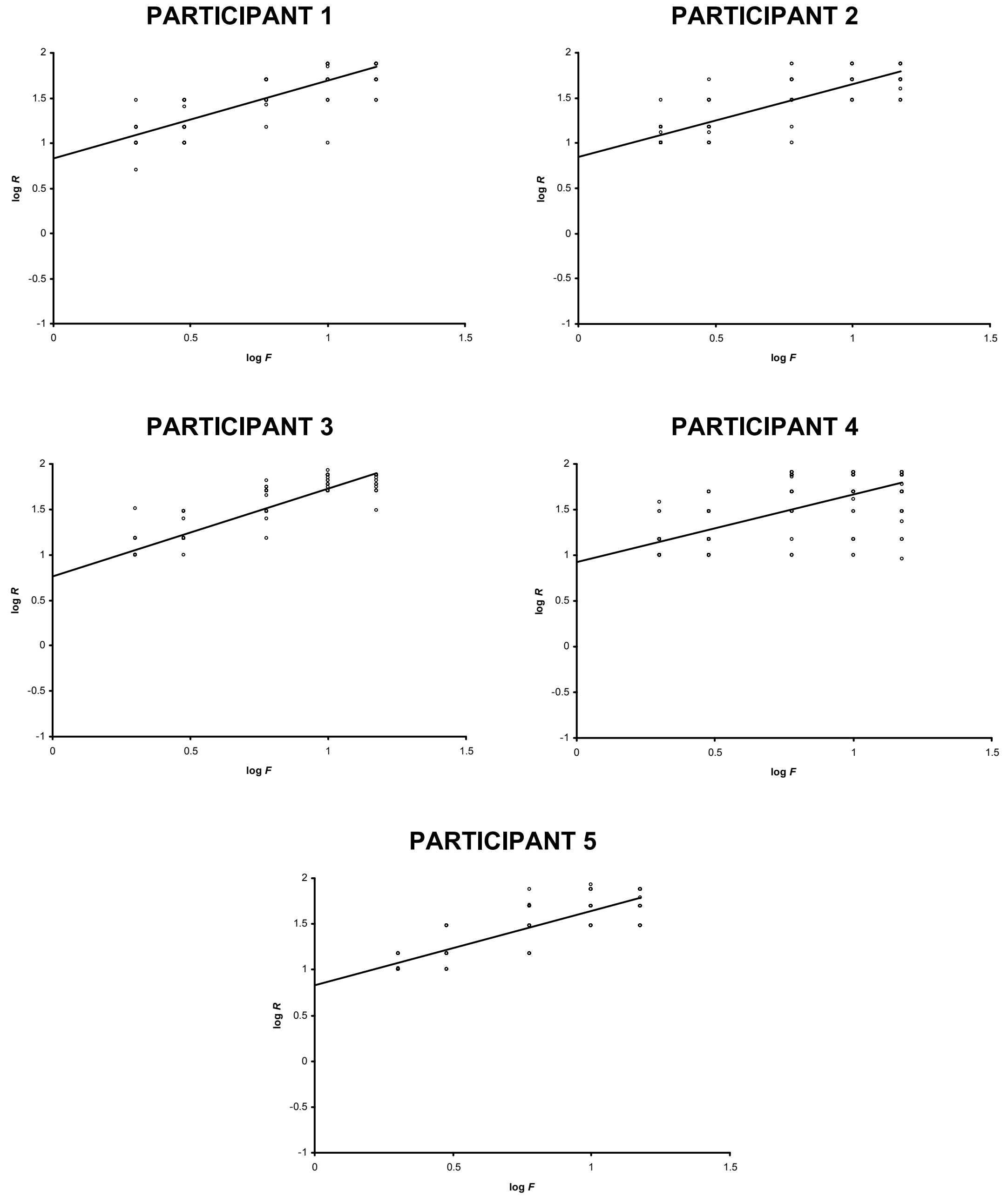
[72] The lower than trained exponent value may be a reflection of a natural scaling exponent that is significantly less than 1.00.

Table 27. Summary of participants' brightness and visual appeal scaling in Experiment 15.

	Medium Action Video + Feedback			Low Action Video + No Feedback			High Action Video + No Feedback		
P	**Exponent**	**Intercept**	***R*²**	**Exponent**	**Intercept**	***R*²**	**Exponent**	**Intercept**	***R*²**
1	0.865	0.828	0.783	1.028	0.656	0.957	0.851	0.817	0.843
2	0.809	0.848	0.769	0.793	0.895	0.679	0.660	0.855	0.687
3	0.967	0.765	0.888	1.734	0.014	0.649	1.031	0.844	0.784
4	0.738	0.925	0.470	0.904	0.786	0.629	0.594	1.016	0.379
5	0.811	0.831	0.750	0.866	0.825	0.804	0.856	0.856	0.820
6	0.923	0.777	0.790	1.007	0.707	0.865	0.924	0.831	0.842
7	0.903	0.779	0.779	1.173	0.578	0.833	0.852	0.818	0.832
8	0.886	0.818	0.848	0.992	0.770	0.886	0.893	0.770	0.873
9	0.995	0.686	0.894	1.174	0.537	0.949	0.734	0.731	0.875
10	0.891	0.795	0.808	0.941	0.751	0.868	0.784	0.841	0.718
M	0.879	0.805	0.778	1.061	0.652	0.812	0.818	0.838	0.765
SD	0.077	0.062	0.119	0.266	0.249	0.120	0.129	0.074	0.149

Table 28. Ratio of motion fluidity ratings for medium action training videos to low and high action testing videos in Experiment 15.

	Ratio of Mid to Low Action Videos			Ratio of Mid to High Action Videos		
P	**Exponent**	**Intercept**	$\boldsymbol{R^2}$	**Exponent**	**Intercept**	$\boldsymbol{R^2}$
1	0.841	1.262	0.818	1.016	1.013	0.929
2	1.020	0.947	1.133	1.226	0.992	1.119
3	0.558	54.643	1.368	0.938	0.906	1.133
4	0.816	1.177	0.747	1.242	0.910	1.240
5	0.936	1.007	0.933	0.947	0.971	0.915
6	0.917	1.099	0.913	0.999	0.935	0.938
7	0.770	1.348	0.935	1.059	0.952	0.936
8	0.893	1.062	0.957	0.992	1.062	0.971
9	0.848	1.277	0.942	1.356	0.938	1.022
10	0.945	1.059	0.931	1.134	0.945	1.125
M	0.854	6.488	0.968	1.091	0.962	1.033
SD	0.127	16.920	0.172	0.142	0.048	0.113



PARTICIPANT 1

PARTICIPANT 2

PARTICIPANT 3

PARTICIPANT 4

PARTICIPANT 5

*Figure 78a. Logarithmic scatterplot and regression line for video training stimuli of variable frame rates (*F*) against average participant rating of video fluidity (*R*) in Experiment 15.*

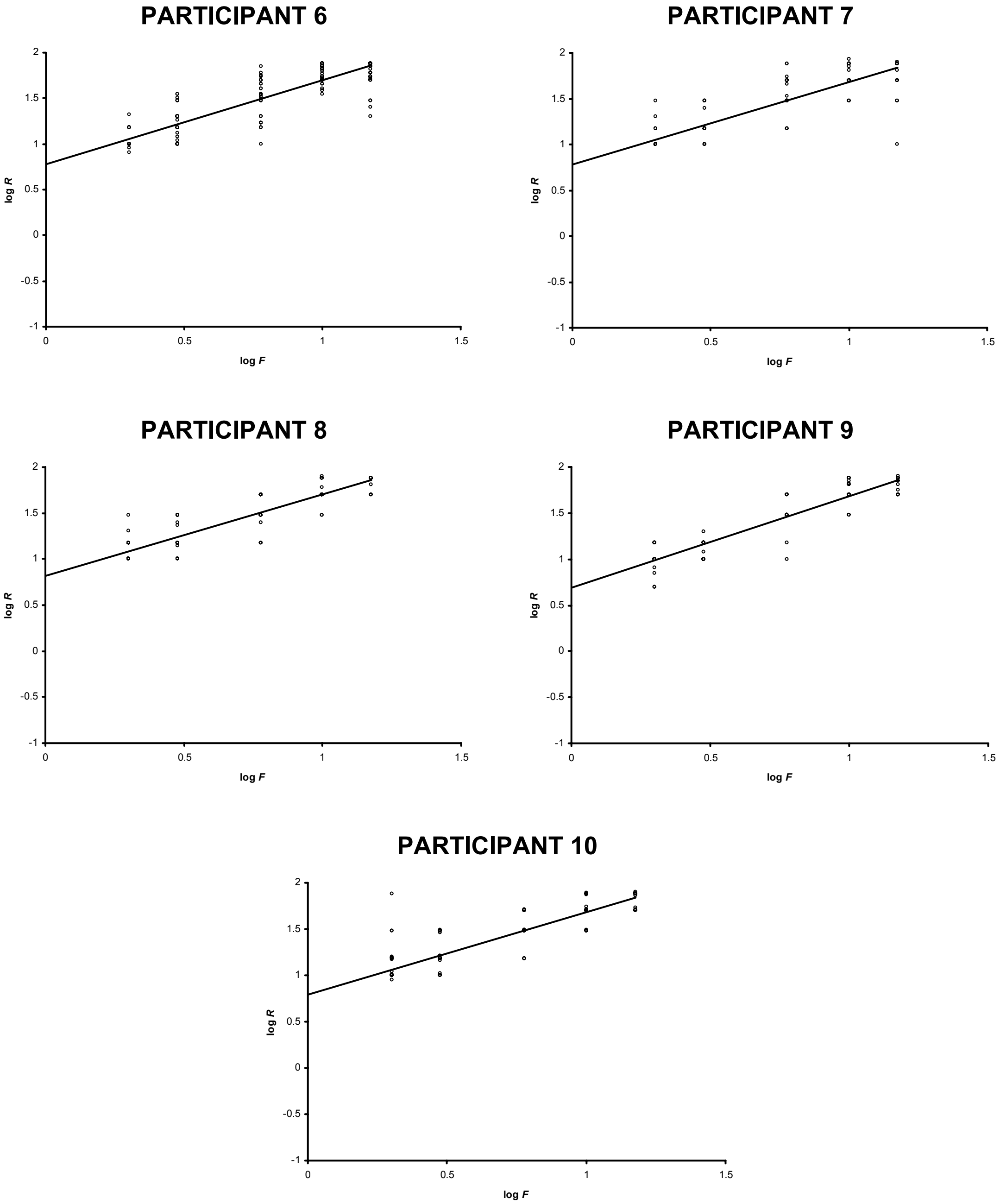


*Figure 78b. Logarithmic scatterplot and regression line for video training stimuli of variable frame rates (*F*) against average participant rating of video fluidity (*R*) in Experiment 15.*

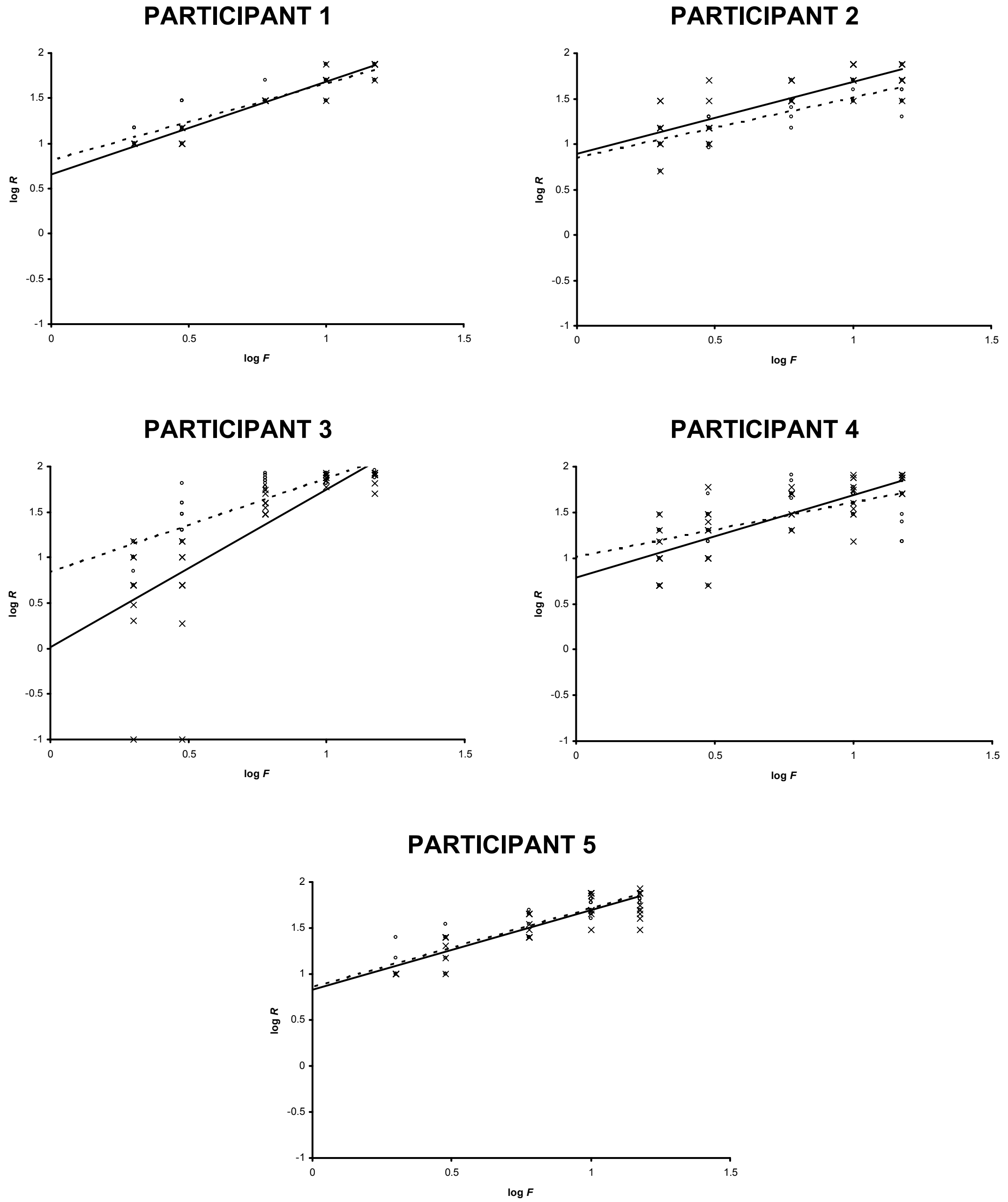


*Figure 79a. Logarithmic scatterplot and regression line for low action (solid line) and high action (dotted line) videos of variable frame rates (*F*) against average participant rating of video fluidity (*R*) in Experiment 15.*

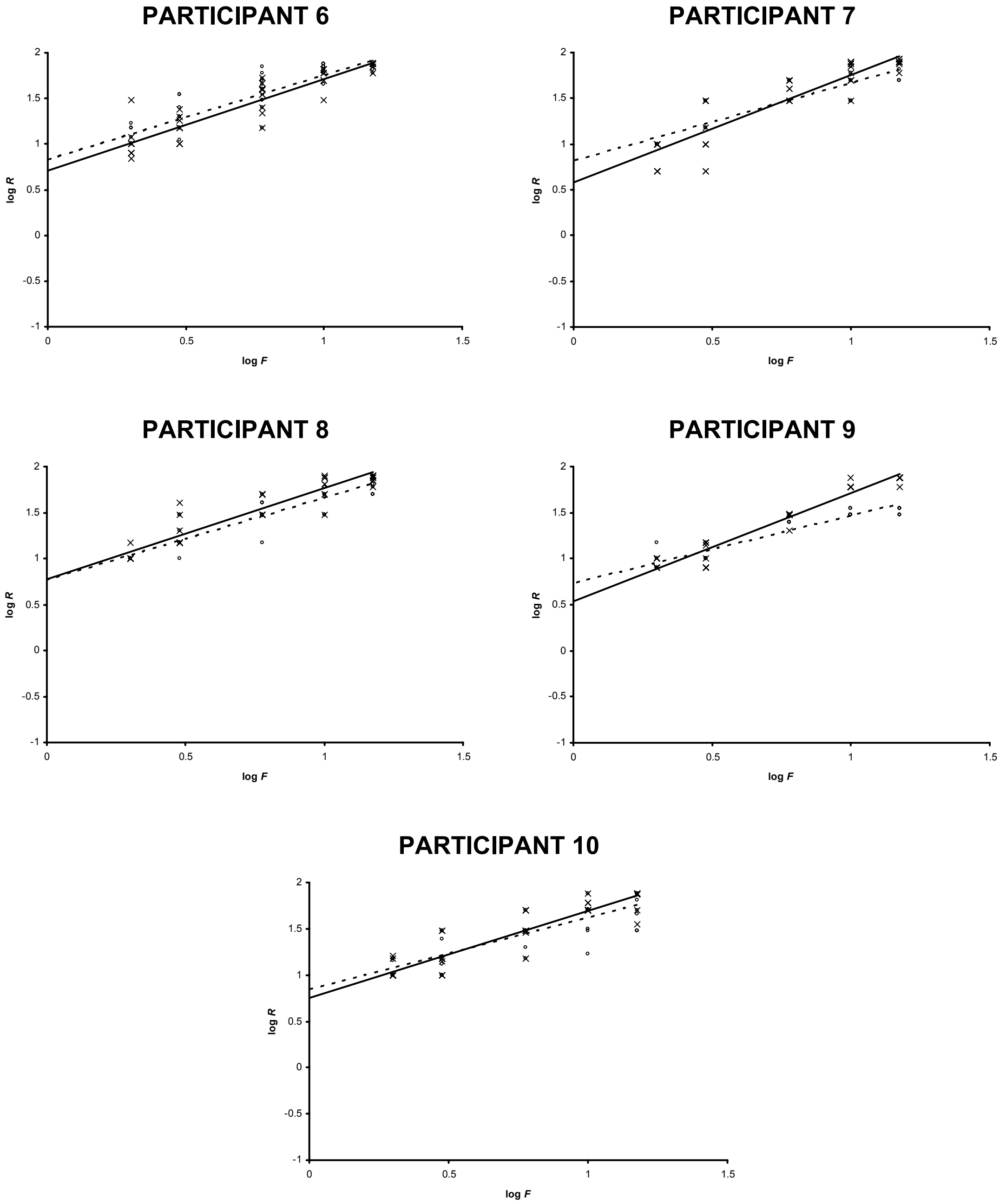


Figure 79b. Logarithmic scatterplot and regression line for low action (solid line) and high action (dotted line) videos of variable frame rates (F) against average participant rating of video fluidity (R) in Experiment 15.

other constrained scaling experiments, suggesting that participants successfully learned the scale, albeit at a diminished response level.

For the low action testing video, the average exponent was 1.061 with *SD*/*M* = 0.250 and *H*:*L* = 2.187:1. For the high action testing video, the average exponent was 0.818 with *SD*/*M* = 0.157 and *H*:*L* = 1.521:1. The apparent pattern suggests that the fluidity of motion slope flattens out as the level of video action increases. For both low and high action videos, the variability was greater than is typical for constrained scaling experiments. The average exponent ratio of the medium action training video to the slow action testing video was 0.854 and to the fast action testing video was 1.091.

Intercept Values

The average intercept value for the medium action training video was 0.805 with *SD*/*M* = 0.077 and *H*:*L* = 1.348:1. The intercept value variability was in line with expected constrained scaling results. For the low action video, the average intercept was 0.652 with *SD*/*M* = 0.382 and *H*:*L* = 63.029:1. Note that the variability is high, especially for the highest-to-lowest exponent value, due to the fact that Participant 3 had a near zero intercept. If Participant 3 were excluded from the data set, the average intercept value would be 0.723 with *SD*/*M* = 0.160 and *H*:*L* = 1.667:1. For the high action video, the average intercept was 0.838 with *SD*/*M* = 0.088 and *H*:*L* = 1.390:1. Omitting Participant 3, there was a slight tendency for the intercept to increase as the level of action in the video increased. With the exception of the intercept for low action video by Participant 3, the variability fell within the expected range for constrained scaling. The average intercept ratio of the medium action training video to the low action testing video was 6.488 (or 1.138 without Participant 3), and to the high action testing video it was 0.962.

Goodness of Fit Coefficients

The average R^2 value for the medium action training video was 0.778 with SD/M = 0.153 and H:L = 1.902:1. For the low action testing video, the average R^2 value was 0.812 with SD/M = 0.148 and H:L = 1.521:1. For the high action testing video, the average R^2 value was 0.765 with SD/M = 0.195 and H:L = 2.309:1. The average goodness of fit was at the expected level for constrained scaling, although variability was somewhat greater than expected. The average R^2 ratio of the medium action training video to the low action testing video was 0.968, and to the high action testing video, it was 1.033.

Post Hoc Analyses

Because the testing video variability was greater than expected for a constrained scaling experiment, additional analyses were conducted to determine the source of the variability.[73] To eliminate scaling noise, the response values were averaged for each frame rate for each participant. Figures 80a and 80b illustrate the relationship between the variable frame rate stimuli and the response values for low and high action videos. Instead of an overall regression line, a connecting line was drawn between each frame rate value.

[73] It should be clear that these post hoc analyses are designed to elucidate the previous results, not supplant them. The post hoc analyses are provided as exploratory avenues for further research and therefore employ a less stringent experimental basis.

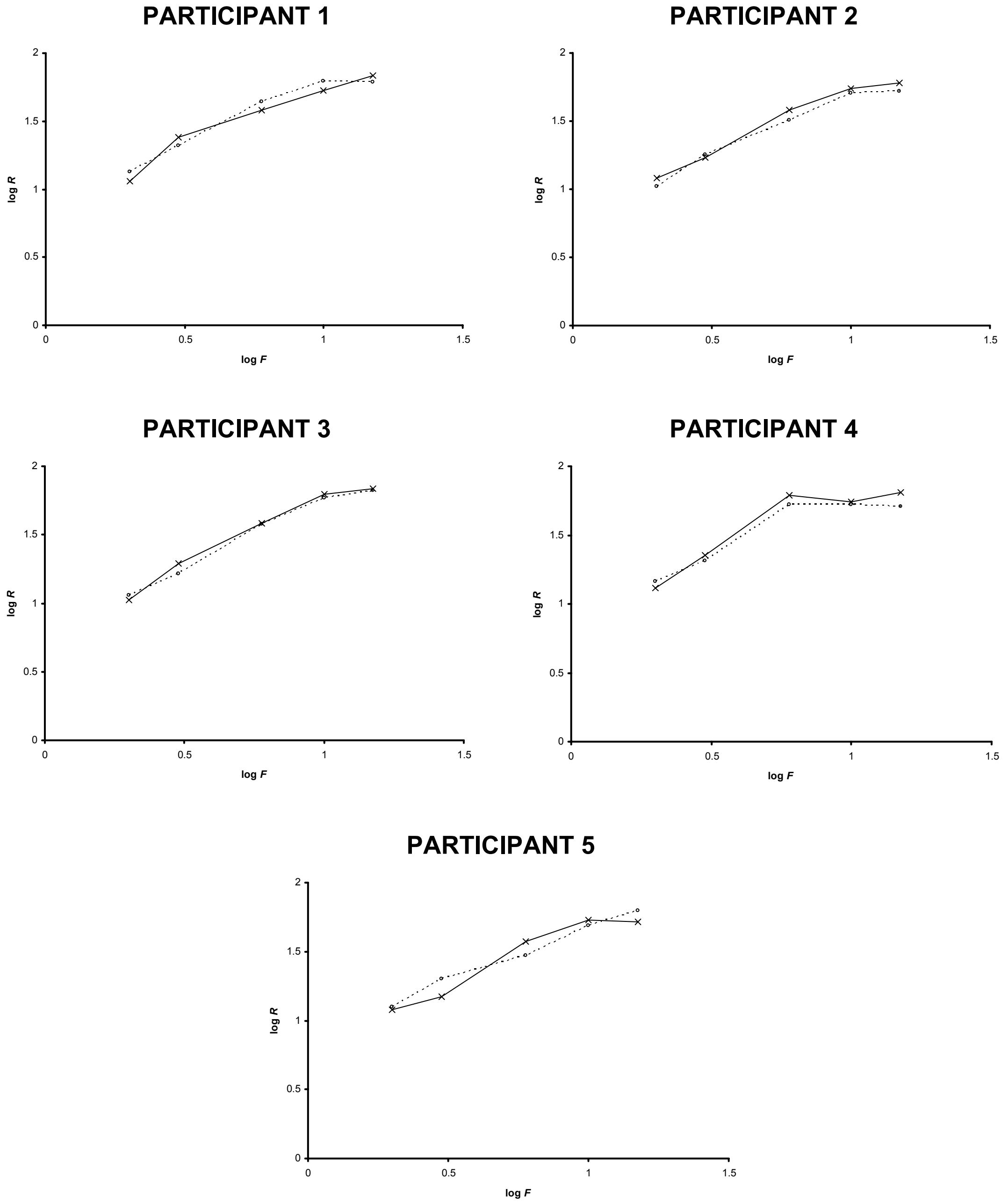
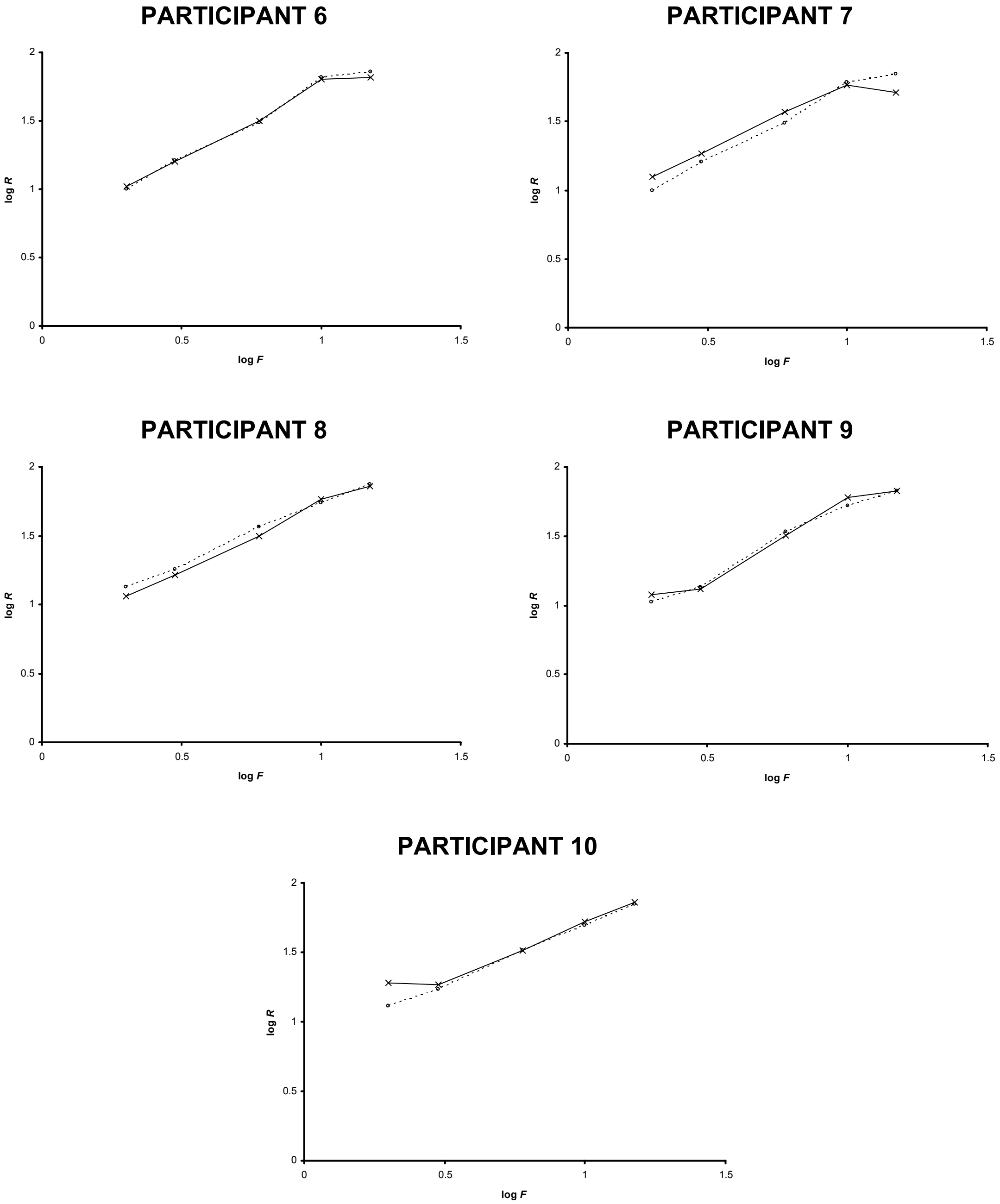


*Figure 80a. Logarithmic scatterplot with connecting lines of the averaged participant rating of video fluidity (*R*) for the five frame rate values (*F*) for low action (solid line) and high action (dotted line) videos in Experiment 15.*



PARTICIPANT 6

PARTICIPANT 7

PARTICIPANT 8

PARTICIPANT 9

PARTICIPANT 10

*Figure 80b. Logarithmic scatterplot with connecting lines of the averaged participant rating of video fluidity (*R*) for the five frame rate values (*F*) for low action (solid line) and high action (dotted line) videos in Experiment 15.*

The connecting line allows a visual inspection of the linearity of the data in logarithmic space. As can be seen, several participants displayed strong deviation from linearity. Nonlinearity is not uncommon when using magnitude estimation (Luce & Mo, 1965), although its occurrence in a constrained scaling experiment is novel. The typical procedure for these outlying participants would be to discard them or average across them. However, since I am interested in individual differences, I note that these four participants were less able than the other participants to exploit the scaling aids offered by constrained scaling. These individual differences may occur in strategy, cognitive ability, or the effort invested by the participants—factors that may not be adequately controlled for simply by using a constrained scaling method. Since these deviations from linearity were not large by magnitude estimation standards, I analyzed the data both with them (as described earlier in this *Results and Discussion* section) and without them. The analysis excluding the outliers follows.

The most prominent departure from linearity was found in Participants 4, 6, 7, 9, and 10. Excluding these participants from the analysis, however, did little to ameliorate variability in the results. For example, in terms of the exponent values, the results were virtually unchanged for medium action [$M = 0.867$, $SD/M = 0.075$, $H{:}L = 1.195{:}1$] and high action [$M = 0.858$, $SD/M = 0.155$, $H{:}L = 1.562{:}1$] videos, but the variability actually increased for low action videos [$M = 01.083$, $SD/M = 0.348$, $H{:}L = 2.187{:}1$].

A further post-hoc analysis indicated that the participants seemed to fall into two different categories—those participants who closely followed the training exponent for the medium action video and those participants who deviated from the training exponent for those same videos. The groups were systematically divided based on the exponent of their regression lines for the medium action training video. The groups were divided depending on their deviation from the exponent 1.0, which indicated perfect learning. The first group included individuals whose learning exponent deviated from 1.0 by 0.10 or less. The second group included individuals whose learning exponent deviated from 1.0 by 0.10 or more. The former group included Participants 3, 6, 7, and 9, while the latter group included Participants 1, 2, 4, 5, 8, and 10. The participants who closely followed the training exponent had an average exponent equal to 0.947 with *SD*/*M* = 0.044 and *H*:*L* = 1.102:1. These same participants showed improved variability for low action [*M* = 1.272, *SD*/*M* = 0.250, *H*:*L* = 1.722:1] and high action [*M* = 0.885, *SD*/*M* = 0.141, *H*:*L* = 1.405:1] videos. For the participants who did not closely follow the training exponent, the average exponent was 0.833 with *SD*/*M* = 0.070 and *H*:*L* = 1.205:1. Again, these participants showed improved variability for low action [*M* = 0.921, *SD*/*M* = 0.093, *H*:*L* = 1.296:1] and high action [*M* = 0.773, *SD*/*M* = 0.156, *H*:*L* = 1.503:1] videos. This latter group exhibited considerably lower variability for low and high action videos than did the group that closely followed the training exponent.

This finding suggests that the two groups followed different strategies for scaling, but both benefited from the training. The variability in both groups was congruent with findings from other constrained scaling experiments. As shown in the original analyses, a failure to recognize that there were two different scaling strategies resulted in noisy, aggregated data that were atypical for constrained scaling results. By properly recognizing the presence of two scaling strategies, it was possible to disentangle the variability levels according to the distinct groups. The resulting variability levels were typical for constrained scaling experiments.

General Discussion

Computer users who need to adjust the quality parameters of streaming media will want to ensure that the adjustments they make are perceptible and meaningful. In this chapter, I demonstrated that a user trained on a scale for medium action video was able to apply that scale to low and high action videos as well. Arguably, this learning would translate directly into making the kind of parameter adjustments required of quality of service in streaming media.

The present experiment provides a compelling case for the utility of constrained scaling for practical, applied dimensions. However, this experiment is only a starting point for other application oriented constrained scaling. Even within streaming media, much follow-on research remains. For example, since different quality parameters scale differently to human perception, it would useful to develop a universal scale to which user judgments about all quality parameters could be calibrated. Future work should show how scales such as the brightness training scale might be applied cross-modally to aid users in making quality judgments across a complete range of streaming media parameters. By reducing variability, such scales would ensure that the selection values users selected matched the actual resultant parameters.

CONCLUSIONS

General Findings

What have these 15 experiments on constrained scaling ultimately revealed? To answer this question, it is necessary to revisit this dissertation in two separate passes. The first pass is at the microscopic or analytic level, reviewing the findings from the individual experiments. The second pass is comprised of a macroscopic or holistic analysis of what recurrent themes emerged across the experiments. This second pass comes at the end (see the *Final Thoughts* section in this chapter). First, I examine the individual experiments. To do so, I consider each experiment according to its contribution as a replication, refinement, extension, or application experiment. A summary of the findings from each experiment is also found in Table 29.

Replication

Only one experiment was, strictly speaking, a replication experiment. Experiment 1 replicated the general loudness constrained scaling design and procedure found in West et al. (2000). The results of Experiment 1 mirrored the results from West et al. While replication is a worthwhile pursuit onto itself, what sets this experiment apart is that it featured a new implementation of the experimental apparatus for constrained scaling. The experimental apparatus was implemented entirely as software within Microsoft Windows, utilizing standard computer hardware. West et al. had required specialized equipment for their experiments. Apart from the sound meter for stimulus calibration and the sound attenuating chamber as a testing environment, Experiment 1 used no special equipment.

Table 29. List of the experiments and corresponding findings.

Experiment Number	Experiment Description	Experiment Finding
1	Basic loudness experiment replicating West et al. (2000)	Current experimental apparatus successfully replicates West et al. (2000) without need for specialized hardware
2	Basic loudness experiment as in Experiment 1, but without sound attenuating chamber	Loudness constrained scaling can be implanted successfully without the need for a sound attenuating chamber
3	Cross-modal matching experiment: Loudness → Brightness	Cross-modal constrained scaling exhibits scale carryover from the training to the testing stimuli
4	Cross-modal matching experiment: Brightness → Loudness	Cross-modal constrained scaling is susceptible to stimulus range effects
5	Cross-modal matching experiment: Short-interval brightness → Loudness	Constrained scaling results are improved for flashed vs. constant brightness stimuli
6	Color brightness magnitude estimation	Conventional computer monitors are effective for displaying color brightness stimuli
7	Color brightness experiment	Constrained scaling significantly improves color brightness scaling reliability compared to magnitude estimation
8	Brightness with categorical learning	Continuous scale stimuli are more effective categorical stimuli for scale training
9	Brightness with categorical learning plus random feedback noise	Random noise added to categorical training stimuli reduces rote memorization and improves scaling performance
10	Brightness with decreased feedback ratio	Optimal ratio of training to testing trials is 1:1
11	Perceptual scale triangulation: Same stimuli trained to different brightness scales	The ratio of training to testing scaling exponents is constant within a modality
12	Subjective triangulation magnitude estimation: Money → Happiness	There are considerable individual differences in scaling the subjective utility of money
13	Subjective triangulation: Same stimuli trained to different brightness scales	Constrained scaling increases the sensitivity to individual differences in psychometric scaling
14	Scaling of Web page visual appeal	Constrained scaling is more sensitive than magnitude estimation to individual differences in subjective visual appeal of Web pages
15	Scaling of video quality of service	Constrained scaling can be applied successfully to aid software users in making parameter selections for streaming media

Refinement

The hallmark of refinement experiments is that they take the existing constrained scaling methodology and incorporate a new element, potentially improving the efficacy of the method. Refinement experiments are not always orthogonal to replication or extension experiments. Within a margin of overlap to these classes of experiments, refinement experiments focus on the method of constrained scaling, often replicating elements of an existing experiment but not extending into novel scaling domains. There are two clusters of refinement experiments, consisting of Experiments 2 and 3 and Experiments 8 – 10.

Experiment 2 is closely related to Experiment 1. It is a complete replication with the exception that it features the relocation of the participant outside the sound attenuating chamber. The results mirrored the results obtained within the sound attenuating chamber, suggesting that it was feasible to conduct loudness constrained scaling experiments without extensive sound dampening efforts.

Experiment 3 was closely related to an experiment in West et al. (2000), in which a learned scale for loudness was applied cross-modally for scaling brightness stimuli. The experiment served as a refinement over the earlier experiment in that it used a calibrated computer monitor for displaying brightness stimuli. The earlier experiment by West et al. had used a calibrated light-emitting diode (LED) for displaying brightness stimuli. The results in Experiment 3 failed to replicate the earlier findings, which was attributed to the diminished luminous intensity of the cathode ray tube display compared to the LED.

The next cluster of refinement experiments (Experiments 8 – 10) explored the nature of the training stimuli. In Experiment 8, a categorical training scale was used instead of the typical continuous training scale. Participants were trained on only five stimulus values along the range of the scale, approximating an ordinal scale. Compared with conventional constrained scaling results using a continuous training scale, categorical constrained scaling increased variability and changed the exponent and intercept values. It was concluded that categorical constrained scaling was not as effective for training as conventional constrained scaling.

Experiment 9 was similar to Experiment 8, except random noise was added to the feedback values given to participants. This random noise served the purpose of eliminating rote memorization as the basis for the decreased effectiveness of categorical constrained scaling. The results of noisy categorical constrained scaling closely matched the results for conventional constrained scaling, demonstrating that it was possible to implement constrained scaling with a reduced number of training stimuli as long as participants weren't able to memorize the training scale. Memorized training scales resulted in decreased scale mastery and generalization to novel stimuli.

Experiment 10 explored a final refinement in the presentation of the training stimuli. In Experiment 10, the feedback trials were halved during the testing phase. Instead of a training trial between every testing trial, a training trial was presented after every second testing trial. While the decreased feedback ratio did not affect the scaling exponent or intercept values, it did significantly increase scaling variability. It was concluded that the optimal feedback ratio should remain at one training trial for every testing trial. Decreasing the feedback ratio decreased the time required for a constrained scaling experiment, but it also decreased the effectiveness of constrained scaling.

Extension

Almost half of the experiments were extensions of constrained scaling into a novel scaling domain. Experiments 4 and 5 extended cross-modal scaling between loudness and achromatic brightness stimuli. Experiments 6 and 7 further extended constrained scaling to color stimuli. Experiments 11 – 13 combined to extend constrained scaling to rate the degree of happiness elicited by different amounts of money. This latter triplet of experiments established the utility of constrained scaling for psychometric scaling.

Experiment 4 reversed the design of Experiment 3 by using achromatic brightness stimuli for training and subsequently testing loudness stimuli. Participants appeared to carry over verbatim the learning exponent from the brightness stimuli to the loudness stimuli, resulting in unexpected exponent and intercept values and higher than expected variability. This experiment revealed limitations on cross-modality constrained scaling, suggesting caution should be exercised when training in one modality and testing in another modality.

Experiment 5 was similar to Experiment 3, except the display time for the brightness stimuli was shortened to test the possible effect of light adaptation on the results. The results approximated the results from Experiment 3, exhibiting substantial scaling carryover from the training modality to the testing modality.

To avoid the complications of cross-modal scaling, Experiments 6 and 7 focused on a single modality. Experiment 6 was a magnitude estimation experiment on the perceived brightness of grayscale, red, green, and blue stimuli. In Experiment 7, participants were trained on a brightness scale for grayscale stimuli and then tested on the color stimuli. The results showed significant improvement in scaling consistency for constrained scaling vs. magnitude estimation. The results were not as conclusive for the blue stimuli. However, confounds were noted for the display of the blue stimuli.

Finally, Experiments 11 – 13 utilized a multiphenomenal within-method triangulation technique to determine the efficacy of constrained scaling for a psychometric domain. Experiment 11 consisted of two phases in which participants were trained on grayscale stimuli and tested on color stimuli. The training exponents were varied across the test phases. Participants were successful at learning both training exponents and using both learned scales on the color stimuli. The purpose of this triangulation experiment was to establish the relationship between scaling exponents. The hypothesized relationship between training and testing exponents held constant across both experimental phases.

Experiment 13 segued directly from Experiment 11. Instead of testing a perceptual stimulus, Experiment 13 tested the subjective utility of various amounts of money. Because the level of happiness elicited by money is a matter of subjective perception, it was not expected that there would be consistency across participants. It was, however, expected that the participants' subjective perceptions would hold relatively constant from one testing phase to the next. Therefore, the scaling ratio of brightness training to money testing was expected to hold across the two testing phases, demonstrating the mapping of a learned perceptual scale to a subjective domain. This ratio did not hold across all participants, bringing into question the role of constrained scaling in psychometric scaling.

The matter is not cut and dry, since there was evidence in support of the value of constrained scaling for psychometrics. Experiment 12 featured a simple magnitude estimation of the relationship between monetary amounts and happiness, which was a baseline experiment against which to compare the results from Experiment 13. Constrained scaling in Experiment 13 resulted in higher variability than constrained scaling in Experiment 12. Further, the results for the brightness training scaling in Experiment 13 exhibited minimal variability and strong evidence of exponent and intercept mastery. These two findings combine to suggest that constrained scaling may have, in fact, helped to capture individual differences in psychometric scaling. Consistent mastery of the perceptual training scale and increased scaling variability for the psychometric domain compared to magnitude estimation point to the possibility that constrained scaling may have acted to increase scaling sensitivity to individual differences. The cross-modal mapping of the perceptual training scale to the subjective domain either failed to function in a linear manner across the two experimental phases. Given the dramatically different performance in psychometric scaling in Experiment 13 compared to Experiment 12, it would be premature to dismiss psychometric constrained scaling.

Application

The final two experiments demonstrated how constrained scaling could be applied. Application is a special case of extension, in which the novel scaling domain corresponds to real-world utility. In this case, both applications were related to software human factors.

In Experiment 14, constrained scaling was used to evaluate the subjective visual appeal of Web pages. The results, when compared to conventional magnitude estimation scaling in Fernandes (2003), revealed successful mastery of the training scale coupled with increased scaling variability for visual appeal. Borrowing on the logic of Experiment 13, the results suggested that constrained scaling increased the sensitivity to true individual differences.

Experiment 15 featured constrained scaling in the rating of frame rate in streaming video. The experiment revealed higher than expected variability. Post hoc analyses showed that there were two apparent scaling strategies employed across participants. Both of these strategies resulted in the low variability typical of constrained perceptual scaling and showed the benefit of training for helping users select quality parameters.

Limitations

The many insights gained through these constrained scaling experiments are necessarily seen against a backdrop of pragmatic limitations. Limitations do not represent critical flaws so much as lessons learned for improving future iterations of the experiments. Four main limitations were present across many of the experiments:

1. *Small sample size.* A priori, I opted for the use of a small sample size for the experiments on the grounds that five participants would be sufficient for descriptive analyses and that the large effect size afforded by constrained scaling allowed sufficient statistical power for inferential analyses. Given the finite budget available for paying participants, the use of small sample sizes allowed more experiments to be conducted and a greater number of evidential insights to be gained than would have been possible if more participants had been required for each experiment. Nonetheless, even with large effect sizes, there are limits in the generalizability of the results from the experiments. However sound the experimental design, there is questionable face validity in generalizing from a sample of five participants to the population. Now that the basic findings have been established, all experiments in this dissertation would certainly benefit from the inclusion of additional participants to reconfirm the results and corroborate the conclusions.
2. *Restricted scale range.* The original constrained scaling experiments (West & Ward, 1994; West et al., 2000) featured a loudness scale that covered a wide range of the decibel scale and mapped well to low and high mental magnitudes. As discussed in Experiments 3 and 4, the grayscale training scale used throughout much of this dissertation was punctuated by hardware limitations in terms of maximal display brightness. As such, the magnitude presentation range was not as wide as in the loudness experiments, potentially creating a bounded mapping of mental magnitudes to the numeric scale. If the participant were subsequently presented with a secondary stimulus range for tested scaling, the learned scale would not necessarily provide a

perspicuous comparison to the novel testing scale. The present experiments failed to determine the transferability of a restricted training scale range to a domain with a larger perceptual or subjective magnitude range.

3. *Alternate scaling strategies.* In the final experiment, there were participants who did not exhibit the same degree of scale mastery as the majority of participants. In the case of psychometric scaling, such results would be expected on the basis of individual differences in subjective experience. The potential reasons for such differences when scaling a perceptual domain are unclear. The present experiments were not operationalized in such a way to highlight possible differences in scaling strategies, and they do not provide clear explanations of alternate scaling strategies adopted by participants. Understanding why, when, and how participants adopt alternate scaling strategies would directly inform efforts to model the cognitive underpinnings of scaling.
4. *Cross-modal constrained scaling.* The experiments failed to provide a definitive account of cross-modal constrained scaling. Experiments 3 – 5, which focused on the cross-modal scaling of loudness and brightness, failed to produce a consistent mapping, with clear evidence of some direct exponent and intercept carryover from the training stimuli to the testing stimuli. Similarly, Experiments 13 and 14 showed a seemingly contradictory set of results, in which participants exhibited excellent mastery of the training scale yet exhibited considerable variability for scaling the test stimuli. This result was attributed to the expected individual differences in the testing stimuli. This compelling argument was undermined by the failure of over half of the

participants to be consistent in their subjective scaling across two phases in Experiment 13. There is considerable ambiguity associated with cross-modal constrained scaling results, making it impossible to provide a parsimonious account of this type of scaling or to come to a verdict about the benefit of constrained scaling for cross-modal research.

Future Research

Future research should address the limitations that are found in the present experiments. As well, the present experiments are but the starting point for the type of questions that can be asked about mental magnitudes and scaling. There are several topics that I foresee as having long-term importance to the enterprise of calibrated mental scaling. These areas of research will drive future efforts at cognitive modeling and theory building in the domain of scaling.

1. There are still many interesting questions about the role of consciousness in scaling that remain unaddressed by the present research. It is unclear, for example, to what extent scaling would be possible without conscious awareness. It is equally unclear what role magnitudes play in consciousness. Future research should aim to determine the interplay of introspective scaling, mental magnitudes, and consciousness.
2. There is currently no definitive neuroscientific account of psychological scaling, nor has there been an attempt to sketch a neuropsychological account of magnitude vs. categorical cognition. It would be fruitful to bridge neuropsychological models of scaling as well as neuropsychological models of classification and categorization. Until current models of cognition and scaling are merged with insights from

neuroscientific studies, there will necessarily remain an element of functionalist black boxism to the theory behind constrained scaling.

3. I have not fully explored the topic of dynamic vs. steady state cognition and its role in scaling. Ward (2002) has sketched a useful account of dynamical cognitive science with interesting links to dynamical psychophysics. At this point, constrained scaling appeals to both dynamic and steady state theorists in cognition, since it both controls for dynamic processes and minimizes constant response biases. With the right manipulations, constrained scaling may ultimately help reconcile these two views or clarify which view better accounts for the mental phenomena of calibrating a mental scale.
4. As more insight is gained into the cognitive constraints of scaling, this information may begin reaching the fertile point where it is possible to model the mental processes involved in translating mental magnitudes into scale values. A good cognitive model of scaling would go a long way toward understanding the effects of constrained scaling on the mind and accounting for categorical and magnitude-based aspects of the mind.
5. Finally, it should be noted that these experiments, like earlier constrained scaling experiments, focused on unidimensional scaling. Much interesting research has been conducted in the field of multidimensional scaling. Future research should attempt to apply constrained scaling within a multidimensional scaling framework to determine its applicability and potentional contribution to this important type of scaling.

There are certainly countless other research questions regarding mental magnitudes or scaling. But, there are clearly many more questions than there is time to answer them in this or other monographs on scaling. In this dissertation, I have simply attempted to pick a few of the most methodologically and theoretically interesting and relevant questions to answer through my research, leaving many inroads for future research.

Final Thoughts

Over the past 312 pages, I have demonstrated the importance of mental magnitude to cognition. I have illustrated the importance of a cognitive-based model of scaling for improving the quality of mental measures cognitive scientists are able to obtain. Through a series of 15 experiments, I clarified and extended constrained scaling as a method and expanded it as a model that incorporates psychometric as well as psychophysical data. Constrained scaling was hypothesized to have a double-edged effect for scaling modalities. For those domains in which there are minimal individual differences, such as perception, constrained scaling works to minimize scaling variability. In contrast, where real individual differences exist, such as in psychometric research, constrained scaling seems to increase the sensitivity for scaling those differences. Finally, constrained scaling was shown to have real-world applicability. Experiments 14 and 15 demonstrated human factors applications of constrained scaling, in which scaling augmented current approaches to investigating novel technological domains.

The experiments in this dissertation make three important points:

1. *Constrained scaling is easy to implement.* Previous research on constrained scaling had focused primarily on the scaling of loudness stimuli in a sound attenuating chamber, a process that was aversive for some participants as well as difficult for researchers to implement. The experiments in this dissertation showed how other, more easily implemented scales such as the brightness scale could be implemented for training purposes and applied to a variety of scaling domains. This dissertation demonstrated that constrained scaling need not exist solely in a highly controlled psychophysical laboratory environment, nor does constrained scaling require an experimental apparatus beyond the standard personal computer that is ubiquitous in psychological research facilities and office work environments. The method was refined and found to be robust even as the experimental constraints on testing were loosened. As a result, the perceived cost in terms of effort required for conducting constrained scaling experiments no longer exceeds the actual benefit of improved scaling reliability.
2. *Constrained scaling is flexible across domains.* This dissertation featured several extension experiments, which were designed to showcase how constrained scaling might be used across a variety of domains. Constrained scaling was applied to color perception, the subjective utility of money, the visual appeal of Web pages, and the frame rate settings for streaming video. These extensions are examples of the potential for constrained scaling to be implemented in a broad range of psychophysical and psychometric domains.

3. *Constrained scaling is an important key to unraveling mental processes.* Psychological measurement, including psychophysical and psychometric scaling, attempt to reveal cognitive functioning. Poor measurement methods can be sentinels to the mind, providing titillating hints of cognition but ultimately failing to reveal the nature of mental processes because of measurement noise, biases, and artifacts. Constrained scaling is unique as a psychological method in that it controls the communicative aspect of cognition while leaving other cognitive processes to proceed along their normal course. It calibrates the production of scale values without impinging on the natural permeation of mental magnitudes. By increasing the reliability of participants' introspective elicitations, constrained scaling brings researchers one step closer to illuminating the black boxes of cognition.

Constrained scaling is not a panacea for psychological measurement; constrained scaling is simply a method for improving human scaling reliability. It infuses a cognitive approach into the scaling literature, modeling the interchange of mental magnitudes and numeric expression. As such, constrained scaling is a useful tool to be employed in the cognitivist's tool chest of methods. It is my hope that this dissertation effectively illustrates the multifaceted capabilities for psychological elicitation available through the constrained scaling method. Constrained scaling has sufficiently improved on magnitude estimation to become the veritable Swiss Army knife of scaling. Perhaps it is poised to take on a similar role in other cognitive research.

REFERENCES

Adelson, E.H. (1982). Saturation and adaptation in the rod system. *Vision Research*, 22, 1299-1312.

Adobe Systems Incorporated. (1998). *Adobe Premiere 5.1*. San Jose, CA: Adobe Systems Incorporated.

Aiba, T.S., & Stevens, S.S. (1964). Relation of brightness to duration under light and dark adaptation. *Vision Research*, *4*, 391-401.

Algom, D. (1992). Memory psychophysics: An examination of its perceptual and cognitive prospects. In D. Algom (Ed.), *Psychophysical approaches to cognition* (pp. 444-513). Oxford: North-Holland.

Algom, D., & Marks, L.E. (1990). Range and regression, loudness scales, and loudness processing: Toward a context-bound psychophysics. *Journal of Experimental Psychology: Human Perception & Performance*, *16*, 706-727.

Apteker, R.T., Fisher, J.A., Kisimov, V.S., & Neishlos, H. (1995). Video acceptability and frame rate. *IEEE Multimedia*, *2*(3), 32-40.

Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In K. Spence & J. Spence (Eds.), *The psychology of learning and motivation. Volume 2.* New York: Academic Press.

Baker, H.D. (1949). The course of foveal light adaptation measured by the thresholod intensity increment. *Journal of the Optical Society of America*, *39*, 172-179.

Bedell, W. (2000). Get accurate results with code-based timers. *Visual Basic Programmers Journal*, *10*(4), 36-43.

Beringer, J. (1992). Timing accuracy of mouse response registration on the IBM microcomputer family. *Behavior Research Methods, Instruments and Computers*, *24*, 486-490.

Berka, K. (1992). Are there objective grounds for measurement procedures? In C.W. Savage & P. Ehrlich (Eds.), *Philosophical and foundational issues in measurement theory* (pp. 181-194). Hillsdale, NJ: Lawrence Erlbaum.

Biederman, I. (1995). Visual object recognition. In S.M. Kosslyn & D.N. Osherson (Eds.), *An invitation to cognitive science. Second edition. Volume 2. Visual cognition* (pp. 121-165). Cambridge, MA: MIT Press.

Blaikie, N.W.H. (1993). *Approaches to social enquiry*. Cambridge, UK: Polity Press.

Boring, E.G. (1950). *A history of experimental psychology. Second edition.* New York: Appelton-Century-Crofts.

Boring, R.L. (2001). User-interface design principles for experimental control software. In *Conference on human factors in computing systems: CHI '01 extended abstracts* (pp. 399-400). New York: ACM Press.

Boring, R.L. (2002). Human-computer interaction as cognitive science. In *Proceedings of the human factors and ergonomics society 46th annual meeting* (pp. 1767 – 1771). Santa Monica, CA: Human Factors and Ergonomics Society.

Boring, R.L. (2003). Improving human scaling reliability. In *Proceedings of the human factors and ergonomics society 47th annual meeting* (pp. 1820-1824). Santa Monica, CA: Human Factors and Ergonomics Society.

Boring, R.L., & Fernandes, G.J. (2004). Measuring visual appeal of web pages. In *Conference on human factors in computing systems: CHI 2004 extended Abstracts* (p. 1557). New York: ACM Press.

Boring, R.L., Gertman, D.I., & Marble, J.L. (2004). Temporal factors of human error in SPAR-H human reliability analysis modeling. In *Proceedings of the human factors and ergonomics society 48th annual meeting* (in press). Santa Monica, CA: Human Factors and Ergonomics Society.

Boring, R.L., West, R.L., & Moore, S. (2002). Helping users determine video quality of service settings. In *Conference on human factors in computing systems: CHI 2002 extended abstracts* (pp. 598-599). New York: ACM Press.

Bos, J.C. (2000). *A user interface for negotiating QoS for video.* Unpublished master's thesis, Department of Psychology, Carleton University, Ottawa, Canada.

Bower, G.H., & Clapper, J.P. (1989). Experimental methods in cognitive science. In M.I. Posner (Ed.), *Foundations of cognitive science.* Cambridge, MA: MIT Press.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P.W. Jordan, B. Thomas, B.. Weerdmeester, and I.L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189-194). London: Taylor & Francis.

Brysbaert, M. (1990). A warning about millisecond timing in Turbo Pascal. *Behavior Research Methods, Instruments and Computers*, *22*, 344-345.

Brysbaert, M., Bovens, N., D'Ydewalle, G., & Van Calster, J. (1989). Turbo Pascal timing routines for the IBM microcomputer family. *Behavior Research Methods, Instruments and Computers*, *21*, 73-83.

Bührer, M., Sparrer, B., & Weitkunat, R. (1987). Interval timing routines for the IBM PC/XT/AT microcomputer family. *Behavior Research Methods, Instruments and Computers*, *19*, 327-334.

Chalmers, E.J. (1996). *The conscious mind. In search of a fundamental theory.* New York: Oxford University Press.

Chapanis, A., & Halsey, R.M. (1955). Luminance of equally bright colors. *Journal of the Optical Society of America*, *45*, 1-6.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis of the behavioral sciences. Second edition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151-216). London: Academic Press.

Commission Internationale de L'Eclairage. (1931). *CIE chromaticy diagram.* Vienna: Commission Internationale de L'Eclairage.

Commission Internationale de L'Eclairage. (1986). *Colorimetric observers (S002).* Vienna: Commission Internationale de L'Eclairage.

Coren, S., Ward, L.M., & Enns, J.T. (1999). *Sensation and perception. Fifth edition.* New York: HBJ College & School Division.

Creative Labs. (2000). *Sound Blaster Audigy*. Milpitas, CA: Creative Technology Ltd.

Dawkins, R. (1998). *Unweaving the rainbow: Science, delusion and the appetite for wonder*. New York: Houghton Mifflin Co.

Dawson, M.R.W. (1998). *Understanding cognitive science*. Oxford: Blackwell Publishers.

Denzin, N. (1970). Strategies of multiple triangulation. In N. Denzin (Ed.), *The research act in sociology: A theoretical introduction to sociological method* (pp. 297-313). New York: McGraw-Hill.

Dillon, R.F. (1983). Human factors in user-computer interaction: An introduction. *Behavior Research Methods and Instrumentation*, *15*, 195-199.

Dreyfus, H.L. (1992). *What computers still can't do*. Cambridge, MA: MIT Press.

Dudek, C., & Lindgaard, G. (2004). Measuring user satisfaction on the web: Stories people tell. *Design and emotion: The experience of everyday things.* London: Taylor & Francis.

Dumas, J.S. (2001). Usability testing methods: Subjective measures—Measuring attitudes and opinions. In R.J. Branaghan (ed.), *Design by people for people: Essays on usability* (pp. 107-117). Chicago: Usability Professionals' Association.

Dumas, J.S., & Redish, J.C. (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex.

Ellis, B. (1968). *Basic concepts of measurement.* Cambridge, UK: Cambridge University Press.

Estes, W.K. (1994). *Classification and cognition.* New York: Oxford University Press.

Fancher, R.E. (1996). *Pioneers of psychology. Third edition*. New York: W.W. Norton & Company.

Fechner, G.T. (1860a). *Elemente der psychophysik. Teil I.* Leipzig: Breitkopf & Härtel.

Fechner, G.T. (1860a). *Elemente der psychophysik. Teil II.* Leipzig: Breitkopf & Härtel.

Fernandes, G.J. (2003). *Judging web page visual appeal*. Unpublished master's thesis, Department of Psychology, Carleton University, Ottawa, Canada.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*, 381-391.

Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J.A. (1998). *Concepts. Where cognitive science went wrong.* Oxford: Oxford University Press.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Conference on human factors in computing systems: CHI 2000 proceedings* (pp. 345-352). New York: ACM Press.

Gardner, H. (1985). *The mind's new science. A history of the cognitive revolution.* New York: Basic Books.

Gazzaniga, M.S. (1989). Organization of the human brain. *Science*, *245*, 947-952.

Gertman, D.I., & Blackman, H.S. (1984). *Human reliability & safety analysis data handbook*. New York: Wiley Interscience.

Gertman, D., Blackman, H., Marble, J., Byers, J., Haney, L., & Smith, C. (2004). *The SPAR-H human reliability analysis method.* Washington, DC: US Nuclear Regulatory Commission.

Gescheider, G.A. (1997). *Psychophysics: The fundamentals.* Mahwah, NJ: Lawrence Erlbaum.

Goldstein, E.B. (2001). *Sensation and perception. Sixth edition.* Belmont, CA: Wadsworth.

Graves, R., & Bradley, R. (1987). Millisecond interval timer and auditory reaction time programs for the IBM PC. *Behavior Research Methods, Instruments and Computers*, *19,* 30-35.

Graves, R., & Bradley, R. (1988). More on millisecond timing and tachistoscope applications for the IBM PC. *Behavior Research Methods, Instruments and Computers*, 20, 408-412.

Haussmann, R.E. (1992). Tachistoscopic presentation and millisecond timing on the IBM PC/XT/AT and PS/2: A Turbo Pascal unit to provide general-purpose routines for CGA, Hercules, EGA, and VGA monitors. *Behavioral Research Methods, Instruments and Computers*, *24*, 303-310.

Heathcote, A. (1988). Screen control and timing routines for the IBM microcomputer family using a high-level language. *Behavior Research Methods, Instruments and Computers*, *20*, 289-297.

Hellman, R.P., & Meiselman, C.H. (1988). Prediction of individual loudness exponents from cross modality matching. *Journal of Speech & Hearing Research*, *31*, 605-615.

Helmholtz, H. (1887/1959). *Die tatsachen in der wahrnehmung: Zählen und messen erkenntnis-theoretisch brachtet.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Hinden, R.M. (1996). IP next generation overview. *Communications of the ACM*, *39*(6), 61-71.

Hubbard, R., & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology—And its future prospects. *Educational and Psychological Measurement*, *60*, 661-681.

Indow, T., & Stevens, S.S. (1966). Scaling of saturation and hue. *Perception and Psychophysics*, *1*, 253-272.

International Electrotechnical Commission. (2001). *IEC 60651 Ed 1.2. Sound level meters.* Geneva: International Electrotechnical Commission.

International Standards Organisation. (1998). *ISO 9241-11. Egonomics requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability.* Geneva: International Standards Organisation.

Intenational Telecommunications Union. (1996). *ITU-T P.800. Methods for subjective determination of transmission quality.* Geneva: International Telecommunications Union.

Intenational Telecommunications Union. (2000a). *ITU-R BT.500-7. Methodology for the subjective assessment of the quality of television pictures.* Geneva: International Telecommunications Union.

Intenational Telecommunications Union. (2000b). *ITU-T P.920. Interactive test methods for audiovisual communications.* Geneva: International Telecommunications Union.

James, W. (1890). *The principles of psychology.* New York: H. Holt & Company.

Johnson-Laird, P. (1993). *The computer and the mind. An introduction to cognitive science. Second edition.* London: Fontana Press.

Judd, D.B. (1951). Basic correlates of the visual stimulus. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 811-867). New York: John Wiley & Sons.

Karmiloff-Smith, A. (1992). *Beyond modularity. A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Karmiloff-Smith, A. (1999). Modularity of mind. In R.A. Wilson & Frank C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 558-560). Cambridge, MA: MIT Press.

Karvonen, K. (2000). The beauty of simplicity. In *Proceedings of the conference on universal usability* (pp. 85-90). New York: ACM Press.

Kirakowski, J. (1996). The software usability measurement inventory. Background and usage. In P.W. Jordan, B. Thomas, B.A. Weerdmeester, & I.L. McClelland (Eds.), *Usability evaluation in industry.* London: Taylor & Francis.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, *24*, 210-212.

Kline, P. (1998). *The new psychometrics. Science, psychology, and measurement.* London: Routledge.

Krueger, L.E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, *12*, 251-320.

Kurosu, M., & Kahimura, K. (1995). Apparent usability vs. inherent usability: Experimental analysis on the determinants of the apparent usability. In *Conference on human factors in computing systems: CHI 1995 Proceedings* (pp. 292-293). New York: ACM Press.

Kyburg, H.E. (1984). *Theory and measurement*. Cambridge, UK: Cambridge University Press.

Laming, D.R.J. (1997). *The measurement of sensation. Oxford psychology series no. 30*. Oxford: Oxford University Press.

Liberman, A.M., Harris, K.S., Hoffman, H.S., & Griffith, B.C. (1957). The discrimination of speech sounds within and across phonemes boundaries. *Journal of Experimental Psychology*, *18*, 201-212.

Likert, R.A. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 44-53.

Lilienthal, M.G., & Dawson, W.E. (1976). Inverse cross-modality matching: A test of ratio judgment consistency for group and individual data. *Perception and Psychophysics*, *19*, 252-260.

Lindgaard, G. (1994). *Usability testing and system evaluation: A guide for designing useful computer systems.* London: Chapman & Hall.

Lindgaard, G., & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with Computers*, *15*, 429-452.

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, *240*, 740-749.

Lyons, J.C. (2001). Carving the mind at its (not necessarily modular) joints. *British Journal for the Philosophy of Science*, *52*, 277-302.

Luce, D.R., & Mo, S.S. (1965). Magnitude estimation of heaviness and loudness by individual observers: A test of a probabilistic response theory. *British Journal of Mathematical and Statistical Psychology*, *18*, 159-174.

Mackworth, N.H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Psychology*, *1*, 6-21.

Marks, L.E. (1989). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 586-602.

Marks, L.E. (1991). Reliability of magnitude matching. *Perception and Psychophysics*, *49*, 31-37.

Marks, L. E., Galanter, E., & Baird, J. C. (1995). Binaural summation after learning psychophysical functions for loudness. *Perception and Psychophysics*, *57*, 1209-1216.

Marks, L.E., & Stevens, J.C. (1966). Individual brightness functions. *Perception and Psychophysics*, *1*, 17-24.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information.* San Francisco: W.H. Freeman and Company.

Meister, D. (2001). Basic premises and principles of human factors measurement. *Theoretical Issues in Ergonomics*, *2*, 1-22.

Meister, D. (2004). *Conceptual foundations of human factors measurement.* Mahwah, NJ: Lawrence Erlbaum Associates.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355-383.

Microsoft Corporation. (1998). *Microsoft Visual Basic 6.0*. Redmond, WA: Microsoft Corporation.

Microsoft Corporation. (2000). *Microsoft Windows 2000 Professional*. Redmond, WA: Microsoft Corporation.

Microsoft Corporation. (2001). *Microsoft Windows XP Home Edition*. Redmond, WA: Microsoft Corporation.

Middleton, W.E.K. (1966). *A history of the thermometer and its use in meteorology.* Baltimore: Johns Hopkins Press.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Murphy, K. (Producer), & Reiner, R. (Director). (1984). *This is spinal tap* [Motion picture]. United States: Metro Goldwyn Mayer.

Nielsen, J. (1993). *Usability engineering*. Boston: Academic Press Professional.

Norman, D.A. (2004). *Emotional design: Why we love (or hate) everyday things.* New York: Basic Books.

Norwich, K.H. (1987). On the theory of Weber fractions. *Perception and Psychophysics*, *42*, 286-298.

Norwich, K.H. (1993). *Information, sensation and perception*. San Diego: Academic Press.

Norwich, K.H., & Wong, W. (1997). Unification of psychophysical phenomena: The complete form of Fechner's law. *Perception and Psychophysics*, *59*, 929-940.

Paap, K.R., & Noel, R.W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research*, *53*, 13-24.

Panek, D.W., & Stevens, S.S. (1966). Saturation of red: A prothetic continuum. *Perception and Psychophysics*, *1*, 59-66.

Pease, V. (1964). The intensity-time relation of a stimulus in simple visual reaction time. *Psychological Record*, *14*, 157-164.

Penzes, W.B. (2002). Time line for the definition of the meter. Retrieved October 16, 2002, from the National Institute of Standards and Technology Web site: http://www.mel.nist.gov/div821/museum/timeline.htm

Petrusic, W.M., Baranski, J.V., & Kennedy, R. (1998). Similarity comparisons with remembered and perceived magnitudes: Memory psychophysics and fundamental measurement. *Memory and Cognition*, *26*, 1041-1055.

Plainis, S., Chauhan, K., Murray, I.J., & Charman, W. N. (1999). Retinal adaptation under night-time driving conditions. In A.G. Gale, (Ed.), I.D. Brown, C.M. Haslegrave & S.P.Taylor (Co-Eds.), *Vision in Vehicles VII* (pp. 61-70). Amsterdam: Elsevier Science.

Poulton, E.C. (1989). *Bias in quantifying judgments.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Puri, A., & Eleftheriadis, A. (1998). MPEG-4: A multimedia coding standard supporting mobile applications. *ACM Mobile Networks and Applications Journal*, *3*(1), 5-32.

Pylyshyn, Z.W. (1984). *Computation and cognition. Toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

Quinlan, P. (1991). *Connectionism and psychology. A psychological perspective on new connectionist research.* Hemel Hempstead, UK: Harvester Wheatsheaf.

Reason, J. (1990). *Human error.* Cambridge, UK: Cambridge University Press.

Rubin, J. (1994). *Handbook of usability testing*. New York: John Wiley & Sons.

Savage, C.W. (1970). *The measurement of sensation*. Berkeley: University of California Press.

Savage, C.W., & Ehrlich, P. (1992). A brief introduction to measurement theory and to the essays. In C.W. Savage & P. Ehrlich (Eds.), *Philosophical and foundational issues in measurement theory* (pp. 1-14). Hillsdale, NJ: Lawrence Erlbaum.

Schulz, D., & Schultz, S. E. (1999). *A modern history of psychology. 7th edition.* New York: Harcourt.

Segalowitz, S.J., & Graves, R.E. (1990). Suitability of the IBM XT, AT, and PS/2 keyboard, mouse, and game port as response devices in reaction time paradigms. *Behavior Research Methods, Instruments and Computers*, *22*, 283-289.

Shiffrin, R.H., & Nosofsky, R.M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, *101*, 357-361.

Simpson, R. (1944). The specific meanings of certain terms indicating differing degrees of frequency. *Quarterly Journal of Speech*, *30*, 328-330.

Smith, B. (1988). Gestalt theory: An essay in philosophy. In B. Smith (Ed.), *Foundations of gestalt theory* (pp. 11-81). Munich: Philosophia.

Sorenson Vision, Inc. (1997). *Sorenson technology white paper*. Salt Lake City: Sorenson Vision, Inc.

Stevens, J.C., & Marks, L.E. (1980). Cross-modality matching functions generated by magnitude estimation. *Perception and Psychophysics*, *27*, 379-389.

Stevens, J.C., & Stevens, S.S. (1960). Warmth and cold: Dynamics of sensory adaptation. *Journal of Experimental Psychology*, *60*, 183-192.

Stevens, J.C., & Hall, J.W. (1966). Brightness and loudness as functions of stimulus duration. *Perception and Psychophysics*, *1*, 319-327.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.

Stevens, S.S. (1966). Duration, luminance, and the brightness exponent. *Perception and Psychophysics*, *1*, 96-100.

Stevens, S.S. (1975). *Psychophysics. Introduction to its perceptual, neural, and social prospects.* New York: Wiley.

Stevens, S.S., & Galanter, E.H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*, 377-411.

Stevens, S.S., & Guirao, M. (1964). Subjective scaling of length and area and the matching of length to loudness and brightness. *Journal of Experimental Psychology*, *66*, 177-186.

Syntrillium Corporation. (2000). *Cool Edit 2000*. Phoenix: Syntrillium Corporation.

Teghtsoonian, R. (1971). On the exponents in Steven's law and the constant in Ekman's law. *Psychological Review*, *78*, 71-80.

Teghtsoonian, R. (1989). The study of individuals in psychophysical measurement. In G. Ljunggren & S. Dornic (Eds.), *Psychophysics in action* (pp. 95-105). Berlin: Springer Verlag.

Teghtsoonian, M., & Teghtsoonian, R. (1983). Consistency of individual exponents in cross-modal matching. *Perception and Psychophysics*, *33*, 203-214.

Teghtsoonian, R., & Teghtsoonian, M. (1997). Range of acceptable stimulus intensities: An estimator of dynamic range for intensive perceptual continua. *Perception and Psychophysics*, *59*, 721-728.

Titchener, E.B. (1911). *A text-book of psychology.* New York: Macmillan.

Thurston, L.L. (1919). A method for scoring tests. *Psychological Bulletin*, *16*, 235-240.

Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. In *Conference on human factors in computing systems: CHI 1997 proceedings* (pp. 115-122). New York: ACM Press.

Tractinsky, N., Katz, A.S., and Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*, 127-145.

Tullis, T.S. (1997). Screen design. In M.G. Helander, T.K. Landauer, and P.V. Prabhu (Eds.), *Handbook of human-computer interaction. Second, completely revised edition* (pp. 503-531). Amsterdam: Elsevier Science.

Ward, L.M. (1974). Power functions for category judgments of duration and line length. *Perceptual and Motor Skills*, *38*, 1182.

Ward, L.M. (1975). Sequential dependencies and response range in cross-modality matches of duration in loudness. *Perception and Psychophysics*, *18*, 217-223.

Ward, L.M. (1982). Mixed-modality psychophysical scaling: Sequential dependencies and other properties. *Perception and Psychophysics*, *31*, 53-62.

Ward, L.M. (1990). Critical bands and mixed-frequency scaling: Sequential dependencies, equal-loudness contours, and power function exponents. *Perception and Psychophysics*, *47*, 551-562.

Ward, L.M. (1992). Who knows? In G. Borg & G. Neely (Eds.), *Fechner day 92. Proceedings of the eighth annual meeting of the international society for psychophysics* (pp. 217-222). Stockholm, Sweden: International Society for Psychophysics.

Ward, L.M. (2002). *Dynamical cognitive science*. Cambridge, MA: MIT Press.

Watson, A., & Sasse, M.A. (1998). Measuring perceived quality of speech and video in multimedia conferencing applications. In *Proceedings of the 6th ACM international conference on multimedia* (pp. 55-60). New York: ACM Press.

West, R.L. (1996). Constrained scaling: Models, methods, and response bias. In S.C. Masin (Ed.), *Fechner day 96: Proceedings of the twelfth annual meeting of the international society for psychophysics* (pp. 423-427). Padua, Italy: International Society for Psychophysics.

West, R.L., Boring, R.L., Dillon, R.F., & Bos, J. (2001). Human computer interaction, quality of service, and multimedia internet broadcasting. In A. Sloane and D. Lawrence (Eds.), *Multimedia internet broadcasting. Quality, technology and interface* (pp. 1-15). London: Springer Verlag.

West, R.L., Boring, R.L., & Moore, S. (2002). Computer augmented psychophysical scaling. In *Proceedings of the twenty-fourth annual meeting of the cognitive science society* (pp. 932-937). Mahwah, NJ: Lawrence Erlbaum Associates.

West, R.L., & Ward, L.M. (1994). Constrained scaling. In L.M. Ward (Ed.), *Fechner day 94. Proceedings of the tenth annual meeting of the international society for psychophysics* (pp. 225-230). Vancouver, Canada: International Society for Psychophysics.

West, R.L., & Ward, L.M. (1998). The value of money: Constrained scaling and individual differences. In S. Grondin & Y. Lacouture (Eds.), *Fechner day 98. Proceedings of the fourteenth annual meeting of the international society for psychophysics* (pp. 377-380). Quebec City: International Society for Psychophysics.

West, R.L., Ward, L.M., & Khosla, R. (2000). Constrained scaling: The effect of learned psychophysical scales on idiosyncratic response bias. *Perception and Psychophysics*, *62*, 137-151.

Wilson, G. (2000). Multimedia conferencing: What cost to users? In A. Pras (Ed.), *Proceedings of sixth European network of universities and companies in information communication engineering open European summer school: Innovative internet applications* (pp. 173-180). Enschede: University of Twente.

Wilson, G.M., & Sasse, M.A. (2000) Listen to your heart: Counting the cost of media quality. In A. Paivia (Ed.), *Affective interactions—Towards a new generation of computer interfaces* (pp. 9-20). Heidelberg: Springer Verlag.

Veryzer, R.W. (1999). A nonconscious process explanation of consumer response to product design. *Psychology and Marketing*, *16*, 497-522.

Zipser, D. (1986). Biologically plausible models of place recognition and goal location. In J.L. McClelland, D.E. Rummelhart, & The PDP Research Group (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 432-470). Cambridge, MA: MIT Press.

Zwislocki, J.J. (1983). Group and individual relations between sensation magnitudes and their numerical estimates. *Perception and Psychophysics*, *33*, 460-468.

APPENDIX A

EXPERIMENTAL CONTROL SOFTWARE

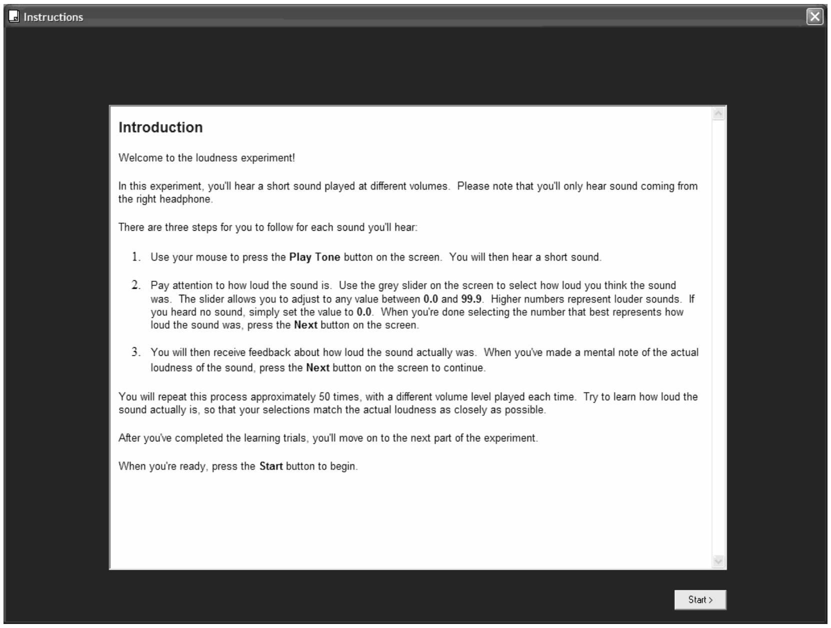
Programming and Interface Description

The experimental control software (ECS) was developed using Microsoft Visual Basic Version 6 (Microsoft Corporation, 1998) with the most current service packs installed. The software was developed in the Windows 2000 operating system platform (Microsoft Corporation, 2000) and executed under a mixture of Windows 2000 and Windows XP (Microsoft Corporation, 2001) platforms. The ECS for the 15 experiments averaged 1700 lines of Visual Basic code split across three independent code modules. While each experiment required slightly different code, there was significant underlying similarity across the experiments, and much of the programming code was reused from one experiment to the next.

Note that in the interest of document parsimony, the code for the individual experiments is not included in this dissertation. The complete code for all 15 experiments, including the nuances that differentiated the individual experiments, requires approximately 425 single-spaced pages to reproduce. The functioning of the ECS is described in detail in this appendix, and suitable screenshots are provided to demonstrate the characteristics of the ECS.

The three ECS code modules included the splash screen, the instructions page, and the main portion of the experiment. The splash screen appeared upon starting the experiment (see Figure A-1). The primary purpose of the splash screen was to verify the

Figure A-1. A sample splash screen used in the experimental control software.



Introduction

Welcome to the loudness experiment!

In this experiment, you'll hear a short sound played at different volumes. Please note that you'll only hear sound coming from the right headphone.

There are three steps for you to follow for each sound you'll hear:

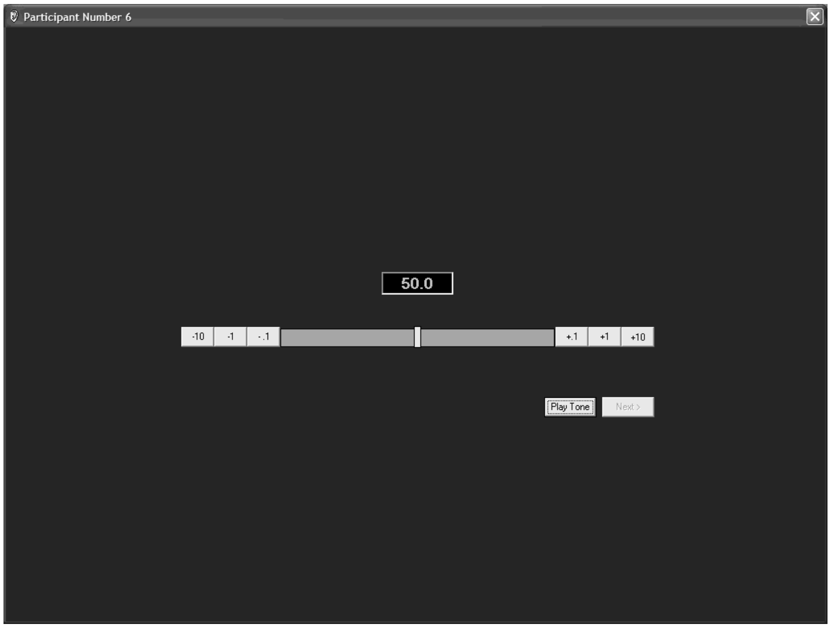
1. Use your mouse to press the **Play Tone** button on the screen. You will then hear a short sound.
2. Pay attention to how loud the sound is. Use the grey slider on the screen to select how loud you think the sound was. The slider allows you to adjust to any value between **0.0** and **99.9**. Higher numbers represent louder sounds. If you heard no sound, simply set the value to **0.0**. When you're done selecting the number that best represents how loud the sound was, press the **Next** button on the screen.
3. You will then receive feedback about how loud the sound actually was. When you've made a mental note of the actual loudness of the sound, press the **Next** button on the screen to continue.

You will repeat this process approximately 50 times, with a different volume level played each time. Try to learn how loud the sound actually is, so that your selections match the actual loudness as closely as possible.

After you've completed the learning trials, you'll move on to the next part of the experiment.

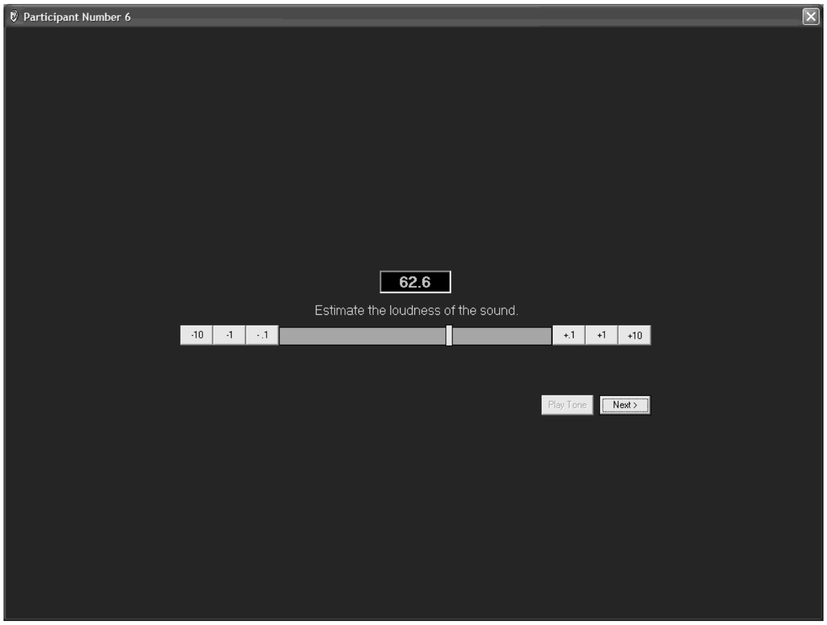
When you're ready, press the **Start** button to begin.

Figure A-2. A sample screenshot produced by the instruction module in the experimental control software.



50.0

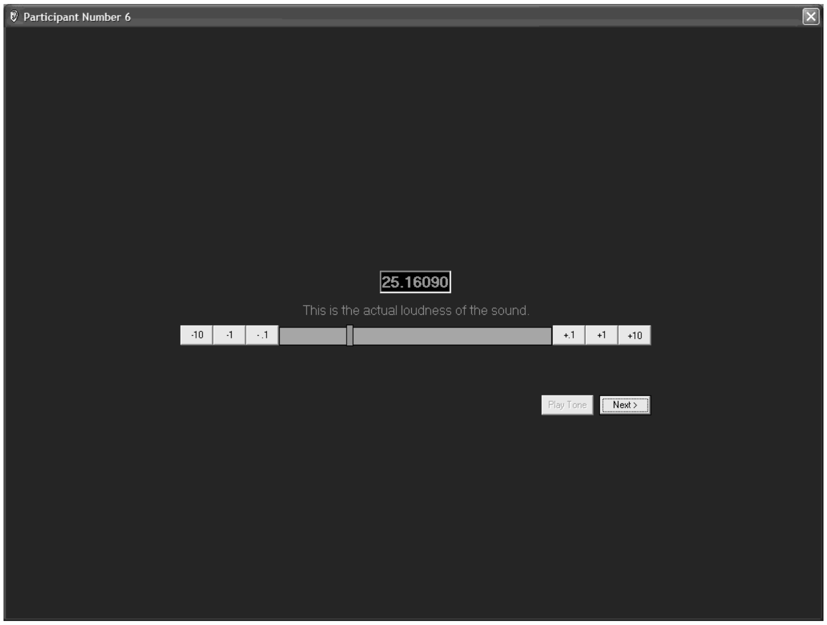
Figure A-3. Sample screen from Experiment 1, in which the participant initiates the stimulus presentation.



62.6

Estimate the loudness of the sound.

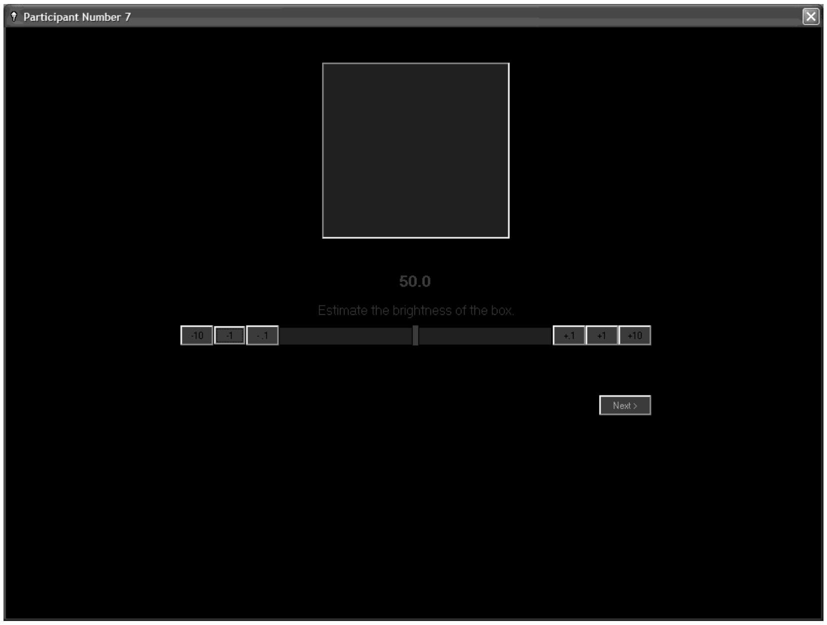
Figure A-4. Sample screen from Experiment 1, in which the participant rates the intensity of the stimulus presentation.



25.16090

This is the actual loudness of the sound.

Figure A-5. Sample screen from Experiment 1, in which the participant receives feedback about the actual intensity of the stimulus.



50.0

Estimate the brightness of the box.

Figure A-6. Sample screen from Experiment 3, in which the participant rated the brightness of a box displayed briefly on the screen.

experiment name for the experimenter. While the splash screen displayed, the ECS loaded stimulus and calibration files while randomizing the stimulus presentation order. During the splash screen, the ECS automatically incremented the participant number and, in some cases, assigned the experimental condition based on the participant number. For example, in Experiments 11 and 13, the odd-numbered participants received the low coefficient stimuli for the first phase of the experiment, while the even-numbered participants received the high coefficient stimuli. The splash screen displayed for ten seconds, after which the ECS proceeded to display instructions.

The instruction module provided an opportunity to display instructions to the participant (see Figure A-2). The instruction module featured an imbedded Web browser capable of displaying any document encoded in HyperText Markup Language (HTML). The use of HTML provided a flexible way to display formatted text instructions and to change the content of the instructions easily from one experiment to another. The instruction module was activated initially to provide the participant with overview instructions for the experiment and was typically activated several times subsequently, whenever the ECS needed to provide additional instructions. A similar textual display was provided at the conclusion of the experiment to debrief the participants on the experiment. When the instructions exceeded the screen's display capacity, the participant was able to scroll down the page for additional text using the scroll bars at the right of the window. When the participant had finished reading the instructions, he or she pressed the "Start" button at the bottom right of the display.

The third and final module of the ECS code contained the actual experiment. This portion of the ECS varied considerably, depending on whether an experiment called for magnitude estimation or constrained scaling, which type of stimulus was used (i.e., loudness of sounds, brightness of squares, value of money, visual appeal of Web pages, or fluidity of video movement), and whether or not a trial included feedback. Figures A-3 through A-5 illustrate the output of the ECS for Experiment 1 during the training trials. Figure A-3 shows the initial screen, in which the participant must press the "Play Tone" button on the screen to hear the sound. The display is deliberately kept sparse, to avoid visual clutter (Tullis, 1997) that might distract the participant from the experimental task. While the onscreen buttons and slider were presented in shades of grey and white, the display background was set to a dark red color in order to disambiguate background and foreground elements in the display. In the center of the display was a custom designed slider control, which was dubbed the IntelliSlider.

The IntelliSlider functions much like a conventional Windows slider, with two exceptions. First, a conventional Windows slider has a single button at each end of the slider bar. These buttons allow the user to decrement or increment the slider position (and value) by a pre-specified amount. Based on previous experience (West, 1996), users prefer to have a finer control over the slider than is present in a conventional slider. Hence, the IntelliSlider features three levels of decrement and increment. The three buttons at either end of the slider bar allow the user to change the value down or up by 0.1, 1, and 10 units, respectively. The user can also click at any point on the slider bar to decrease or increase the position indicator (and value), or the user can click and drag the slider position indicator to a new position. The value is signified by both the position of the slider indicator and by the synchronized numeric value presented in the small window located on the display directly above the slider. The second unique feature of the IntelliSlider is that it scales to an approximated 101-point scale. The scale ranges from 0.1 to 99.9, and scale values are adjusted to this scale regardless of the stimulus modality. Thus, whereas in the loudness experiments, the scale is calibrated to an underlying dB scale, in the brightness experiments, the scale represents an underlying scale in cd/m^2.

In order to maintain display simplicity, no help instructions were initially provided on the experimental screen. However, to assist the participant, brief instructions were flashed on the display after six seconds of inactivity. The instructions remained in place for another six seconds. The instructions were repeated following another six seconds if the participant had not selected one of the appropriate onscreen buttons. In this manner, the participant benefited from additional but unobtrusive usage cues.

After the stimulus presentation in Figure A-3, the participant was presented with a screen in which he or she was instructed to rate the intensity of the previously presented stimulus (see Figure A-4). The participant used the IntelliSlider to set the value and then pressed the “Next” button on the screen. The IntelliSlider was always positioned at the midpoint value of 50.0. As in the previous screen, the ECS flashed brief instructions to assist the user after six seconds.

Following selection of the stimulus intensity value in screen A-4, the participant would either advance to the next stimulus presentation or receive feedback in accordance with constrained scaling methodology. In a typical feedback screen (see Figure A-5), the

ECS displayed the actual stimulus value according to the IntelliSlider scale. The value was displayed positionally on the IntelliSlider as well as numerically in the display above the IntelliSlider. To differentiate between data entry mode and feedback mode, the position indicator on the IntelliSlider as well as the numerical display were colored light green. The participant acknowledged seeing the actual stimulus intensity value by clicking on the "Next" button on the screen, which advanced the participant to the next trial or next portion of the experiment.

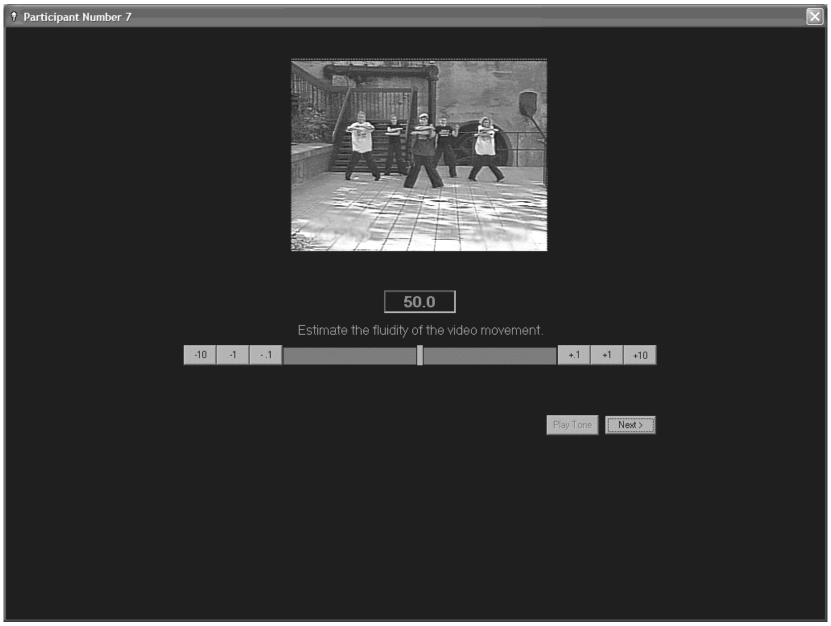
As mentioned earlier, the main experimental interface varied depending on the type of experiment. For loudness experiments, the stimulus was presented audibly, with no visible stimulus presentation element. Most of the experiments featured an on-screen stimulus display element. For example, experiments involving grayscale featured the display found in Figure A-6. Since the scaling in question involved the brightness of an object displayed on the screen, the colors of the display were darkened. The background was black, while on-screen objects such as the slider were set to varying levels of grey. The stimulus was displayed directly above the IntelliSlider and value indicator box. Except for in Experiment 3 and 4, this square of varying intensities was flashed on the display for 1000 ms.

The square was framed by a lighter colored border. When the box was not illuminated with a greyscale box, it matched the black background of the display window. To reflect the nature of the scaling task, the on-screen instructions reminded the participant to estimate the brightness of the on-screen box.

Similar customizations were implemented to accommodate the display and scaling requirements for the full range of the experiments. For example, for the color brightness scaling experiments, the on-screen box displayed red, green, and blue colored squares in addition to the grayscale squares. For the subjective utility of money experiments, the display featured a sum of money above the IntelliSlider. The IntelliSlider, in turn, served as a rating scale for subjective happiness. In the experiment to determine the visual appeal of Web pages, the individual Web pages were flashed at the full dimensions of the screen, after which the ECS displayed the standard IntelliSlider display with a rating scale for judging the visual appeal of the Web page. Finally, in the experiments involving video frame rate, the ECS featured a box, similar to the grayscale box in Figure A-6, in which a 320x240 pixel video was displayed (see Figure A-7).

Timing Accuracy

Several reaction time measures were obtained throughout the experiments. However, obtaining accurate reaction time measures in Microsoft Windows operating system environments has been a problematic programming endeavor. As a preemptive multitasking operating system, the Windows family allots slices of time to each concurrent process. The result is that, when chained together, these individual time slices give the appearance of simultaneous processing of multiple processes. From a timing standpoint, time slicing introduces considerable variability into processing time, meaning that applications that are time sensitive are best implemented in an operating system environment that does not utilize preemptive multitasking.



Participant Number 7

50.0

Estimate the fluidity of the video movement.

Figure A-7. Sample screen from Experiment 15, in which the participant rated the fluidity of video displayed on the screen.

Factors beyond the operating system also significantly hinder the timing accuracy of standard personal computers. Highly accurate timing presents an obstacle for the hardware architecture established by the IBM PC in 1981 and compatibly maintained by subsequent generations of personal computers that are based on the Intel 80x86 chipset, including current generation Pentium processors. Although this architecture includes a highly accurate timer chip (the Intel 825x series of chips) integral to every system board, its purpose is primarily to maintain the time-of-day status for the computer. The 825x chip is invariably set at a frequency of 18.2 Hz, which gives a timer resolution equivalent to 55 ms. Although this resolution is more than sufficient to maintain the time-of-day clock (Bührer, Sparrer, & Weitkunat, 1987), such a resolution is nonetheless too slow for the exacting millisecond accurate reaction time experiments common in psychology. Given this hardware constraint, the challenge is whether or not this single timer may be harnessed in a way so as to achieve higher accuracy without the need for additional timer hardware. Two solutions to this challenge have emerged.

The first solution is to increase the frequency of the timer. The frequency of the 825x timer may be increased via software control to 1000.1522Hz, which provides accuracy around 1 ms (Brysbaert, Bovens, d'Ydewalle, & Van Calster, 1989). This solution at first seems ideal, but closer consideration reveals some limitations of this approach. Fine-tuning the frequency of the timer chip results in significantly slowed overall system performance, because the hardware timer interrupt is issued 60 times more frequently than normally. This ultimately results in the operating system and main program being able to perform their operations 60 fewer times per second. Especially in

Windows-based programs, the dependence on timing routines is critical to overall system functioning. Upsetting the frequency of Windows time slices can easily halt system operation. Furthermore, by manipulating the 825x default parameters, the time-of-day routines are rendered inoperable to normal system operation, since the system advances the internal time of day 60 times faster than normal.

The second and typically preferred solution maintains the frequency of the timer. Extended timer resolution is possible by setting the 825x timer chip mode to maintain an internal residual time count between two timer events. Graves and Bradley (1987, 1988) first devised assembly language routines that read the time-of-day counter and the residual counter to achieve millisecond accuracy. By this method it is possible to derive the elapsed time since midnight in milliseconds.

Bedell (2000) offers an alternative method to derive accurate timing measures. Since most Windows based personal computers are equipped with a multimedia player, and since these players require millisecond accuracy timing for the playback of Musical Instrument Digital Interface (MIDI) files, it is possible to access the multimedia timer to achieve accurate timing. Bedell presents a series of routines in Visual Basic that can be used to derive elapsed time or to trigger events at frequency intervals faster than that otherwise achieved in Windows.

The routines in Bedell (2000) served as the basis for the timing code in the ECS. While specific routines recorded the elapsed time since a starting point (i.e., the time at the presentation of a stimulus), other routines were used to set the trigger point for the presentation of on-screen help, and still other routines were used to add a delay function to the ECS. This delay function was used for timing the duration of the initial splash screen as well as the duration of stimulus presentations.

It should be noted that the reaction time measures in this dissertation do not purport true millisecond accuracy. Despite efforts to control for timing accuracy, several possible confounds exist. Considerably timing latency was introduced due to the input devices. The typical personal computer keyboard, for example, has a built-in delay of between 10 ms (Segalowitz & Graves, 1990) and 36.7 ms (Brybaert, 1990). Moreover, the circuitry of keyboards is highly variable, resulting in different delays in different computers. Because of the large number of keys on a keyboard, the delay is also variable depending on the particular key pressed. Nonetheless, mean values do exist, and most PCs scan the keyboard for a keypress about every 15 ms, with a variance up to ±7.5ms (Segalowitz et al., 1990).

In contrast to the keyboard, the response time of the serial PC mouse has been shown to be highly invariable (Segalowitz et al., 1990; Beringer, 1992). The serial PC mouse transmits information consistently to the PC at a rate of 1200 baud or greater. There is a consistent 31 to 33 ms response delay between the actual press of the mouse button and the issuance of a hardware interrupt. Thus, the actual reaction time may be calculated by subtracting 32 ms from the recorded reaction time. Similar characteristics can be expected of mice designed to work with the Universal Serial Bus (USB). Although the serial PC mouse affords great timing accuracy, a dominant mouse design in personal computers is the PS/2 mouse, which connects directly to the keyboard port. As expected, the PS/2 mouse shows the same timing inconsistencies as the keyboard (Beringer, 1992).

The PC mouse, whether serial, USB, or PS/2, transmits information in variable blocks. If there is only movement or only a button press, the mouse does not need to transmit a very large block of data to the PC in order to establish the change in its status. If, however, there are mouse movements and button presses, the block of information to be transmitted is increased substantially, thereby increasing the reaction time proportionately (Segalowitz et al., 1990; Beringer, 1992). Thus, the overlap of mouse movement and button pressing as required to use the IntelliSlider results in considerable variability in the amount of information and the time to transmit that information.

Another issue limiting the timing accuracy of the ECS stems from the video screen refresh cycle. Achieving millisecond accuracy in the system time or the response input apparatus is of no use if the screen output is not synchronized with these millisecond counters. Screens are not refreshed every millisecond. Taking a typical monitor refresh rate of 70 Hz, the display is updated once every 14.3 ms (Segalowitz et al., 1990). Hence, it is possible that on-screen stimulus information may be written to video memory and the ECS may start the timer during the refresh cycle, resulting in up to a 14.3 ms delay between the activation of the timer and the time when the experimental participant actually sees the stimulus. If that potential delay is not considered, the overall average variability of an experiment is increased by 14.3 ms for each stimulus presentation. This problem is further compounded if the stimulus takes longer than a single refresh cycle to display, since then it will not be displayed in its entirety and the participant may register only part of the image. This would increase the overall variability of the experiment by 28.6 ms or more for each stimulus presentation (Haussmann, 1992).

The solution to this problem is to display information as quickly as possible during the screen refresh cycle (Heathcote, 1988; Haussmann, 1992). The memory address at segment 64 (hexadecimal 0040) and offset 99 (hexadecimal 0063) contains the address of the 6845 video chip commonly found in video adapters and maintained for backward compatibility in most PC video systems. Six bytes from the address provided therein is the byte that contains the video display vertical retrace status, located on bit 3. If the bit is on, then the screen is in vertical retrace. During vertical retrace, information may be safely written to the screen, because no new information will be displayed until the retrace is complete (Haussman, 1992).

Windows display programming is complex, because the amount of information to be displayed graphically is much greater than in DOS text mode. The short duration of the vertical retrace makes it difficult to transfer large amounts of information to the screen memory. Further, since the standard Windows graphical interface libraries do not control for screen refresh, the programmer must code custom graphical controls for all on-screen functions. Alternately, it is possible for the programmer to use an extended graphics library like Microsoft's DirectDraw. The use of such a library adds considerable complexity to the coding required to perform tasks in the ECS.

In the present ECS, screen refresh rates were not controlled for. An attempt was made to generate stimuli in the computer's random access memory (RAM), from which it could be transferred rapidly to video memory. This process eliminated the generation of visual stimuli directly in video memory, where they would likely be generated over one or more screen retrace periods. The transfer of stimulus information from system RAM to video RAM did not, however, guarantee that stimuli would be displayed within a single screen retrace. The accuracy of the reaction time measures should be viewed in light of this potential display latency.

Usability and Design Considerations

Boring (2001, p. 400) offers six guidelines for designing experimental control software (ECS). These are quoted below:

- *Follow standard user-interface design principles when designing for ECS.* Computer controlled psychology experiments should be viewed as simplified interfaces that are subject to the same need for usability as would be more complex interfaces.
- *Design the ECS to support both the experimental participant and the experimenter.* Make it easy for the participant to take part in the experiment, but don't forget to make it easy for the experimenter to administer the experiment.
- *Automate tasks as much as possible.* Currently, much ECS does not fully utilize automation and is thereby error prone. When feasible, provide on-screen instructions to the experimental participants; automate the numbering of experimental participants; assign participants to experimental conditions automatically.
- *Make the data appropriate to the analysis tools.* Formatting the output of the ECS to match the standard input of the analysis software can save time and reduce data handling errors.
- *Simplify the mode of response.* Participants in experiments will perform best if the mode of response is intuitive and natural. As the mode of response becomes less natural, the artificiality of the results increases.
- *Give feedback.* Creating situations in which there is little, no, or punitive feedback establishes a barrier between the participant and the ECS. This interaction barrier can impact the experimental findings.

The ECS used in the experiments in this dissertation was designed specifically to adhere to these guidelines. The following considerations influenced the design of the ECS.

Standard Microsoft Windows user-interface conventions were used throughout the software. The IntelliSlider control, for example, followed closely the conventions of the standard Windows slider, in that software users could click on the appropriate arrows to decrement or increment the slider value, click on an area of the slider to change the slider value, or click and drag the slider to a new position. Other Windows conventions included the use of a software wizard format to guide the software user through successive steps in the experiment. Special attention was paid to using standard button labels such as “Next” and “Ok.”

The ECS supported both the participant and the experimenter. The ECS was designed to include a minimum of items on the screen, so that the on-screen interface elements afforded simple and natural interaction. To this end, the experimental windows were displayed such as to fill the entire screen. On-screen help was provided to assist the participant during the experimental task. The experiment was started easily by double-clicking on an icon on the screen. The ECS splash screen clearly identified the experiment to the experimenter.

Many tasks were automated to minimize the change of experimenter error. Participant numbering and the selection of experimental conditions was handled automatically by the ECS. The instructions were incorporated into the ECS as was the experimental debriefing after the experiment.

The data were collected automatically throughout the experiment and were recorded in a simple columnar format that was compatible with all statistical software packages that were used to analyze the data.

A simple mode of response was offered in the form of the IntelliSlider. The IntelliSlider functioned identically to existing Windows slider controls with the exception that the IntelliSlider offered additional functionality to simply the scaling task for participants.

Three types of feedback were implemented in the ECS. Instructions were provided at the beginning of experimental sessions and at periodic junctures throughout the experiment. Feedback was also provided in the form of unobtrusive help messages that were displayed on the screen after a period of inactivity. These help messages guided the participant on the proper course of action. Finally, in the constrained scaling conditions, feedback was provided while participants learned the scale. This feedback, as was demonstrated throughout this dissertation, provided not only a usability advantage but also allowed participants to calibrate their scale usage in a manner that could be generalized to new stimulus presentations.

These design considerations helped mitigate usability issues that might otherwise have resulted in experimenter or participant error. These design considerations also minimized the possibility of experimental artifacts and confounds that might have arisen if interface characteristics of the ECS had detracted from the basic scaling task being investigated in this series of experiments.

APPENDIX B

LOUDNESS CALIBRATION

Pure tone sinusoidal waveform files were generated as stimuli for the loudness scaling experiments using Cool Edit 2000 software (Syntrillium Corporation, 2000). The files consisted of compact disk quality 16-bit samples at a sampling rate of 44.1 kHz monaurally over the right sound channel. Sound files were generated for 65 and 1000 Hz tones with amplitude settings in 1 dB increments from 50 to 99 dB SPL and in 0.1 dB increments from 99 to 100 dB SPL. The tones were 1 s in duration with 6 ms ramp-up and ramp-down times.

Figure B-1 illustrates the configuration for calibrating the loudness of the stimuli for use in the loudness experiments. The amplitude of the playback tones was calibrated using a Sper Scientific IEC 60651 Type 1 certified sound level meter (International Electrotechnical Commision, 2001) fast time-weighted to the "A" frequency range[74] [see (i) in Figure B-1], a custom acoustic coupler [see (ii) in Figure B-1], sealed circumaural headphones by Sennheiser [see (iii) in Figure B-1], and a sound insulating container [see (iv) in Figure B-1]. The custom acoustic coupler was manufactured out of ceramic modeling clay and plaster, with a felt-lined opening for securely inserting the sound level meter's microphone. The acoustic coupler featured a narrow opening from the headphone earpiece to the sound level meter's microphone, simulating the human

[74] The IEC 60651 Type 1 designation specifies that the sound level meter has a measurement accuracy of ± 1 dB. Time-weighting refers to the sound frequency sample rate, whereby the "fast" setting results in a fast measurement response. Frequency weighting refers to the sensitivity of the sound level meter to sound intensity across the frequency spectrum, whereby the "A" weighting approximates human hearing.

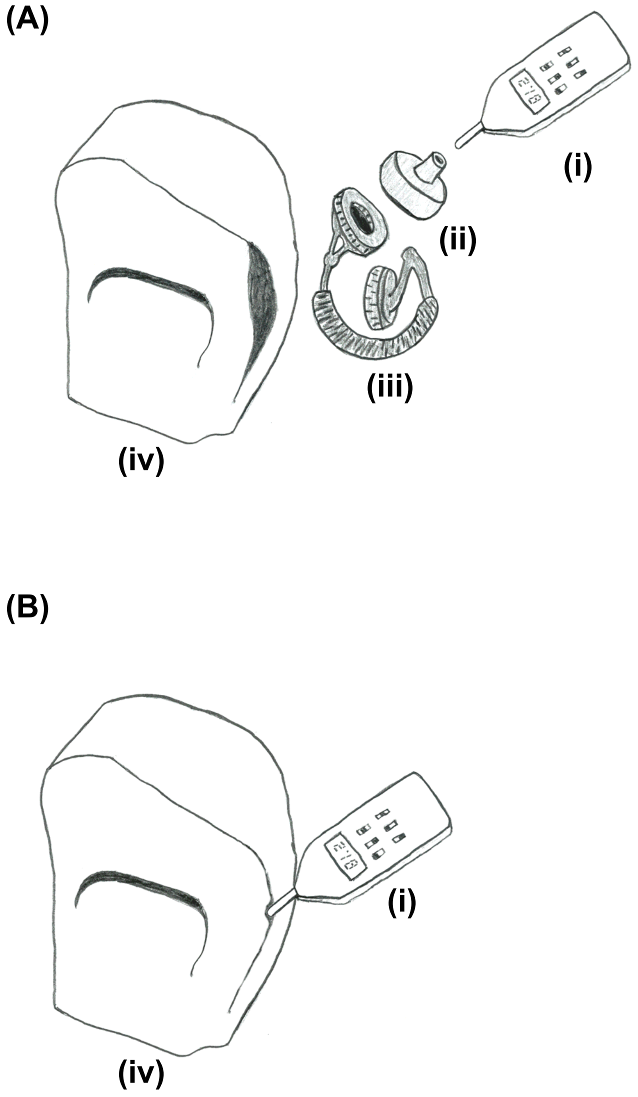


Figure B-1. Configuration for calibrating the loudness experiment stimuli.

auditory canal. The sound insulating container consisted of a canvas backpack filled with a heavy, 800 fill down coat. When the headphones and acoustic coupler were wrapped in the down coat and placed inside the backpack (see Part B of Figure B-1), the backpack could be zipped closed while allowing a sufficient aperture for the stem of sound level meter. The sound insulating container offered approximately 40 dB sound attenuation while making it possible to read the display of the sound level meter. In this manner, it was possible to calibrate the loudness stimuli in an environment without special sound dampening.

The sound files were played through a Sound Blaster Audigy (Creative Labs, 2000) sound card, providing a sound-to-noise ratio of 100 dB. The Sound Blaster Audigy, like preceding Sound Blaster sound cards, offers three volume controls for sound playback within Windows operating systems. The sound card features an overall volume control called *main volume* with a resolution of 16 bits, or 65536 levels. The main volume is fed by mixer input from multiple sound channels, including the right and left volume channels for waveform file playback called *wave volume*. The wave volume similarly features a resolution of 16 bits. Within wave volume, the waveform file is characterized by sound samples of varying amplitudes. The current sound stimulus files featured 16-bit samples, corresponding to a volume resolution of 65536 levels.

By combining the three independent volume levels, it was possible to adjust the output sound level of the sound card with a resolution greater than the measuring accuracy offered by the sound level meter. Main and wave volume levels have a direct relationship to the sound output volume in dB. Using the calibration method depicted in

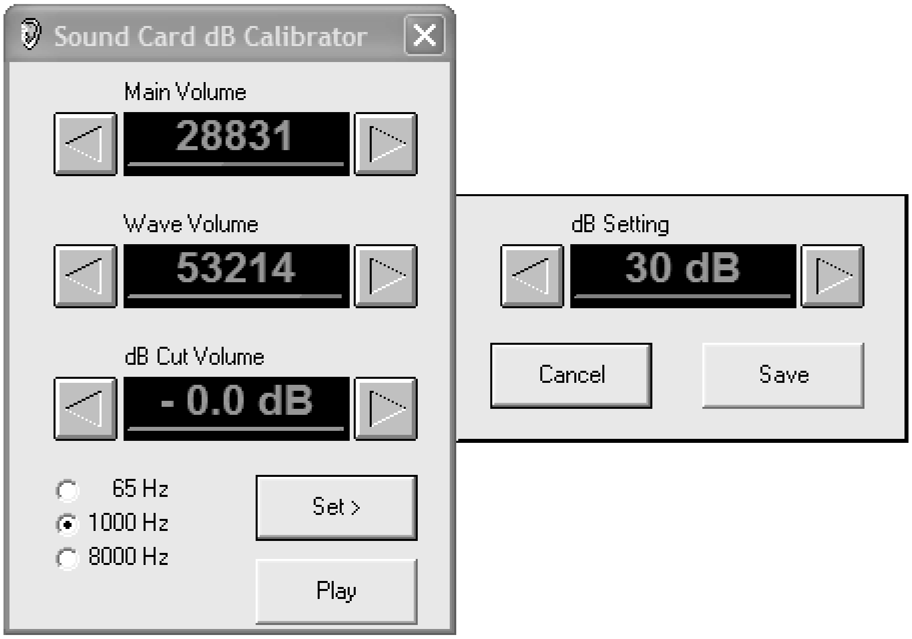
Figure B-1, the volume level in dB was found to be a function of the main or wave volume level setting (V) according to the following equation:[75]

$$\begin{aligned} \text{dB} &= \frac{V}{2184} - 1 \quad \text{for } V \leq 32768 \\ &= \frac{V}{990} - 1 \quad \text{for } V > 32768 \end{aligned} \tag{45}$$

When tested across a variety of system configurations, it was found that in some cases the main and wave volume output levels followed a step function for Equation 45, while in other cases the main and wave volume levels followed a linear function. However, in no configurations did the waveform file volume levels follow a step function. Therefore, some configurations required the inclusion of waveform file volume levels in order to make output adjustments smaller than whole dB steps.

A program was created in Microsoft Visual Basic 6 (Microsoft Corporation, 1998) to create sound calibration lookup tables for use by the experimental control software. The program permitted the experimenter to set the volume levels of the main volume, the wave volume, and the sound file volume samples and to peg those values to a specific dB setting. The settings were, in turn, stored in a data file that was used by the experimental control software when playing sounds. Figure B-2 illustrates the interface for making the sound level adjustments. The three volume levels could be set, the tone played through the headphones, the volume level read from the sound level meter, and this value set and saved in the interface. Table B-1 shows the three volume settings for

[75] I thank Matthew Rutledge-Taylor of Carleton University's Institute of Cognitive Science for identifying this relationship.



Sound Card dB Calibrator

Main Volume: 28831

Wave Volume: 53214

dB Cut Volume: - 0.0 dB

○ 65 Hz
● 1000 Hz
○ 8000 Hz

dB Setting: 30 dB

Figure B-2. The sound card calibration software used for loudness experiments.

each dB level used in the experiment for 65 Hz tones. Table B-2 shows the corresponding settings for 1000 Hz tones.

Table B-1. Loudness calibration values for 65 Hz tones.

Actual Loudness	Master Volume	Wave Volume	File Volume
30dB	24358	30738	45
31dB	24358	30738	44
32dB	24358	30738	43
33dB	24358	30738	42
34dB	24358	30738	40
35dB	24358	30738	38
36dB	24358	30738	36
37dB	24358	30738	35
38dB	24358	30738	33
39dB	24358	30738	32
40dB	24358	30738	31
41dB	24358	30738	29
42dB	24358	30738	28
43dB	24358	30738	27
44dB	24358	30738	26
45dB	24358	30738	25
46dB	24358	30738	24
47dB	24358	30738	23
48dB	24358	30738	22
49dB	24358	30738	21
50dB	24358	30738	20
51dB	24358	30738	19
52dB	24358	30738	18
53dB	24358	30738	17
54dB	24358	30738	16
55dB	24358	30738	15
56dB	24358	30738	14
57dB	24358	30738	13
58dB	24358	30738	12
59dB	24358	30738	11
60dB	24358	30738	10
61dB	24358	30738	9
62dB	24358	30738	8
63dB	24358	30738	7
64dB	51036	53936	28
65dB	51036	53936	27
66dB	51036	53936	26
67dB	51036	53936	25
68dB	51036	53936	24
69dB	51036	53936	23
70dB	51036	53936	22
71dB	51036	53936	21
72dB	51036	53936	20
73dB	51036	53936	19
74dB	51036	53936	18
75dB	51036	53936	17
76dB	51036	53936	16
77dB	51036	53936	15
78dB	51036	53936	14
79dB	51036	53936	13
80dB	56836	53936	14
81dB	56836	53936	13
82dB	56836	53936	12
83dB	56836	53936	11
84dB	56836	53936	10
85dB	56836	53936	9
86dB	56836	53936	8
87dB	56836	58576	7
88dB	56256	65535	6
89dB	56256	65535	5
90dB	56256	65535	4
91dB	56256	65535	3
92dB	56256	65535	2
93dB	56256	65535	1
94dB	59735	65535	1
95dB	59735	65535	0
96dB	65535	65535	1
97dB	65535	65535	0
98dB	65535	65535	0
99dB	65535	65535	0
100dB	65535	65535	0

Table B-2. Loudness calibration values for 1000 Hz tones.

Actual Loudness	Master Volume	Wave Volume	File Volume
30dB	9279	4060	19
31dB	9279	4060	18
32dB	9279	4060	17
33dB	9279	4060	16
34dB	9279	4060	15
35dB	9279	4060	14
36dB	9279	4060	13
37dB	9279	4060	12
38dB	9279	4060	11
39dB	9279	4060	10
40dB	9279	4060	9
41dB	9279	4060	8
42dB	9279	4060	7
43dB	9279	4060	6
44dB	9279	4060	5
45dB	9279	4060	4
46dB	9279	4060	3
47dB	9279	4060	2
48dB	9279	8699	7
49dB	9279	8699	6
50dB	9279	8699	5
51dB	10439	8699	5
52dB	10439	8699	4
53dB	46976	44077	46
54dB	46976	44077	45
55dB	46976	44077	44
56dB	46976	44077	43
57dB	46976	44077	42
58dB	46976	44077	41
59dB	46976	44077	40
60dB	46976	44077	39
61dB	46976	44077	38
62dB	46396	42917	37
63dB	46396	42917	36
64dB	50456	44077	37
65dB	50456	44077	36
66dB	50456	44077	35
67dB	50456	44077	34
68dB	50456	44077	33
69dB	50456	44077	32
70dB	50456	44077	31
71dB	50456	44077	30
72dB	50456	44077	29
73dB	50456	44077	28
74dB	50456	44077	27
75dB	50456	44077	26
76dB	50456	44077	25
77dB	50456	44077	24
78dB	50456	44077	23
79dB	50456	44077	22
80dB	50456	44077	21
81dB	53356	45817	22
82dB	53356	45817	21
83dB	53356	45817	20
84dB	53356	45817	19
85dB	53356	45817	18
86dB	53356	45817	17
87dB	53356	45817	16
88dB	53356	45817	15
89dB	53356	45817	14
90dB	53356	44657	13
91dB	53356	44657	12
92dB	53356	44657	11
93dB	53356	44657	10
94dB	52208	48136	9
95dB	52208	48136	8
96dB	52208	48136	7
97dB	52208	48136	6
98dB	52208	48136	5
99dB	52208	48136	4
100dB	52208	48136	3

APPENDIX C

BRIGHTNESS CALIBRATION

In accordance with CIE color standards (Commission Internationale de L'Eclairage, 1986), the brightness stimuli consisted of achromatic squares of 4° of visual field displayed on the screen directly in front of the participant. As depicted in Figure C-1, given the distance, *d*, between the participant and the display, the height, *h*, of a visual field, *α*, was calculated according to the following equation:

$$h = d \tan \alpha \tag{46}$$

Given *α* equal to 4° and *d* measured in cm, it was possible to calculate the height, *h*, of the square in cm. By definition, the width, *w*, of the square was equal to *h*.

A program was written in Visual Basic 6 (Microsoft Corporation, 1998) in order to simplify the process of calibrating the screen for the display of the appropriately sized squares (see Figure C-2). This screen calibration tool required four steps in order to determine the proper location and size of the onscreen squares. In the first step, the experimenter moved the location of the program window around the screen until the crosshatch was directly centered between the eyes and at eye level. In this manner, the program determined the screen coordinates, *x* and *y*, of the center of the screen relative to the experimenter's field of vision. In the second step, the experimenter measured the distance, *d*, between the eyes and the crosshatch. This was readily accomplished through the use of a standard tape measure. In the third step, the experimenter adjusted the size of an onscreen square to measure 1 cm wide by 1 cm high. Again, this measurement was readily determined through the use of a standard tape measure. In this manner, the ratio

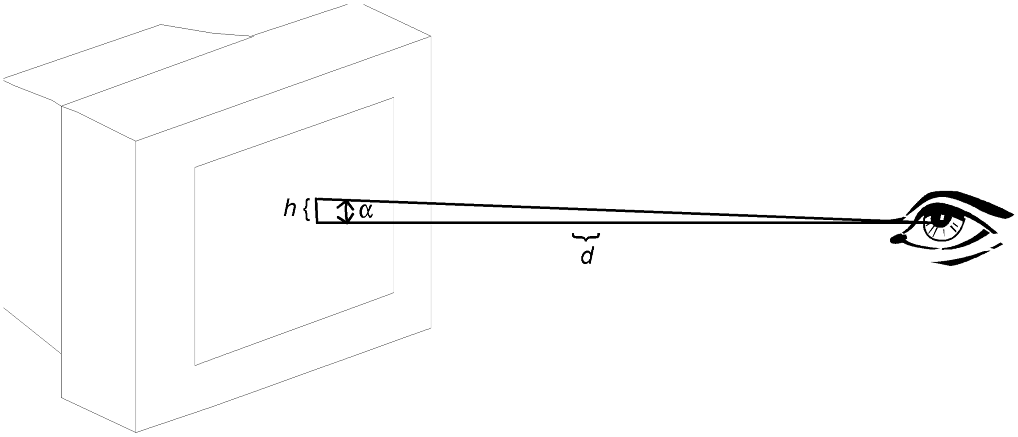
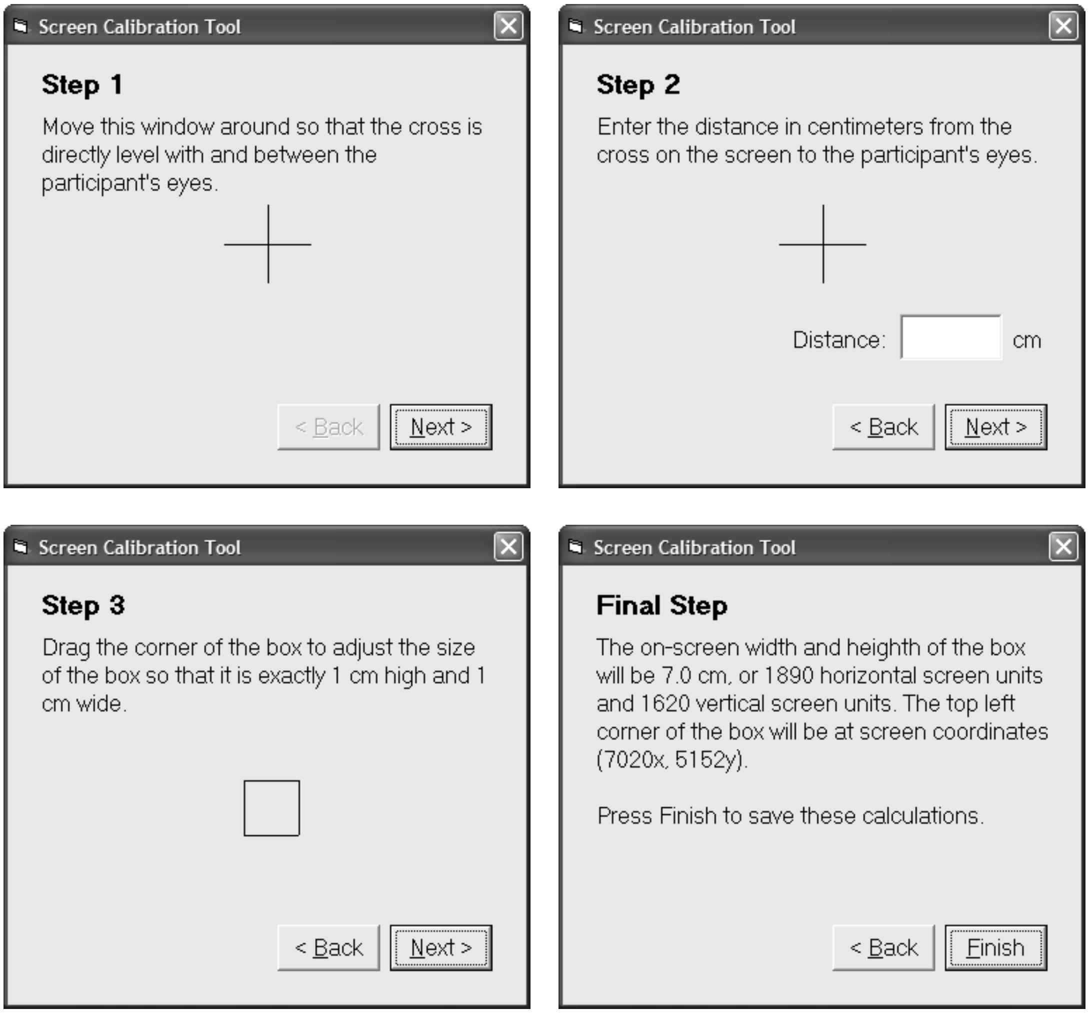


Figure C-1. Measurements used for determining squares of 4° field of vision on a computer display.



Screen Calibration Tool

Step 1

Move this window around so that the cross is directly level with and between the participant's eyes.

Screen Calibration Tool

Step 2

Enter the distance in centimeters from the cross on the screen to the participant's eyes.

Distance: cm

Screen Calibration Tool

Step 3

Drag the corner of the box to adjust the size of the box so that it is exactly 1 cm high and 1 cm wide.

Screen Calibration Tool

Final Step

The on-screen width and heighth of the box will be 7.0 cm, or 1890 horizontal screen units and 1620 vertical screen units. The top left corner of the box will be at screen coordinates (7020x, 5152y).

Press Finish to save these calculations.

Figure C-2. Program for calibrating the coordinates of onscreen squares.

of horizontal and vertical screen pixels to cm, p_x and p_y respectively, was obtained. Using Equation 46, the height and width, *h* and *w*, of the square were obtained. Multiplying *h* by p_x and *w* by p_y gave the dimensions of the square in screen pixels, which, along with the center coordinates of the screen, were saved as a file for subsequent retrieval by the experimental control software.

The squares were color calibrated in accordance with the 1931 CIE Standard Observer Model (Commission Internationale de L'Eclairage, 1986). According to this standard, color is classified into tristimulus values, *XYZ*, which are transformed to the chromaticity values, *Yxy*, where *Y* represents perceived luminance and *x* and *y* represent a two-dimensional color chromaticity classification scheme (see Figure C-3). The CIE color model affords a device independent unified color classification scheme based on the color sensitivity of standard human observers. Any color that can be perceived by standard human observers can be classified according to the CIE chromaticity diagram by mixing the red ($\overline{r_\lambda}$ = 700 nm), green ($\overline{g_\lambda}$ = 546.1 nm), and blue ($\overline{b_\lambda}$ = 435.8 nm) primary color wavelengths. The CIE color model also allows for an achromatic color space in which all colors are perceived with equal intensity. The CIE standard illuminant D65 denotes neutral daylight chromaticity, represented by $x = 0.313$ and $y = 0.329$ on the CIE *Yxy* chromaticity diagram. Similarly, the CIE standard illuminant E denotes true achromaticity, represented by the diagram midpoint at x = 0.333 and y = 0.333. Any achromatic color space remains color neutral even as the luminance value, Y, is increased. The chromaticity coordinates for red, green, blue, and grayscale stimuli are depicted in Figure C-4. Note that with the exception of grayscale stimuli, it was

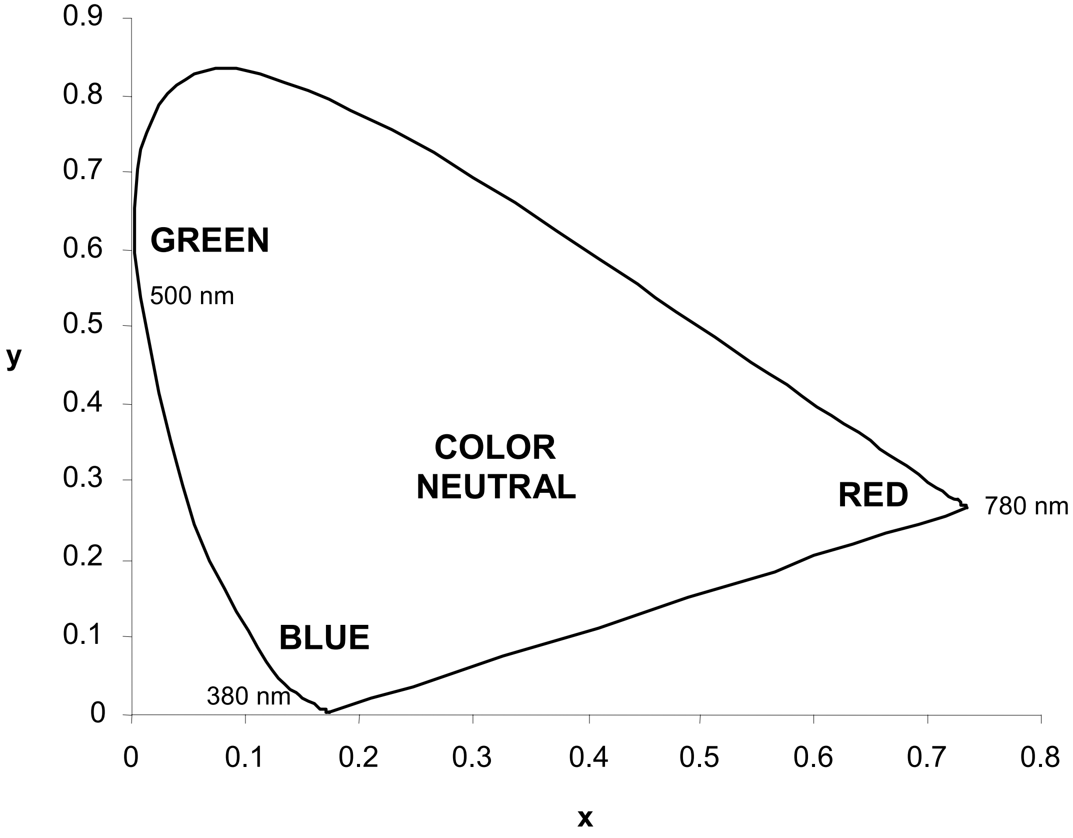


Figure C-3. CIE 1931 Yxy chromaticity diagram with major wavelengths and color areas indicated.

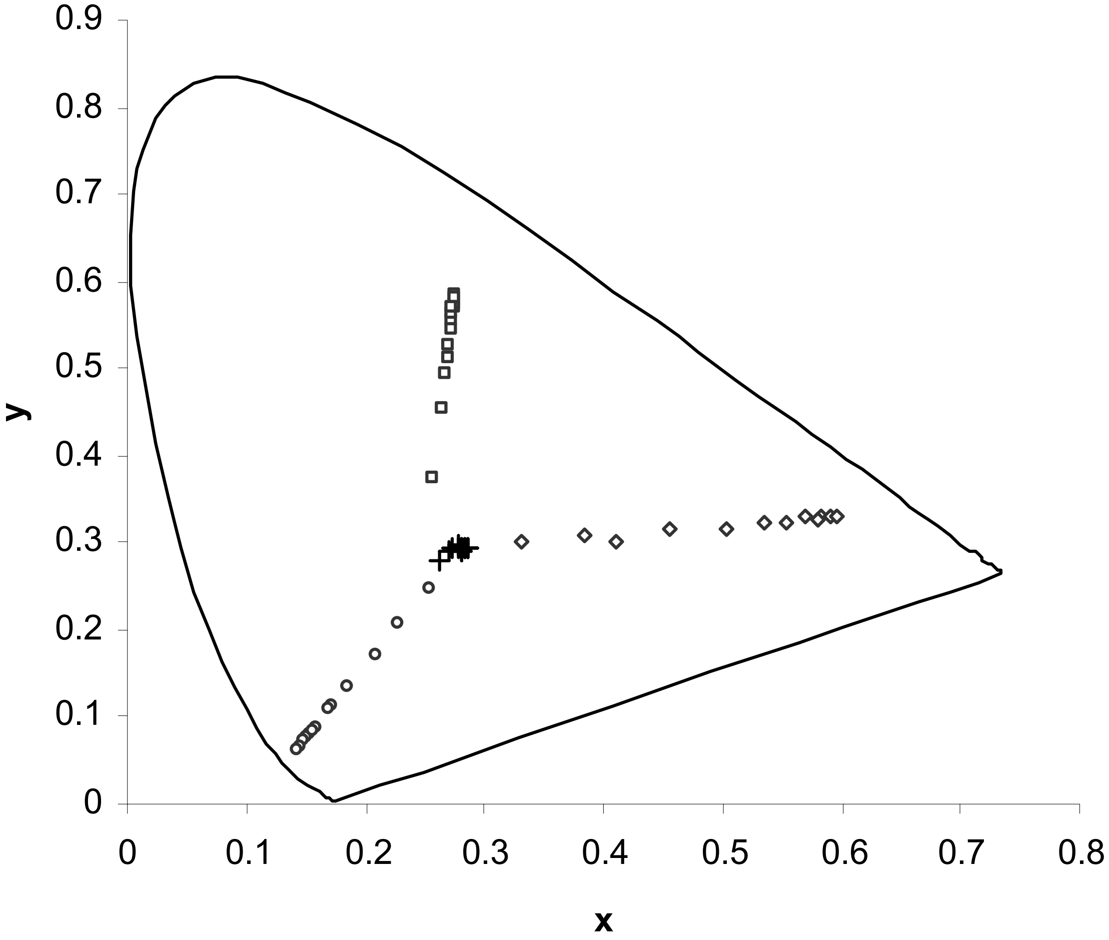


Figure C-4. Chromaticity coordinates in CIE Yxy color space for red (◇), green (□), blue (○), and grayscale (+) stimuli.

impossible to retain chromaticity constancy as the luminance of the stimulus was increased. As the luminance of the stimulus increased, chromaticity tended to become more achromatic.

On a calibrated display, the CIE *Y* values are a close approximation of physical luminance as measured in cd/m^2. Nonetheless, it should be noted that the CIE *Yxy* values are not direct physical measures of luminance and color, but rather a perceptually constant ratio of color and luminance with respect to the gamut of the display device. The luminance of the *Y* value on a calibrated monitor can be approximated by converting the device independent *Yxy* values to device specific red, green, and blue (*RGB*) phosphor values corresponding to the electrical current required to drive the CRT's three electron guns. The mapping of chromaticity coordinates to red, green, and blue phosphor values is accomplished through a color lookup table (CLUT). The CLUTs for the red, green, blue, and grayscale stimuli used in the brightness scaling experiments are found in Tables C-1 through C-5. Note that Table C-1 represents the CLUT for grayscale stimuli in Experiments 3 – 6 and 14, and Tables C-2 through C-5 represent the CLUT for color and grayscale stimuli in the remaining brightness experiments.

For the squares, the computer's graphic display adapter was configured to 24-bit color resolution, corresponding to 8 bits (256 levels) for each of the *RGB* color channels. The CRT display was calibrated using a ColorVision Spyder colorimeter puck to a standard daylight temperature of 6500° K with a black point luminance targeted at 0.00 cd/m^2 and with a white point luminance targeted at 100.0 cd/m^2. For experiments 3 – 6, which only involved grayscale stimuli, a CLUT was generated for grayscale squares

Table C-1. Color lookup table for grayscale stimuli in Experiments 3 – 6 and 14, featuring luminance (L) measured in cd/m^2 and red (R), green (G), and blue (B) screen phosphor values.

L	***R***	***G***	***B***
0	0	0	0
1	34	34	34
2	46	46	46
3	54	56	56
4	62	63	62
5	69	69	69
7	79	79	79
10	93	92	93
14	107	107	107
19	120	121	121
27	140	140	140
37	161	161	161
52	187	187	188
72	216	217	216
100	251	250	250

Note: The CIE chromaticity coordinates, *x* and *y*, were not recorded for each luminance level for Experiments 3 – 6 and 14.

Table C-2. Color lookup table for red stimuli, featuring luminance (L) measured in cd/m^2; red (R), green (G), and blue (B) screen phosphor values; and CIE x and y chromaticity coordinates.

L	***R***	***G***	***B***	***x***	***Y***
1	33	0	0	0.411	0.302
2	62	0	0	0.502	0.317
3	83	0	0	0.535	0.323
4	100	0	0	0.554	0.324
5	115	0	0	0.569	0.329
7	141	0	0	0.581	0.327
9	160	0	0	0.583	0.329
13	193	0	0	0.591	0.332
18	227	0	0	0.595	0.329
24	255	18	18	0.579	0.328
34	255	92	92	0.456	0.314
46	255	139	139	0.385	0.307
64	255	189	189	0.330	0.300

Table C-3. Color lookup table for green stimuli, featuring luminance (L) measured in cd/m^2; red (R), green (G), and blue (B) screen phosphor values; and CIE x and y chromaticity coordinates.

L	***R***	***G***	***B***	***x***	***Y***
1	0	13	0	0.258	0.373
2	0	29	0	0.265	0.455
3	0	41	0	0.268	0.492
4	0	51	0	0.270	0.513
5	0	59	0	0.271	0.526
7	0	74	0	0.272	0.543
9	0	88	0	0.274	0.554
13	0	109	0	0.274	0.564
18	0	132	0	0.275	0.571
24	0	154	0	0.275	0.575
34	0	185	0	0.275	0.583
46	0	218	0	0.275	0.582
64	26	255	0	0.274	0.568

Table C-4. Color lookup table for blue stimuli, featuring luminance (L) measured in cd/m^2; red (R), green (G), and blue (B) screen phosphor values; and CIE x and y chromaticity coordinates.

L	***R***	***G***	***B***	***x***	***Y***
1	0	0	46	0.172	0.114
2	0	0	91	0.158	0.086
3	0	0	122	0.153	0.079
4	0	0	149	0.151	0.075
5	0	0	172	0.149	0.072
7	0	0	213	0.146	0.067
9	0	0	253	0.142	0.062
13	46	46	255	0.155	0084
18	76	76	255	0.169	0.108
24	102	102	255	0.185	0.134
34	137	137	255	0.208	0.172
46	167	167	255	0.228	0.206
64	205	205	255	0.254	0.246

Table C-5. Color lookup table for grayscale stimuli, featuring luminance (L) measured in cd/m^2; red (R), green (G), and blue (B) screen phosphor values; and CIE x and y chromaticity coordinates.

L	***R***	***G***	***B***	***x***	***Y***
1	10	10	10	0.262	0.280
2	22	22	22	0.270	0.289
3	32	32	32	0.273	0.295
4	40	40	40	0.277	0.296
5	48	48	48	0.278	0.295
7	60	60	60	0.280	0.295
9	71	71	71	0.281	0.292
13	89	89	89	0.281	0.292
18	107	107	107	0.282	0.294
24	125	125	125	0.283	0.295
34	151	151	151	0.283	0.295
46	178	178	178	0.285	0.295
64	211	211	211	0.285	0.294

ranging from 1 to 100 cd/m^2 using equal logarithmic spacing. For subsequent experiments, the CLUTs were generated for the red, green, blue, and grayscale squares ranging from 1 to 64 cd/m^2 using equal logarithmic spacing. The restricted brightness range for the latter experiments was necessary because it was not possible to produce red, green, and blue squares with the same maximum brightness as the grayscale squares.

The stimulus values, *S*, were calculated according to the following equation:

$$S_{1..N} = \left[\left(S_N \right)^{\frac{1..N}{N}} \right], \tag{47}$$

Where *N* is the total number of logarithmically spaced units desired, $S_{1..N}$ represents a vector containing the stimulus values, and S_N is the maximum stimulus value. It is assumed that the starting stimulus corresponds to 1 and that *N* is greater than 1. For the present purposes, S_N was assumed to be 100 cd/m^2 for Experiments 3 – 6 and 14, and 64 cd/m^2 for the other brightness experiments. Note that the square brackets, [ ], signify rounding down, $\lfloor\,\rfloor$, or up, $\lceil\,\rceil$, to the nearest integer. The maximum number, *N*, of logarithmically spaced units possible corresponds to the greatest number for which no stimulus values overlap. For example, if S_{10} and S_{11} both equaled 42, it would be necessary to decrement *N* until they no longer equaled the same value and no other *S* values were equal.

The CRT display was allowed a one-hour warm-up period prior to beginning experimental trials. As is shown in Figure C-5, the one-hour warm-up period allowed the phosphor brightness levels to stabilize in order to minimize potential brightness fluctuations over the duration of an experimental session. The settings shown in Figure

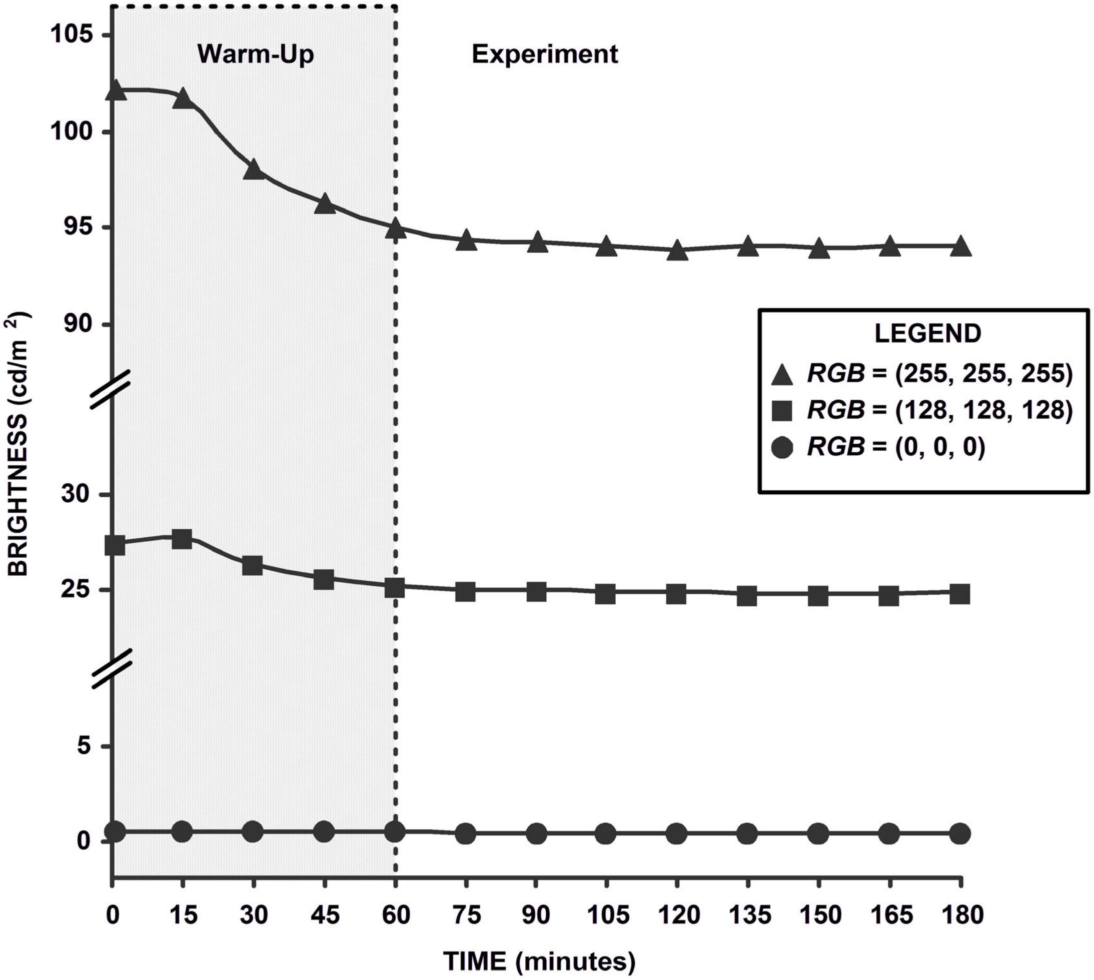


Figure C-5. CRT display phosphor luminance over time across low (●), medium (■), and high (▲) RGB settings.

C-5 were taken with a Samsung 19-inch SyncMaster 950P CRT display attached to an ATI Radeon VE graphics display adapter. The measurements were obtained with a display resolution of 1024 x 768 pixels at a screen refresh rate of 100 Hz.[76] Participants in the brightness experiments were seated in a dark room for five minutes prior to the beginning of the experiment. The room was reflectively lit from behind the participant, resulting in an approximately 10 cd/m^2 light reflectance in the area around the display.

The level of lighting ensured that participants retained photopic visual sensitivity comparable to normal daylight vision. In photopic vision, the cones are maximally sensitive, ensuring full color vision. Often, brightness experiments make use of dark adaptation, in which the eyes have scotopic or mesopic visual sensitivity. In scotopic vision, which occurs in near to full darkness from 10^{-6} to 10^{-3} cd/m^2 (Plainis, Chauhan, Murray, & Charman, 1999), the rods are maximally sensitive and the cones are minimally sensitive, resulting in monochromatic vision. In mesopic vision, which occurs in dusk-like lighting conditions from 10^{-3} to 3 cd/m^2 (Plainis et al., 1999), the rods and cones share color sensitivity, resulting in a generally degraded perception of colors. In avoiding dark adaptation, the brightness experiments reflected the participants' typical daylight sensitivity to brightness and color.

[76] Similar results were obtained for Experiments 3 – 6 and 14, although the screen refresh rate was set at 75 Hz, which produced brighter display values across the color gamut.