# Vision and the Visual System

Davida Y. Teller
and
John Palmer

March 23, 2016

Davida Y. Teller

# Palmer's Preface

Davida Teller worked on this book until 2009. At that time her health prevented her from finishing and she passed away in 2011. The book was well along but incomplete. Beginnning in 2010, I took over the book with the goal of making it accessible to others. As much as possible, my approach was to preserve Davida's voice by removing material that was incomplete rather than adding material.

This printing includes the 26 chapters written by Davida. I have preserved some of her hand drawn figures when possible. Many other figures include material that has been previously published and I have included a citation of the source. Permission to reproduce these materials will be requested when the book is complete.

I intend to do one more round of editing and then send the book to the publisher. At the time of updating this preface, Chapters 1 to 7 are essentially done. There are 19 to go. Figures are essentially complete for Chapters 1-12 and 22-26. Work remains for figures in Chapters 13-21. For those reading it now, please tell me about any errors you find (jpalmer@uw.edu). I do not expect to do any major revisions, but I do wish to remove as many errors as possible.

The editing of this book was done together with my colleagues and friends. The first round of copyediting, gathering figures, and finding references was done by Serap Yigit-Elliott. Thanks Serap! It was supported by the Department of Psychology at the University of Washington at the direction of its chair, Sheri Mizumuri. The second round of editing was done in conjunction with teaching "Davida's vision course" at the University of Washington in the summer of 2012. I thank all of the students of that course for their suggestions. The third round of editing was based on feedback from a seminar conducted in 2015 by Cathleen Moore at the University of Iowa. In addition to Cathleen's wise advice, it was eye opening to get feedback from students that were being taught by someone other than the book's authors.

Throughout I was aided by comments from many colleagues including David Brainard, Angela Brown, Michael Crognale, Norma Graham, Greg Horwitz, Ken Knoblauch, Laurence Maloney, Suzanne McKee, Orin Packer, Cathleen Moore, Zygmunt Pizlo, Joel Pokorny, Dina Popovkina, Maureen Powers, Erik Runeson and Jason Webster. Thanks also go to Maria Pereverzeva for detailed comments and proofing of every chapter. Special thanks go to my wife, Zelda Zabinsky, for her suggestions and enduring support. In addition, I thank Davida's husband, Tony Young, for his encouragement and help assembling the materials. Most importantly, I thank Davida Teller for her trust in allowing me to edit her work.

John Palmer
Seattle, Washington
23 March 2016

# Teller's Preface

Vision science can be defined as the study of vision, the visual system, and the relations between the two. When we study vision, we use psychophysical and perceptual techniques to describe what and how well we see: how good are our spatial and temporal resolution? What is our color vision like? What are the properties of motion perception, form perception, object recognition? When we study the visual system, we use the techniques of neuroscience to describe the properties of the neural machinery – the optics, photochemistry, anatomy and/or physiology of the visual system – that makes seeing possible. And when we study the relationships between the two, we try to answer the question: How do the properties of the neural machinery leave their marks on the properties of perception?

In the terminology of this book, a *linking theory* is an attempt to answer a question of this kind. For example, the statement that the shape of the scotopic spectral sensitivity curve is caused by the absorption spectrum of the photopigment rhodopsin is a linking theory. (Most consider this story to be correct). Another linking theory would be the statement that visual acuity is limited by the optical quality of the eye. (It would be incorrect, but a linking theory nonetheless). I argue below that the most fundamental goal of vision science lies in the discovering and testing of linking theories.

Almost no one starts his or her intellectual life as a vision scientist. Most of us are trained in one or another of the classic disciplines: physiology, psychology, medicine, philosophy, physics, engineering, computer science, and so on. But eventually we may find ourselves working on a problem defined in the parent discipline, but whose answer impinges on the properties of vision and/or the visual system. One day our hearts suddenly beat faster with the insight that we might be able to contribute to the understanding of how people actually *see*. We're drawn to thinking of the visual system as a *system*, with vision as one of its high-level properties. We suddenly want to know how the topic we are studying fits into an understanding of the visual system as a whole. We're hooked – we've just become a vision scientist.

The conceptual base of vision science is remarkably varied and remarkably rich. This is partly because each new vision scientist brings along facts and concepts from his or her parent disciplines. Every few years new ways of thinking arrive from the parent disciplines. With a lot of intellectual work and many graduate seminars, the new concepts are eventually either found to enhance the old or are weeded out, and the discipline is the richer for it. An addiction to new concepts can keep a person in vision science for a lifetime.

But by the same token, vision science can be difficult for the beginning student to penetrate. This is partly because the conceptual base is so broad, and the factual base so extensive. But it's also because, as I mean to convince you, linking theories – explaining perceptual facts on the basis of neural facts – are a philosophically tricky matter.

In writing this book, I have had two goals: an initial goal which I chose, and a later one that forced itself upon me. The initial goal was to write a standard textbook – an introduction to vision science. I particularly wanted to weave together the concepts from the different parent disciplines. I wanted to make them mutually consistent, and accessible to beginners migrating into vision science from other disciplines. In particular, whenever a concept came up with which I had initially had particular trouble, I have tried to explain it in detail, with particular attention to the aspect with which I initially had trouble.

But as time and drafts went on, a second goal forced its way into the book. Again, the vision scientist's question is the question of linking theories: How do the properties of neural signals at the various levels of the visual system cause the properties of our perceptions? More deeply, where do hypotheses about linking theories come from, how are they tested, and by what criteria are they judged? How can we tell a good one from a bad one? I had done some prior work in this area (Teller, 1984), and found myself inevitably drawn back to it.

As it turns out, there has been remarkably little explicit analysis of these questions in the vision literature. Consequently, linking theories can be difficult for the newly arriving student to evaluate. The second goal of the book, then, is to attempt to provide an extended, consistent analysis of the logic and the forms of argument used in vision science in general, and in linking theories in particular.

The format and content of this book are as follows. We begin with an explicit treatment of some of the kinds of propositions that enter into arguments about linking theories. Then, after an introduction to psychophysical techniques, we step through the visual system in the usual order – optics, photochemistry, photoreceptors, retinal processing, and so on. In each case, I provide at least a thumbnail sketch (and often a more extended treatment) of the properties and workings of the particular stage of processing. I then build upon this material to tell and evaluate one or more linking theories about how each level of visual processing leaves its marks on our perception.

For instructors, each chapter is meant to correspond to an individual lecture. The book has been used in teaching a 10-week course with 26 lectures of an hour and a half each. The material is suitable for graduate students or advanced undergraduates with a previous introductory course on sensation and perception.

Finally, a note about myself: As all of its practitioners know, science and philosophy are intensely personal passions. I find I can communicate that passion best to students by including personal anecdotes and making personal appearances in the book. But use of the first person in written work was beaten out of me in the third grade, and makes me uncomfortable still. Accordingly, I have added comments and footnotes about "Teller's" thoughts and experiences. As my professional alter ego, Teller makes her presence known throughout the book. She feels free to express her opinions, and to suggest that the reader stop and think at certain points. Also, she feels free to just stop and *wonder* about things. Of course I do not claim that all the questions Teller wonders about are original – surely most of them have been treated better by others. The goal is to model the sense of wonder that science engenders, and expose the students to the siren song of the next question down the road.

Davida Y. Teller
Seattle, Washington
September 2007

# Teller's Acknowledgments

In 1970, Tom Cornsweet published an introductory text on visual science, entitled *Visual Perception*, (Academic Press). It is still read for its lucid accounts of the relationships between physics, physiology and perception. Although I have had to depart from his leisurely style of explanation because so much more is known by now, Tom's writing has nonetheless provided a model for this book.

Since 1995, several new books on vision science have been published by friends and colleagues. I have used them shamelessly as reality checks, and to educate myself on the parts of visual science that I knew the least about, and I thank the authors for their contributions to my education:

Brian Wandell's *Foundations of Vision* (1995), Sinauer Associates,
D. Milner and M. Goodale's *The Visual Brain in Action* (1995), Oxford University Press,
Bob Rodieck's *The First Steps in Seeing* (1998), Sinauer Associates,
Bruce Goldstein's *Sensation and Perception* (1999), Brooks/Cole,
Clyde Oyster's *The Human Eye: Structure and Function* (1999), Sinauer Associates,
Stephen Palmer's *Vision Science: Photons to Phenomenology* (1999), MIT Press,
Mike Levine's *Fundamentals of Sensation and Perception* (2000), Oxford University Press,
Martin Regan's *Human Perception of Objects* (2000), Sinauer Associates.

Special mention should also be made of the collections of papers on vision science in *The Cognitive Neurosciences* (1995) and *The New Cognitive Neurosciences* (2000), Michael Gazzaniga (ed), MIT Press.

Personal thanks must begin with Dr. Maureen (Mo) Powers, who started this book with me, and who produced early drafts of some of the chapters. Unfortunately, changes in her life led her to withdraw from the book early in the writing process. Mo brought activation energy and enthusiasm to the project, along with the conviction that writing a book was actually possible. Well begun is half done. (Well, not really, but it made all the difference.) Without Mo's enthusiasm, this book wouldn't have happened. Thanks, Mo.

I also especially thank my long-time colleague and friend, John Palmer, for challenging my thinking at many junctures over the years.

I thank a number of other colleagues for discussions, email conversations and reading chapters of the manuscript: Steve Buck, Tom Cornsweet, Dennis Dacey, Anita Hendrickson, Don Hood, Temy Kennedy, Mike Landy, Barry Lee, Walt Makous, John Maunsell, Suzanne McKee, Matt McMahon, Bill Newsome, Orin Packer, Joel Pokorny, Fred Rieke, Micheal Rudd, Michael Shadlen, Julie Schnapf, Steve Shevell and forgotten others. In addition, I thank the students in my vision class who have read whatever chapters were available, and who annually rekindled my will to continue.

In my career I have had four mentors who most effectively challenged my intellect. They

are the Gestalt psychologist Hans Wallach; the philosopher Michael Scriven; the engineer-turned-vision-scientist Tom Cornsweet, and the optometrist-turned-vision-scientist Gerald Westheimer. All communicated to me their passion for ideas. And their collective wisdom can be summarized in two words: Think harder.

Finally, I thank my husband, Tony Young, for making the usual sacrifices an author demands of her family. I also thank him for using his skills as a photographer to provide some of the illustrations included in this book. We have had happy times searching the world for just the right image to illustrate one perceptual concept or another. The series of pictures on the development of vision in infants exists because of his creative skills and imagination (Teller, 1997).

# Contents

# Part I

# INTRODUCTIONS, OPTICS AND TRANSDUCTION

# Chapter 1

# The Domain of Visual Science

**Contents**

## 1.1 What is vision science?

Vision science is the study of vision, the visual system, and the relations between the two.

When vision scientists study *vision*, we study what and how well people see. The scientific goal is to describe and quantify our sensory and perceptual capacities – our capacities to respond to physical stimuli by using our eyes. What is the dimmest light we can detect? How fine are the finest details we can resolve? What is our color vision, or our stereovision, or our perception of motion like? How accurate are our perceptions of the sizes, shapes, and locations of objects? How well can we recognize objects? What can we see, and what can we *not* see?

When we study the *visual system*, we study the "machinery" – the optics, photochemistry, anatomy, and physiology of the eye and the parts of our brains that serve vision. How fine an image is made by the optics of the eye? How is light absorbed? Through what processing stages does incoming information pass? How is information recoded – what computations are performed – as we go from one stage of neural processing to the next? To what physical stimuli or aspects of the visual scene are neurons at different levels of the visual system tuned to respond?

What about the relations between the two? For many vision scientists, studying vision and the visual system separately is not the ultimate goal. Rather, the ultimate goal is to *explain why we see as we do, on the basis of the properties of the machinery that makes seeing possible.* Why can we detect some lights and not others; resolve some spatial details and not others; see objects accurately under some conditions but not others? Vision scientists want to understand how the computations carried out by the visual system both enable and limit our visual capacities, and how

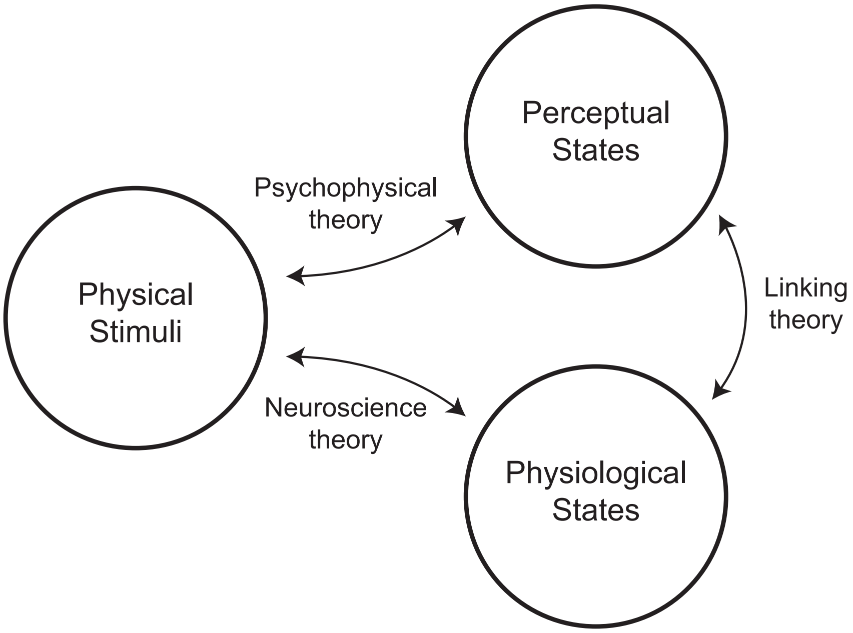


Figure 1.1: The domain of vision science includes three types of entities, three types of mapping rules and three kinds of theory. The entities are physical stimuli, physiological states, and perceptual states. The mapping rules are: (a) from physical stimuli to perceptual states; (b) from physical stimuli to physiological states; and (c) from physiological states to perceptual states. Each kind of mapping rule leads to a different domain of theory: psychophysical, neuroscience and what we will call linking theory.

they leave their marks on our visual perception. We will call these attempted explanations *linking theories*. They link vision and the visual system.

### 1.1.1 Entities, mapping rules and theories

Let us now put these questions in a slightly broader context. As shown in Figure 1.1, vision science is concerned with three kinds of entities, three kinds of mapping rules, and three kinds of resulting theory.

Let us start with the three kinds of entities. The first kind is *physical objects*, or *physical stimuli* – the physical objects and light sources that send light to our eyes. The second is *physiological states* – the states of the many varieties of neurons in the visual system, occurring in response to the physical objects and light sources that lie in front of us. And the third is *perceptual* states of phenomenal experience. These conscious experiences usually correspond remarkably well to the objects and other stimuli in the physical world. Typically, we will accept appropriate behavioral reports as a stand in for perceptual states. These three kinds of entities are shown by the circles in Figure 1.1.

Between each pair of entities there is a set of *mapping rules*. We are looking for rules of correspondence of the form, entities $x_1, x_2$, in the physical domain occurring in conjunction with entities $y_1, y_2$, in the neural domain, and with entities $z_1, z_2$, in the perceptual domain. The fundamental goal of vision science is to determine the mapping rules between each pair of entities, by whatever techniques are required, and however simple or complex these mapping rules might be. The three kinds of mapping rules are shown by the three arrows in Figure 1.1.

The three kinds of mappings are studied by three very different kinds of techniques. We study *physical-perceptual mappings* by means of the discipline of *visual psychophysics*, using sophisticated behavioral techniques to ask human subjects what they see when they view particular stimuli. In this discipline the assumptions about entities and mappings are elaborated into *psychophysical theory* that relates the physical world to the perceptual world.

We study *physical-physiological mappings* with the techniques of visual neuroscience, such as presenting particular stimuli and recording the activities of particular neurons at various levels of the visual system. In this discipline the assumptions about entities and mappings are elaborated into *neuroscience theory* that relates the external physical world to the physiological world.

What about *physiological-perceptual mappings*? Suffice it to say that at this point the techniques used for exploring these mappings are harder to define. At the same time, for many vision scientists, these are the heart of the matter, because as stated earlier, the ultimate goal of many vision scientists is to explain why we see as we do, on the basis of the properties of the neural machinery that makes seeing possible. These mappings link the vision and the visual system. For that reason, we will refer to theories built upon such mappings as *linking theories.*

The world as perceived is a strikingly accurate representation of the physical world, allowing us both to perceive objects and to carry out appropriate motor activity with respect to them. But the existence of psychophysics notwithstanding, there are no direct causal mappings between physical and perceptual entities. The perceptual representations we have of the physical world are created by passing through the other two legs of the triangle: first through physical-physiological and then through physiological-perceptual mappings.

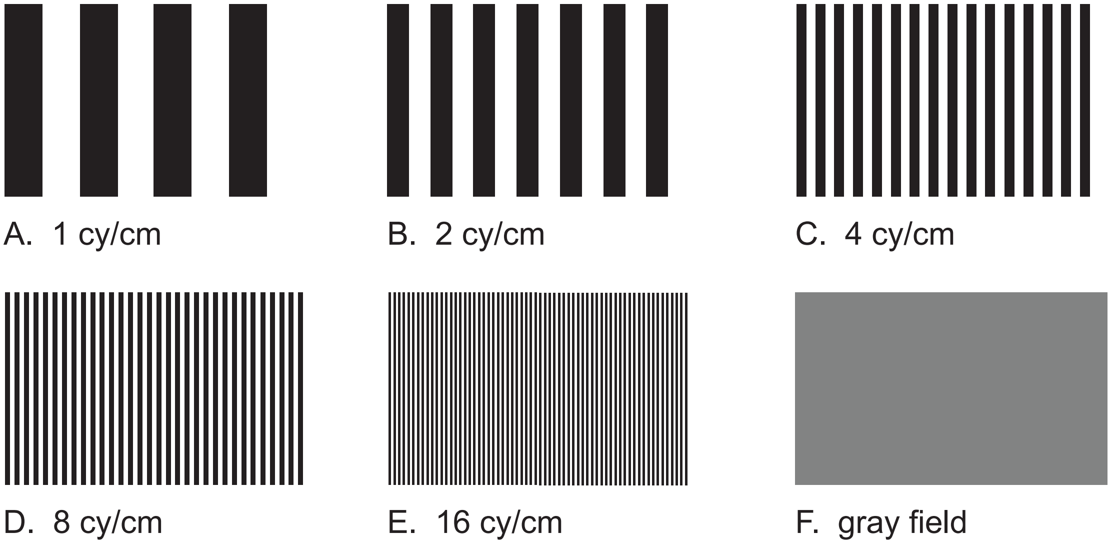
In the next few sections of this chapter, we will expand on each of these types of mappings. Then, in later sections of the chapter, we will expand at length on the mappings between physiological and perceptual states – because of their philosophical complexity and because of the fascination they carry for the authors of this book.

### 1.1.2 An example: Grating acuity

Let's take a concrete example of a physical-perceptual mapping from psychophysics. Figure 1.2 shows five sets of regular black and white stripes, called *square-wave gratings*, and one homogeneous gray field. The stripes in each grating are half as wide as the stripes in the next coarser grating. At normal reading distance, you can probably see the spatial variation of light across all of the gratings (e.g., A-E). Now view the gratings from a 3 meter distance: what is the finest grating that you can see?

Somewhere between gratings D and E, your perception of black and white stripes probably fades perceptually into a uniform gray, and cannot be distinguished, or *discriminated*, from the homogeneous field F. Your *grating acuity* is defined as the finest stripes that you can just barely perceive as stripes, or discriminate from the homogeneous field on the basis of their spatial pattern. Grating acuity is a measure of the limit of detail you can resolve in space; it defines the *spatial resolution* of your visual system.

We now elaborate on the three sets of entities and the three sets of mapping rules that make up the domain of vision science, using grating acuity as our example.



A. 1 cy/cm

B. 2 cy/cm

C. 4 cy/cm

D. 8 cy/cm

E. 16 cy/cm

F. gray field

Figure 1.2: Five square wave gratings and a blank (homogeneous) field. The gratings are labeled A, B, C, D, E. The gray field, F, approximately matches the gratings in average light intensity. If you have normal vision, at 3 meters distance you will probably be able to resolve gratings A-D, but not grating E. The finest grating you can discriminate from the gray field, F, defines your grating acuity. Unfortunately, imprecise printing often creates irregularities in the finer grating that make it more visible than if it was printed precisely. If this is the case, try the experiment from 6 meters and perhaps Grating D will not be resolvable.

## 1.2 Three kinds of questions, three kinds of theory

### 1.2.1 Psychophysics: Mapping from physics to perception

What are physical-perceptual mappings? Vision scientists want to know how and how well people see – to measure and quantify human sensory and perceptual capacities. To find out, we bring people (usually called *subjects* or *observers*) into the laboratory, and use well-controlled physical stimuli and sophisticated behavioral, or *psychophysical*, techniques to measure their visual capacities. The results of such experiments yield objective descriptions of the facts about visual acuity, color vision, distance perception, object recognition, and so on. We will look in detail at psychophysical techniques in Chapter 2 and 3.

The most immediately interesting fact about grating acuity is that it is so readily definable. There is a range of coarse gratings that you can resolve, an abrupt transition, and then a set of finer gratings that you can't resolve. Notice the first of many mismatches of properties between the physical and the perceived. The physical variation is continuous – there is nothing in the stimulus continuum that suggests a basis for any break of perceptual properties – but the change of perception from seeing to not seeing is abrupt.

At the perceptual level, grating acuity has several additional interesting properties. The finest grating you can see on the page varies with the viewing distance, the light level, and the part of your field of view in which the grating is presented. Try these experiments. First, prop up the book across the room, perhaps 6 meters away. From this distance, you will probably be able to resolve only a few of the coarsest gratings. Now walk toward the book. Every time you cut the distance to the book in half, you should be able to resolve one more grating.

Second, turn down the lights in the room in progressive stages, or prop the book up in the light of a window in the evening, when the outdoor light is steadily decreasing. As the light level decreases, your grating acuity will decrease, and you will need to move closer to resolve gratings that were easily resolvable at your original distance in full daylight. And third, instead of looking directly at the gratings, look above them by various amounts while still trying to resolve the stripes. The greater the *eccentricity* of the grating – the greater the displacement from your center of vision – the lower will be your grating acuity.

By picking out the finest grating you can see under a variety of conditions, you have just been a subject in an (informal) psychophysical experiment. You have made a series of measurements of the mappings from physical to perceptual entities. You have encountered the perceptual phenomenon of grating acuity, and you have learned about several important parameters – distance, light level, and eccentricity – that influence it.

### System Properties: Bumblebees can fly

Grating acuity and its variations with distance, light, and eccentricity can be called *system properties* of vision. Such system properties are interesting in and of themselves. But they become more interesting when we realize that system properties provide us with logically compelling information about some of the physiological properties of the visual system, without a single physiological experiment having been done. Oddly, we are arguing that perceptual results imply physiological conclusions. A fancier way of saying this is, system properties place important constraints on models of the visual system. The constraints depend on whether you (the subject) resolve the grating or not.

#### If you resolve the grating, information physically present in the stimuli must have been retained.

When you discriminate a grating from the homogeneous field, it follows that information that the two stimuli differ is retained from the physical stimulus, all the way through every one of the series of anatomical/physiological stages that make up your visual system. It is retained through every link of a *causal chain*, right up to your conscious perception, and right out through whatever motor system you use to tell the experimenter you resolve the grating. The question then becomes: How is the information carried, or *coded*, at each anatomical stage?

#### If you don't resolve the grating, information physically present in the stimuli must have been lost.

When a grating is present but you don't discriminate it from the homogeneous field, it follows that information that the two stimuli differ must be lost somewhere, at one or more stages of processing in your visual system. Like information retention, information loss implies a physiological conclusion – there exists a stage or stages of visual processing at which the information is lost. The question then becomes: at what anatomical stage (or anatomical *locus*) is the information lost? Questions of this kind are sometimes called *locus questions.*

Because they specify the limits of actual visual function, system properties like spatial resolution exert a great deal of power over mathematical and physiological models. A classic joke based on the importance of system properties concerns some early aerodynamic engineers who tried to make a mathematical model of the bumblebee, to find out how it flies. They tried and tried, but the model bumblebee fell out of the air every time. Eventually, the engineers concluded that bumblebees can't fly! But in fact, bumblebees can fly, so immediately we know that something about the model had to be wrong. Similarly, if a vision scientist makes a model of the visual system, and that model predicts that human beings can't resolve gratings as fine as the ones you could resolve in Figure 1.1, we know immediately that something is wrong with the model.

In more abstract terms: the system properties of vision are "black box" measurements made on a highly complex, multi-stage physiological system, and rarely reveal the details of the machinery inside the box. But even in the absence of knowledge about the inside of the box, system properties put fundamental constraints on models of how it works. If the person can do X, then the underlying visual system must be such as to allow X to occur. Any model that claims that the person can't do X must be wrong. We will refer to arguments of this form as *bumblebees can fly* arguments. On the other hand, if the person can't do X, then information is lost, and there must exist a locus (or loci) of information loss.

In addition to such logically compelling implications, system properties often play a less formal but very important theoretical role in vision science. That is, system properties can encourage speculation and theory-building concerning information processing within the visual system. Such speculation, and the computational models it generates, can provide the motivation for visual neuroscientists to go and look for particular kinds of elements or processing circuits within the eye and visual system. We will see many examples of this pattern of argumentation, from psychophysics to physiology, as we go along.

### 1.2.2 Neuroscience: Mapping from physics to physiology

In this section we provide a brief preliminary encounter with physical-physiological mappings. We begin with a simplified cartoon of the anatomy of the eye and the early parts of the visual system.

Figure 1.3 and Figure 1.4 show an overview of the early parts of the human visual system. As shown in Figure 1.3A, the optics of the eye create an optical image of the visual scene, called the *retinal image*, at the back of the eyeball. The major optical elements that form the image – the *cornea*, the *iris* (and the hole in its center, the *pupil*), and the *lens* – are shown in Figure 1.3B. The *retinal image* is focused on the *retina*, a thin sheet of neural tissue that lines the back of the eyeball. The *fovea* is a small, central region of the retina, specialized for high acuity. When you looked at each acuity grating in Figure 1.2, you turned your eyes to place your fovea under the retinal image of that grating.

Figure 1.4A shows a cartoon of a small piece of the retina, and some of the neurons it contains. The *photoreceptors* at the back surface of the retina catch the incoming light, and initiate a set of neural signals. Several other types of cells within the retina process these signals before they arrive at the *retinal ganglion cells.* The output processes (axons) of the ganglion cells make up the *optic nerve.*

Oddly, the retina is "in backwards" – the photoreceptors lie at the back (or outer) surface of the retina. In consequence, the light must pass through all of the other neurons in the retina before getting to the photoreceptors for absorption; and the optic nerve must pass through the retina to

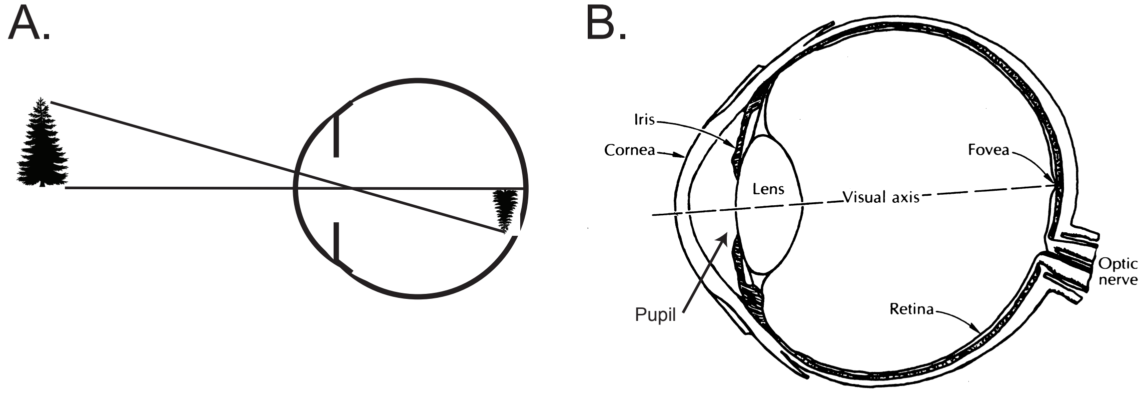


Figure 1.3: Overview of the eye and its optics. A. The optics of the eye form an image of the physical world on the back of the eyeball. B. This sketch shows a horizontal section through the right eye, labeling the major optical elements of the eye that work together to form the retinal image: the cornea, pupil, and lens. It also shows the retina, a thin sheet of neural tissue that lines the inside of the eyeball; the fovea, a central region of the retina specialized for high acuity; and the optic nerve, the nerve bundle that leaves the eye and carries visual signals toward the brain. (Modified from Cornsweet, 1970, Fig. 3.11, p. 40)

get out of the eye, creating a blind spot in your visual field. The probable reason for having the retina in backwards is that (as we will see) the photoreceptors are highly active metabolically, and need to be near a blood supply; and a blood supply that traversed across the front of the retina would itself get in the way of the retinal image.

As shown in Figure 1.4B, the optic nerve projects to a way station – *the lateral geniculate nucleus (LGN)*, deep within the brain – before signals are sent on to the *primary visual cortex*, a cortical region located at the very back of the brain. From there, signals project outward and forward to many higher levels of cortical processing (see Figures 17.5 and 17.4 in Chapter 17 for an intimidating preview).

When we ask about physical-physiological mappings, we are trying to trace the features of each physical stimulus through the visual system. We want to understand and quantify the characteristics of the eye's optics, and the means whereby light is transformed into physiological signals. We want to know the anatomy and physiology of each stage of processing in the early visual system and the cortex, and the computations – information losses, retentions, and recodings – that take place at each anatomical locus within this multistage information processing system. These are the questions of visual neuroscience.

### 1.2.3 Linking theory: Mappings from physiology to perception

But what about physiological-perceptual mappings? Which anatomical structures and which physiological computations are critical to producing any particular system property of perception? And how, exactly, do they produce it? We have now come to the theory that links vision and the visual system.

In order to give an example of exploring this question in depth, we now ask: What limits grating

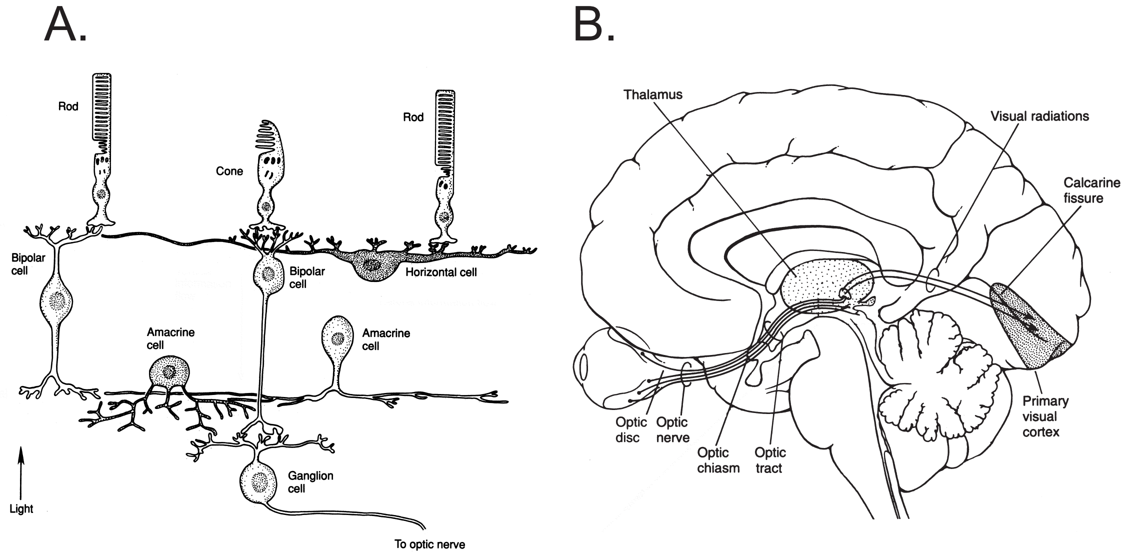


Figure 1.4: Overview of the visual system. A. A schematic of a small section of the retina showing its neurons. Light enters from the bottom of the diagram. The photoreceptors capture the light, and pass on neural signals (via several other types of neurons) to the ganglion cells. The output processes (axons) of the ganglion cells travel across the inside surface of the retina to combine into the optic nerve. B. A sketch of the pathway from the eye to the visual cortex. The eye is at the lower left. In the primary projection from the eye, the optic nerve and the optic tract carry the signals sent by the ganglion cells to the lateral geniculate nucleus (LGN). Axons from the LGN project to the primary visual cortex and from there to other cortical areas (not shown) for further processing. [Modified from Kandel, Schwartz, and Jessell (2000, A. Fig. 28-6, p. 409; B. Fig. 27.4, p. 527).]

acuity? At what stage or stages of visual processing is limited spatial resolution imposed on the incoming sensory signal? At which anatomical *locus* is the information lost? From Figure 1.3 and Figure 1.4 you can already intuit that the limit could be imposed at any of several stages.

In this section, we will raise four hypotheses (or linking theories) that relate the observed spatial resolution to the underlying physics and physiology. Each is an example of a physiological-perceptual mapping. We emphasize that the treatment of these possibilities is qualitative and intuitive at this stage – we just want you to end up believing that the answer is not obvious, and there are several very different and plausible candidate explanations. In later chapters, we will take up each of these possibilities in much more detail. In the meantime, just imagine you are looking out at Figure 1.2 through the neurons in your visual system, and trying to use the incoming spatial pattern of light or neural activity to tell the difference between a grating and a homogeneous field. Which level or levels of the processing system limits your spatial resolution, and what features of processing impose that limit?

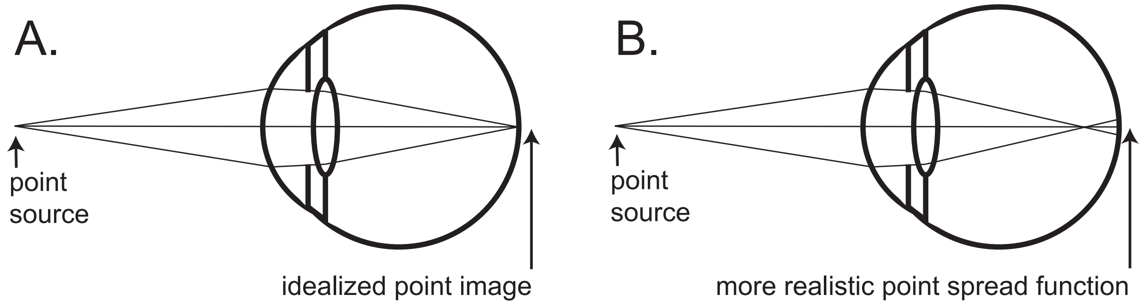


Figure 1.5: Idealized and realistic optical point spread functions. A. A point source and an idealized point image, drawn on the premise that all rays that leave the point source and enter the eye are perfectly bent by the optics, so as to end up at a single point on the retina. B. A more realistic, imperfect image – an extended "blob" of light – caused by optical imperfections in the eye. Some of the rays originating from the point source are bent too much or too little, so that they arrive near but not exactly on the idealized image of the point source. The distribution of light in the image of a point source is called a point spread function.

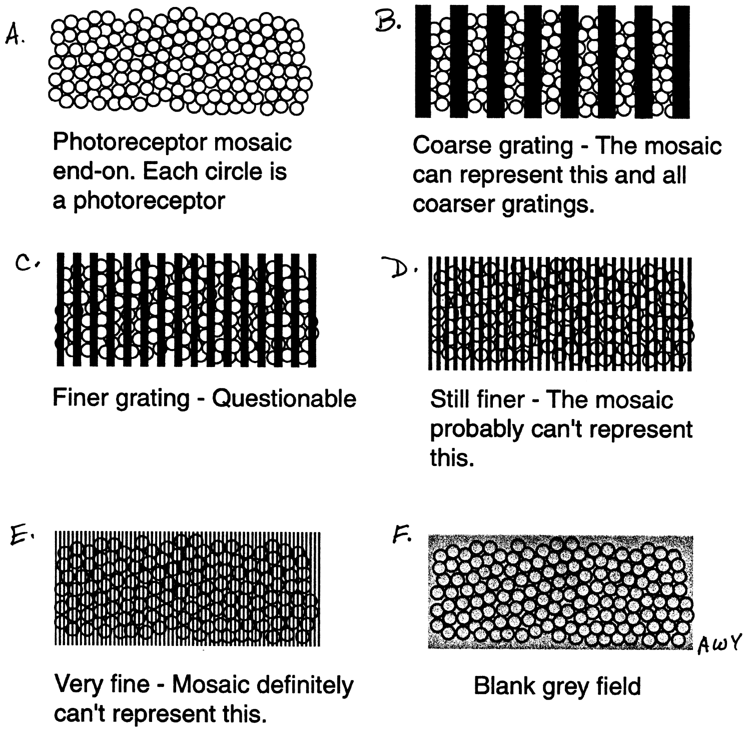
### Hypothesis 1: The optics of the eye

The black and white gratings in Figure 1.2 are patterns of light that exist in the physical world. When you look at a grating, the optics of the eye make an image of that grating on your retina. The process is straightforward, as shown in Figure 1.5A. Rays of light coming from a particular point on the grating leave that point and travel in straight lines in all directions. An image is formed because the optical system captures a subset of those rays, and (ideally) bends each ray just enough so that all of the rays that start at a point on the object are reunited at a point in the image. Neighboring points on the object are represented at neighboring points in the image, with the result that an image of the grating is formed at the back of the eye.

But real optical systems are not perfect. As shown in Figure 1.5B, in reality the rays from a single point on the object do not converge perfectly to a single point in the image. They are slightly spread out in the image, forming a small irregular blob. Technically, the optical image of a point of light is called a *point spread function* because it describes how much the rays from a single point in the object are spread over the image. Rays from neighboring points form neighboring point spread functions, and these imperfect blobs of light will soften the boundaries of the stripes of the grating in the retinal image.

Intuitively, what consequences would such optical imperfections have for grating acuity? A coarse grating will be represented faithfully, still recognizable and resolvable, but with slightly fuzzy edges. But we can intuit that when the stripes in the image of the grating are about equal to the width of the point spread function, the blobs from neighboring stripes will begin to overlap, so the stripes will become less resolvable. As the stripes become even finer, the overlapping point spread functions could eventually produce a homogeneous wash of light, and we wouldn't be able to tell the striped field from the homogeneous field in Figure 1.2.

Now, remember the argument you already know: bumblebees can fly. If you can resolve grating D (say) in Figure 1.2, you know immediately that the optics of your eye must be of at least sufficient quality to make a perceptible image of grating D on your retina. However, if you can't



Photoreceptor mosaic end-on. Each circle is a photoreceptor

Coarse grating - The mosaic can represent this and all coarser gratings.

Finer grating - Questionable

Still finer - The mosaic probably can't represent this.

Very fine - Mosaic definitely can't represent this.

Blank grey field

Figure 1.6: Discrete sampling by the photoreceptor mosaic. A shows a schematic of the photoreceptor mosaic, viewed face on. Each circle is a photoreceptor. Each photoreceptor catches light from a small but spatially extended region of the retinal image, and sums the light over this region. B, C, D and E show images of four different square wave gratings, from coarse to fine. Intuitively, it seems likely that the grating in B will make a signal that varies systematically across the matrix of photoreceptors, but the finer gratings, especially the very fine grating schematized in E, might make only a homogeneous signal across the matrix. If so, the grating represented in E would not be discriminable from the blank field represented in F, and information about the spatial structure of the grating would be lost.

resolve grating E (say), you know that information from grating E is lost somewhere within your visual system. The optical imperfections argument suggests intuitively that the optics *could* be the level that imposes the limit on your grating acuity, and prevents you from resolving grating E. To find out, we'd have to find a way to measure the actual quality of the optics of the human eye, and develop a quantitative theory of the effects of optical quality on vision. We will return to this task in Chapter 4 and 5.

### Hypothesis 2: Photoreceptor spacing

Within the eye, at the back of the retina, lies a layer of tightly packed, highly specialized neurons called *photoreceptors*. The retinal image falls on the matrix of photoreceptors, and each photoreceptor captures the light from its particular region of the two-dimensional retinal image. However, the photoreceptor sums the light it catches over its whole extended region, and doesn't keep track of where each bit of light came from within that region. So the continuous retinal image is sampled piece-by-piece; that is, the photoreceptors perform a *discrete sampling* of the retinal image.

What consequences does discrete sampling have for grating acuity? As shown in Figure 1.6, each of the stripes in the optical image of a coarse grating covers many photoreceptors, so coarse gratings will yield variations of outputs across the *matrix* (or *mosaic*) of photoreceptors as a whole. By analyzing the spatial pattern of those outputs, we could readily tell that the input differed from that created by a homogeneous field. But a grating so fine that it puts more than one stripe on each photoreceptor is in danger of destruction, because it may yield no regular variation of output across the matrix of photoreceptors. In other words, a fine enough grating may produce the same spatial pattern of photoreceptor signals as does the homogeneous field of light, and thereby not be discriminable from it. At this point we might guess that the finest grating that gets through a discrete sampling matrix will have stripes just wide enough to put one stripe on each photoreceptor.

Again, bumblebees can fly. We already know that the limit imposed by discrete sampling at the photoreceptor mosaic can't be worse than the behaviorally measured acuity. But if the optics don't impose the resolution limit, the receptor matrix might. To find out, we'd need to know the sizes of the photoreceptors and their spacing, and make a quantitative model of the effects of discrete sampling. We return to this task in Chapter 5.

### Hypothesis 3: Neural convergence within the retina

The concept of neural convergence has already been illustrated in Figure 1.4A: there are many more photoreceptors than ganglion cells. In fact, across the retina as a whole there are about 100 million photoreceptors but only about 1 million ganglion cells. That is, on average, over the retina as a whole, there is a 100:1 spatial *convergence* of neural signals. Intuitively, it is easy to imagine that unless special provisions are made, spatial resolution could be compromised here.

Bumblebees can fly. Even despite this average 100:1 convergence of photoreceptors onto ganglion cells, you already know that the ganglion cell layer, like all of the other layers of the visual system, must pass on information allowing us to resolve the finest gratings we do resolve. But if the optics and the photoreceptor spacing don't limit grating acuity, maybe neural convergence in the retina does. We'll look at this question again in Chapter 13.

### Hypothesis 4: Later levels of the visual system

Beyond the retina, at the cortical level, there is a long series of anatomical stages and physiological recodings of visual information. In looking for the limits of grating acuity, our question about each level would be the same. How does that level manage to preserve and pass on information about the finest grating we can resolve? How is information about the grating carried (or coded) at this level? And if the resolution limit is not imposed before this level, might it be imposed here, and if so, by what computational process? The second half of the book deals with these later levels of the visual system. [Before you go on, why not lay a bet as to which level imposes the resolution limit, and give the best justification you can at this stage for your answer.]

## 1.3 Linking theories

We now describe in more detail the idea of a *linking theory.* Linking theories are proposed physiological explanations of perceptual events. In short, they are theories linking vision (perception) and the visual system (physiology). For example, the attribution of the grating acuity limit to the optics of the eye, or to the properties of the retinal mosaic, or to a combination of both, are all linking theories. Linking theories can be speculative, or they can be argued on the basis of quantitative theory.

*Locus questions* relate to a particular kind of linking theory: Where within the neural information processing system is information lost, or importantly recoded, in such a way as to bring about the correspondence between physical stimuli and perception? The four options for the locus of information loss in grating acuity represent four possible answers to a locus question. The usefulness of these concepts will become more meaningful through examples encountered throughout the book.

Of further interest are *design questions.* Design questions are *why* questions. These questions are concerned with why human vision and the human visual system take the form they take. For example, why is our grating acuity as good as it is, and why is it not better? Design constraints are imposed from many sources: the laws of physics, the physiological properties of neurons, and the effectiveness of various visual coding schemes for various purposes. In addition, the design of the visual system is shaped by the competing evolutionary pressures that combined to shape the organism as a whole. Many competing pressures have acted upon the design of the visual system, and the current features of the visual system are doubtless historical compromises among these pressures. The answers to design questions are usually speculative, but often instructive and interesting as well.

The fundamental design question about grating acuity is: What factor or combination of factors necessitates that visual resolution have the limits that it has? Is it that the optics can't be any better? Or can't the photoreceptors be any smaller? Or can't there be any more ganglion cells? Or that is there a constraint imposed on some later level of the system? Or might it be that no one level is to blame for the spatial resolution limit, but rather that the limit is imposed by conflicting design necessities, and many levels of the system conspire to impose this limit in a more complex way? What would have to be changed in order to improve our acuity by a factor of two, and what would be the cost?

### 1.3.1 Linking theories and the mind/brain problem

The mappings between physiological and perceptual entities are a central topic in the philosophy of mind. Vision science can be particularly perplexing from a philosophical perspective, because it seems that through linking theories, vision scientists hope to explain mental events (visual perception) on the basis of physiological events (neural activity). This hope brings us to close encounters with the mind/body or mind/brain problem.

For centuries philosophers have argued about the nature of the relationship between mind and brain. In particular they have argued about whether mind and brain are a single physical entity (a position called *materialism*), a single mental entity (a position called *idealism*), or two separate entities (a position called *dualism*); and if two entities, whether one of the two holds a causal priority over the other. Many variants of each of these positions have been formulated, and the debate continues to fascinate philosophical audiences across the centuries (Chalmers, 1996; Metzinger, 2000).

Most vision scientists are probably most comfortable with a materialist perspective. That is, most of us probably believe, implicitly if not explicitly, that between mind and brain, the brain is the primary causal agent. Moreover, perceptual events become less mysterious when they are viewed simply as high-level properties of the brain. To support this view, an analogy can be made between the properties of chemicals and chemical compounds, and the properties of brains and conscious states. Just as water can be viewed as a high-level property of hydrogen and oxygen, so a conscious perception can be viewed as a high-level property of a complex neural network. From this point of view, linking theories are all about making a *causal story* of how perception arises from physiology.

Given the inevitable continuation of these debates, our point of view is rather than developing a science that depends upon a single view of the mind/brain problem, vision science would be wise to finesse it. That is, use a formulation of the questions of vision science that will be robust, and survive across many or all of the different philosophical stances on the mind/brain problem.

In 1870, Ewald Hering (of whom you will hear more later), laid out the classic finesse for of the mind/brain problem for visual science. The idea is to identify the mappings between physiological and perceptual states without assigning causality. Instead simply specify the correspondence between the two domains. Hering argues that we can set aside the philosophical problem, and get on with finding the lawful relationships (*mapping rules*) that hold between neural and perceptual states. This is a very sensible position and is appropriate for many of the purposes of this book. However, we will not entirely give up the materialistic idea of a causal story and will discuss a few examples with interesting arguments for causal relations.

## 1.4 Linking propositions

We now turn to the major philosophical theme of this book: the topic of *linking propositions.* Let's take the next logical step beyond Hering's assertion that lawful relationships – mapping rules – exist between visual perception and visual neurophysiology. Can anything more be said about the properties of these mapping rules? The question isn't just: do neural states map to perceptual states? It's which neural states map to which perceptual states?

### 1.4.1 Mueller's axioms of psychophysical correspondence

Interestingly, an elaborate set of mapping rules was explicitly formulated right at the beginning of the discipline of psychophysics. The 19th century scientists who founded the discipline were motivated not just by an interest in sensations and perceptions, but also by a desire to use perceptual observations as a tool for drawing inferences about the workings of the brain. They argued that perceptual (mental) events and brain (material) events were of two different kinds, described by language from two different realms of discourse. Therefore, if conclusions about brain events were to be drawn from facts about perceptual events, some kind of special linking statements would be needed. Their attempts to specify the necessary arguments were concisely formulated by G.E. Mueller in 1896, and are known as *Mueller's axioms of psychophysical correspondence* (Boring, 1942, p. 89).

Mueller's axioms have several important properties. Mueller called these linking statements *axioms* – statements that could not be proved, but that had to be assumed to be true if the discipline was to be pursued. With these axioms in place, one could use perceptual facts to deduce some aspects of the workings of the brain. Today we would be more likely to call these statements premises or assumptions. We argue that they are a special class of assumptions, and they enter into all claims about physiological-perceptual mappings in vision science. Moreover, they have a huge impact on the science we choose to do, and they govern the kinds of arguments that we entertain.

Mueller's first axiom states the very general premise that all perceptual processes arise from material processes, with the material process taking causal priority. However, it says nothing further about the forms these physiological-perceptual correspondences might take. So far, the possibility is left open that any material process, or brain state, could give rise to any mental process, or perceptual state. However, since the rest of the universe is lawful, it makes sense to assume that mappings between neural and perceptual states are lawful too.

The second and third axioms state several specific lawful relationships that might be assumed to hold between perceptual and neural states. The starting point concerns identity relations: he stated that identical perceptual states imply identical neural states, and vice versa. Further statements assumed that similar perceptual states imply similar neural states, and vice versa.

Statements like these are sometimes called *isomorphisms – similarities of form* – between neural and perceptual states. More specifically, we call the mapping rules in Mueller's second and third axioms *relational isomorphisms.* In a relational isomorphism, perceptual states are compared to perceptual states, neural states to neural states; and the isomorphism is that the same *relationship* that holds between perceptual states is assumed to hold between neural states (for a formal development of this view of theory see Coombs, Dawes, and Tversky, 1970). The propositions are that identical perceptual states imply identical neural states, and vice versa; similar perceptual states imply similar neural states, and vice versa; and so on.

The discussion of the possible limits on grating acuity illustrates the use of relational isomorphisms. If you review each of the four cases, you will find in each case the assumption that due to a particular spatial processing imperfection within the visual system, as the grating gets finer and finer, the spatial distributions of signals produced by the grating and the homogeneous field become more and more similar, and eventually become *identical.* The linking proposition that enters each argument is that an identity of neural processes creates an identity of perceptions; and vice versa, the identity of the two perceptions implies that the neural identity has been reached.

In general, relational isomorphisms are not logical necessities. Two identical perceptions could in principle arise from different brain states if the mappings from brain states to perceptual states were chaotic, or if they were many-to-one instead of one-to-one (Teller and Pugh, 1993). Other rules of relational isomorphism, such as those arising from similarity, are also easily challenged. The relational isomorphisms that enter into our beliefs about physiological/perceptual mappings are premises, not logical necessities. But they are certainly convenient!

### 1.4.2 Updating Mueller's axioms: Linking propositions

There has been surprisingly little work on the axioms of mental/material correspondence since 1896. Brindley (1960) treated the topic briefly, proposing the name *linking hypotheses* as a name for these rules of correspondence. Since they are rarely actually hypotheses, Teller (1980) suggested the name *linking propositions.* A linking proposition is defined as *a claim that a particular mapping* *occurs, or a particular mapping principle applies, between neural and perceptual states.* She argues that linking propositions lie at the heart of vision science. Any linking theory – any attempt to explain perceptual events on the basis of neural events, or vice versa – will necessarily include a linking proposition. Moreover, linking propositions are often implicit rather than explicit, and part of the fun of vision science is ferreting them out and examining them.

### 1.4.3 Analogies and nothing mucks it up

The term *isomorphism* has also been used in another way in the context of linking propositions (cf. Pessoa, Thompson, and Noe, 1998). We call this second category of isomorphisms *analogies.* These are isomorphisms that bridge between perceptual and physiological domains, by making an analogy between some aspect(s) of perceptual and neural states. For example, think about your perception of a set of broad black and white stripes, such as those in Figure 1.2A. In speculating on the neural state that underlies this perceptual state, you might assume (or show) that there is a region of the retina across which the firing rates of neurons take on a similar pattern – a set of neurons firing more slowly, say, to provide the neural correlate of a black bar, and a similar set of neurons, displaced by the equivalent of one bar width, firing more rapidly to provide the neural correlate of a white bar. The plots of whiteness/blackness against space, and firing rate against space, would look very similar. In other words, there is a visual similarity, or *analogy* between the perceptually and neurally defined patterns.

An interesting feature of analogies as linking propositions is that they often enter vision science as nothing more than that two pictures look alike. But typically, many pieces need to be added to the argument to make a compelling theory or explanation of the perception on the basis of the neural activity. In fact, Teller (1980) argues that such arguments must depend on implicit assumptions that she called "*Nothing Mucks it Up*" provisos. These assumptions are of the form that nothing within the visual system, between the neural pattern and the perception, interferes with the control of the particular neural pattern over the perception. In general, the earlier in the visual system the neurons on which the analogy is based occur, the more complicated and tenuous the required "Nothing Mucks it Up" provisos would seem likely to be.

### 1.4.4 From the sublime to the ridiculous

An interesting property of linking propositions is that they are often implicit. But once a linking proposition is made explicit, there is often a surprisingly good consensus among most vision scientists about its acceptability as a premise. At the one extreme lie some relational linking propositions, like Mueller's axiom of identity. Most vision scientists would probably regard this proposition as clearly true perhaps even analytically true or tautological (cf. Brindley, 1960). We will return to elaborate it further in Chapter 2.

At the other extreme, there are some candidate linking propositions that we would doubtless all reject. For example, we would probably be uncomfortable with the assumption that the neural code for seeing a three-dimensional object must be (literally) a neural circuit with the same three-dimensional shape as the object, somewhere within the brain. We would doubtless deny that a neuron that signals redness would have to be literally red (but then, what would a redness neuron have to be like?). And we would think it silly to argue that when we perceive a dance we must do so with dancing neurons, as shown fancifully in Figure 1.7.

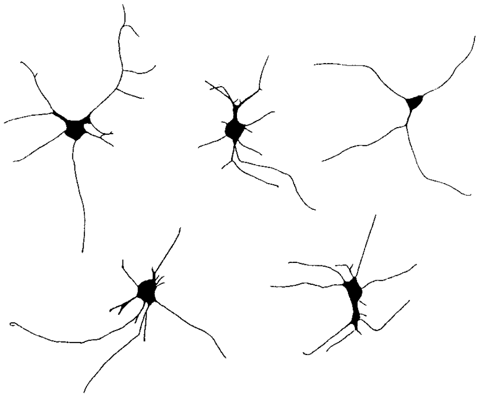


Figure 1.7: Dancing neurons? (From Cohen-Cory, Dreyfus, and Black, 1991)

These examples are chosen because they define the ends of a continuum from high acceptability to silliness. But the credibilities of other kinds of linking propositions, such as analogies and computational linking propositions, fall between the two extremes, and are worth some thought.

The moral of the story is this. Whenever a linking theory is proposed, a linking proposition lies within it. Is it a linking proposition that most vision scientists would readily accept, or readily discard? Would there be a consensus, or might different vision scientists differ in the kinds of linking propositions they are willing to incorporate as premises into their linking theories?

## 1.5 A more perceptual perspective

Historically, vision has been studied from two different perspectives. The first is a sensory perspective, in which vision scientists tend to emphasize the simpler aspects of vision, and (in general) attempt to account for them on the basis of the codings and recodings of information that take place within the early processing stages of the visual system. The second is a perceptual perspective, in which we tend to emphasize the more complex aspects of perception, and (in general) attempt to account for them on the basis of the more complex and higher-level aspects of visual processing.

So far in this chapter we have been taking a classically sensory approach to vision. But let's switch briefly to a more perceptual approach. From the perceptual perspective the interesting parts of vision science lie not in the sensory details such as grating acuity, but rather in the complex system properties of perception. And the most fundamental phenomena are not illustrated by drawings on the pages of a book, but by looking at the world around you. [Look at the world around you!]

As you look around, you see a scene that contains three-dimensional objects of particular sizes, shapes and colors in particular three-dimensional locations. These objects may move, but their essential characteristics of size, shape and color tend to remain constant across viewing conditions (that is, we as perceivers have good size, shape, and color *constancy*). You are often able to recognize objects across time and across contexts. The perception scientist wants to describe and quantify these more complex system properties, and to understand the properties of the neural processes that make these high-level features of perception possible. We will argue later, for exam-

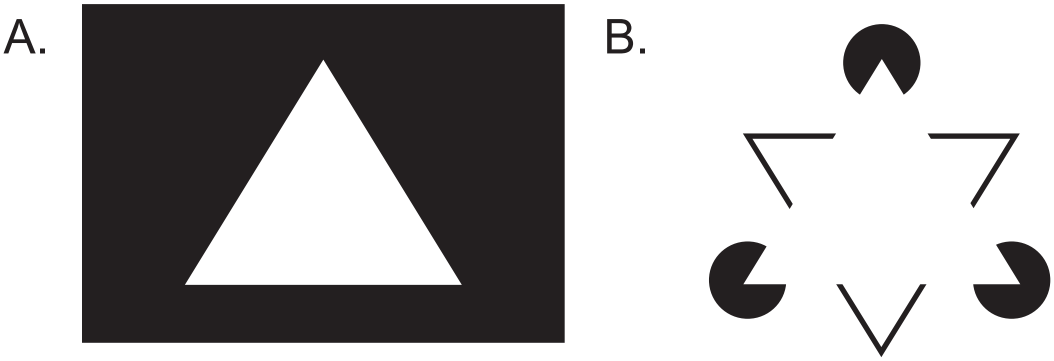


Figure 1.8: Illusory contours. A. A white triangle with a black background. B. A white triangle placed over an outline black triangle and three ”pac-men” with a white background. The continuous contours at the left and the illusory contours at the right both give rise to similar perceptions of a white triangle. Why?

ple, that incoming sensory information must be combined with stored information and processed with complex computational algorithms before it can provide the neural basis of high-level visual perception.

Until relatively recently, physiological study of high-level visual processing was still in its infancy. Since little relevant information was available, many perception scientists were not much interested in knowing the details of information processing within the visual system. In fact, some have denied the value of understanding anatomy and physiology for understanding perception. In terms of our introduction to the domain of visual science, these scientists have chosen to study only physical-perceptual mappings.

But this situation is changing. By now, a great deal of information about the anatomy and physiology of cortical processing is well established, as we will show. Moreover, in the past few years a number of vision scientists have undertaken studies of single neurons, searching for the neural basis of particular perceptual processes and phenomena. As this knowledge comes in, perception scientists are beginning to invent linking theories about perceptual events and processes based on neural events and processes. Moreover, recent, more global analyses arising from neural imaging techniques such as functional magnetic resonance imaging (fMRI), have also drawn the interest of perception scientists toward neuroanatomy and neurophysiology, and toward linking theory relating perception to neurophysiological activity.

### 1.5.1 Linking theories for perception

Let us develop an example of a linking theory in high-level perception. First, take the case of the triangles illustrated in Figure 1.8. This figure illustrates the phenomenon of *illusory contours*. The perception of a set of three borders can arise from at least two very different stimuli: three physical borders formed from solid lines; or a set of “pacmen” at three corner locations. And in both cases, similar triangles can be perceived. Suppose we were to decide to search for the neural cause of this kind of perceptual similarity. Where would we start, and why?

One linking proposition we could adopt would be a similarity (or identity) proposition: that

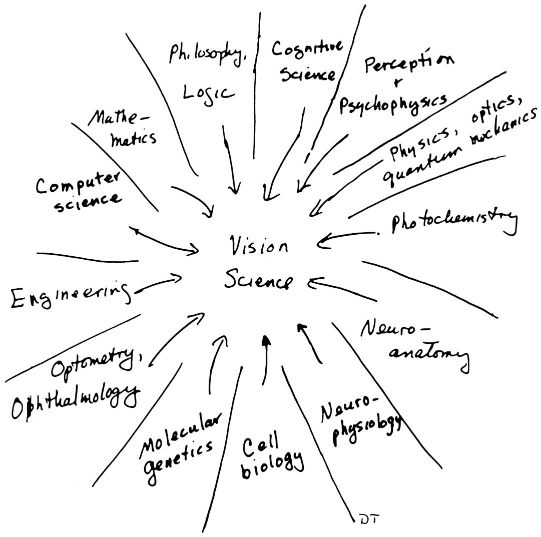


Figure 1.9: The major disciplines that contribute to vision science.

within the visual system, there will exist neural elements that respond similarly (or identically) to these two stimuli – the lines and the pacmen. Such a speculation will lead us to examine individual neurons at various levels of the visual system, and test them with both kinds of stimuli. The goal would be to try to find such neurons, and to locate the earliest level at which they occur. But the whole enterprise rests on a relational isomorphism: the premise that the two similar perceptual states – the two perceptions of a border – indicate the presence of two similar states of the same neurons arising from the two physical stimuli. The potential result of this line of questioning is a linking theory that accounts for the perception of illusory contours by a particular set of neurons that process all kinds of contours in the same way.

## 1.6 An interdisciplinary field

It should be obvious by now that vision scientists cannot afford to respect the boundaries between classical scientific disciplines, much less be chauvinistic toward any of them. Instead, we first define the questions of interest, as we have done in this chapter. Then we look around to see what kinds of classical disciplines can provide us with the expertise we need to address our questions. Specialists of many different kinds are invited – we need all the help we can get! The disciplines that unite to form the field of vision science, and the expertise we need them for, include at least the following.

**Psychophysics and perception:** To describe and quantify the system properties of vision.

**Physics:** To describe the nature of light and the optical quality of the eye.

**Photochemistry:** To describe the interaction of light with matter in the photoreceptors.

**Neuroanatomy:** To describe the structure of the various parts of our visual systems.

**Neurophysiology:** To describe the neural activity at each stage of the visual system.

**Cell biology:** To describe the internal workings of cells and their mechanisms of communication.

**Molecular genetics:** To describe the genetic control of the various parts of the visual system.

**Optometry, ophthalmology:** To describe the disorders of vision and the visual system.

**Engineering:** To provide conceptual and mathematical tools for describing the visual system.

**Computer science:** To discover the algorithms of information processing in the visual system.

**Mathematics, statistics:** To provide mathematical models for the three kinds of mapping rules.

**Philosophy, logic:** To provide logical analyses of our arguments (e.g. linking propositions).

**Cognitive neuroscience:** To provide models of the physiological processes that affect perception.

In fact, over the years specialists from all of these fields have been drawn into vision science. As a consequence, the field is conceptually rich and sophisticated, and new ideas are always arriving from different sources. For us, the excitement has lasted a lifetime. Everyone in the field comes from somewhere else; everyone has some areas of deep expertise, and some areas where he or she is an amateur.

## 1.7 Summary

In this chapter we defined vision science as the study of vision, the visual system, and the relations between the two. We argued that vision science spans three domains: physical stimuli, physiological states, and perceptual states; and that vision scientists are interested in the mapping rules among these three domains.

The questions in which vision scientists are interested were illustrated by using the example of *grating acuity*. We defined grating acuity as a physical-perceptual mapping. We then provided a brief description of physical-physiological mappings, mostly within the retina. Finally, we posed questions about a physiological-perceptual mapping. What stage or stages of visual processing limit grating acuity? We suggested four candidate answers. Each of the hypotheses about what limits grating acuity are examples of a physiological explanation or what we have called *linking theories*. Such theories will figure prominently throughout this book.

In addition, we have introduced an abstraction of the arguments within linking theories. Such *linking propositions* provide the philosophical underpinnings of the theories that relate physiology to perception. We will elaborate on such linking propositions as the book proceeds.

This book has had two broad goals. First, vision science is a complex interdisciplinary field, influenced by concepts from many kinds of science and by many kinds of scientists, and the beginning student is likely to have some difficulty with the parts that are the farthest from home. The first goal of this book is to provide a united, self-consistent set of tutorials, using simple examples, to make the science as a whole accessible. It is hoped that the tutorials in the various chapters of this book will slow down the moving train just enough for students with many different backgrounds to jump on.

Second, we have argued that vision science is a sophisticated discipline, spanning across physical, perceptual and physiological realms. To our knowledge there is no deliberate, consistent exposition of the forms of argumentation common in vision science. The arguments and premises are often implicit. It is hoped that making them explicit will demystify them, and thereby help students get on the train. And, of all the implicit elements, linking propositions are perhaps the most consistently hidden in the shadows, and therefore the most fun to bring out into the light. With luck, these efforts should encourage the seamless integration of sensory and perceptual approaches to vision science.

In Chapter 2 and 3 we examine the methodological tools with which vision scientists study physical-perceptual mappings (psychophysics), and a sample of the results they have found. In Chapter 4 and 5, we examine the optics of the eye, and the marks that the optical system leaves on our perceptions. Then in Chapter 6 and 7 we examine the workings of photoreceptors, both individually and in sets, and begin to analyze the code for color vision.

# Chapter 2

# Psychophysics: Identity Experiments

## Contents

*Psychophysics*[1] is the study of quantitative relationships between physical stimuli and perceptions. Alternatively, we can define psychophysics as the science of quantifying the system properties of vision: What and how well do we see?

Somehow, information originating from physical stimuli arrives in our retina, is encoded and recoded in our visual systems, and eventually maps to our perception of those stimuli. These mappings are remarkably regular and lawful – the same stimulus usually brings about the same perception – as they must be if we are to respond properly and consistently to stimuli and objects in the physical world. Our first concern in this chapter is how scientists quantify the relationships between physical and perceptual realms. To that end, we introduce some of the measurement techniques used in psychophysics.

Measurement techniques become more interesting as they are used to discover the system properties of vision. As soon as we introduce a measurement technique, we want to put it right to work. In this chapter, after we discuss psychophysical techniques, we will use them to define a new set of system properties having to do with the effect on our visual perception of variations in the *wavelength and intensity of light.* As you will notice, many of the questions are the same as they were for grating acuity, but translated into the realm of wavelength. What wavelengths of light can we see, and at what intensities? Can we tell different wavelengths of light apart?

[1] Teller once tried to define psychophysics for a physicist. He listened attentively for about five minutes, and then said, "Oh, I get it. But why don't you just call it crazy physics?"!

But first, how do we study the characteristics of human vision? Historically, the approach has been to present the stimuli to the subject via an optical system. Such optical systems allow the precise specification of light at the retina. They often allow careful control of the wavelength of light and the ability to vary the intensity over many orders of magnitude. However, they cannot display a wide range of spatial and temporal patterns. Alternatively, one can present the stimuli on a video display system. These systems have relatively limited range of light levels (2 orders of magnitude) and very limited freedom to vary wavelength. But they have the advantage of almost unlimited freedom in displaying spatial and temporal patterns. In either case, the psychophysicist varies the physical characteristics of the stimuli, and the subject reports what she sees, either by turning a knob or pressing a key. Now, more specifically, how do we quantify the lawful relationships between the physical stimulus and the subject's perception?

## 2.1 Two kinds of experiments: Identity and appearance

In 1960, in his now-famous chapter on linking hypotheses, Giles Brindley made a distinction between what he called *Class A* and *Class B* psychophysical observations. *Class A* observations are those in which a subject is asked to judge whether two perceptions are identical. Three common examples of identity experiments are: detection, discrimination and identity matching. In *detection experiments*, one judges the presence or absence of a stimulus. In *discrimination experiments*, one judges whether two stimuli are identical; in *identity matching*, one adjusts one stimulus to be identical with a second stimulus in all perceptual aspects. As we will see, these kinds of observations are measurements of *detection thresholds*, *discrimination thresholds* or *identity matches*, respectively. We use the term *identity experiments* as a more intuitive label than Class A experiments.

In contrast, *Class B* observations are those in which the subject judges how some aspect of his perception varies with the physical properties of the stimulus. Common examples include matching the brightness of two lights that can vary in color, comparing two lights in brightness when they differ in color, or naming the color of a light. By Brindley's definition, any judgment that depends on more than identity is considered Class B. We will focus on a subset of Class B experiments that are called *appearance experiments*: they depend on the appearance of the percepts internal to an observer and nothing else (e.g. external feedback). Class B experiments that do not depend on appearance (e.g. categorization of externally-defined categories) are beyond the scope of this book.

In the present chapter we confine our attention to identity experiments (Class A) – primarily detection and discrimination thresholds. Appearance experiments (subset of Class B) will be discussed in Chapter 3.

### 2.1.1 Detection and discrimination experiments: What is a threshold?

In everyday English, a *threshold* is a boundary between one thing and another – between the inside and the outside of a house, for example. In psychophysics, a threshold is the boundary between conditions under which a stimulus is seen and conditions under which it is not seen. The term threshold captures the idea that the transition from seeing to not seeing, like the transition from inside to outside a house, is relatively abrupt. But in fact, the precise place at which we should say one enters the house is slightly ambiguous. Is it the porch steps, or the front door, or halfway between? A visual threshold is similar – relatively abrupt, but with a small region of ambiguity that requires further consideration.

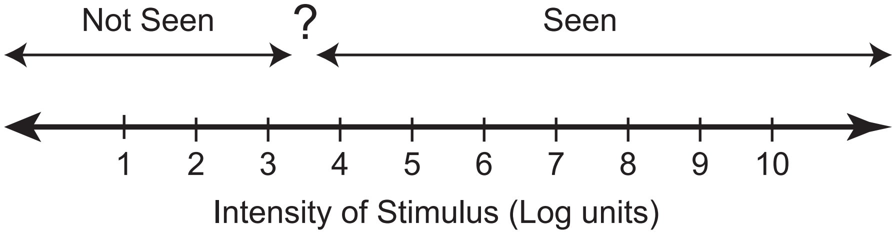


Figure 2.1: The threshold region. The horizontal line represents the intensities of a series of stimuli. The human eye can function over an enormous range of intensities, here represented as ten orders of magnitude (or 10 *log units*). Under a given set of conditions, however, the threshold – the boundary between seeing and not seeing – is remarkably narrow, perhaps a factor of two or three. In the diagram, the threshold region marked with a question mark, occupies the range between about 3.3 and 3.6 log intensity units (its width is 0.3 log units, or a factor of two).

Suppose that we arrange our laboratory equipment so that we can provide spots of light of many intensities, covering a range of (say) $10^{10}$, or 10,000,000,000 to 1 (a realistic estimate of the range of intensities that the human eye can handle)[2]. As shown in Figure 2.1, as we look at these different stimuli, letting our eyes adjust to changes of intensity as necessary, we will find informally that there is a large range at the low intensity end where we never see the test spot, and a large range at the high intensity end within which we always see it. In between there is a remarkably small region of uncertainty, within which we see the stimulus only some of the time. This region of uncertainty points to the location of the subject's threshold.

## 2.2 Classical psychophysical methods applied to detection

As psychophysicists, our first goal is to make quantitative estimates of thresholds. How shall we go about it? In the following paragraphs we give examples of three different psychophysical methods. In choosing a psychophysical method, there are at least three interrelated design factors: The *stimuli* to be used; the *task* the subject is asked to perform; and, the *responses* the subject is allowed to use. As will be seen, these three factors all vary among the three psychophysical methods we will describe. Many more combinations, of course, are possible and have been used.

### 2.2.1 The method of adjustment

The first and most intuitively obvious method is the *method of adjustment.* To apply the method of adjustment to light detection, we present the subject with a stimulus such as a spot of light. We

[2] Vision scientists generally plot the intensity of light in logarithmic units. This is because the range of intensities over which the visual system operates is enormous – about $10^{10}$ from the dimmest star to a snowy ski slope when the sun is shining. The use of logarithmic units is a convenient way to compress the range into something that is manageable graphically. It also allows easy comparison between threshold curves and sensitivity curves, as will be discussed later.

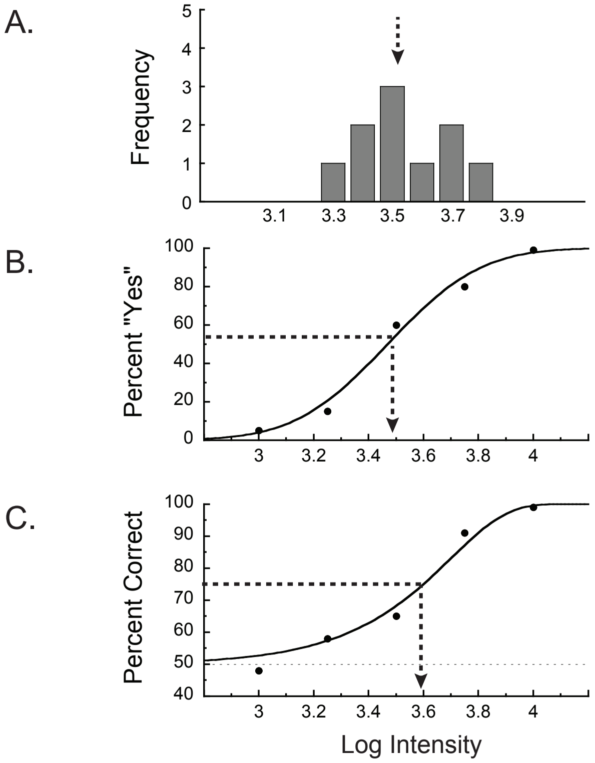


Figure 2.2: Illustrative data from three psychophysical methods. A. The method of adjustment. B. The Yes-No method of constant stimuli. C. The forced-choice method of constant stimuli. The arrows show the threshold estimates from each of the three methods.

give him a knob to turn or a computer key to press, and ask him to adjust the physical intensity[3] of the spot until he can just barely see it. The subject is asked to repeat the adjustment some number of times; say, ten.

Hypothetical results of a method of adjustment experiment are shown in Figure 2.2A. In this figure, the abscissa shows the physical intensity of the spot of light, on an expanded, arbitrary intensity axis. The ordinate shows the frequency with which the subject sets the light to each of the different intensities in a set of ten trials. Subjects do this task quite reliably – the range of intensities might typically span only about a factor of about two or three. The mean or median of the set of intensities is the response measure, and is used to characterize the intensity required for threshold – that is, to specify quantitatively the location of the threshold region along the intensity axis.

The main advantage of the method of adjustment is that it is quick and efficient – each setting might take, say, 10 seconds. Thus, ten settings of a single threshold might take a couple of minutes, and a set of ten thresholds could be measured readily in a 20 minute session. Because of its speed, the method of adjustment is extremely useful in preliminary work, or in cases in which large effects

[3] The term *intensity* has two uses in vision science. In this book, it is always used informally, to refer to variations in either the physical or the quasi-physical quantity of light (see Chapter 3). But it is also used technically, as part of a formal specification system for the physical intensity of light.

are being measured and/or only rough estimates of thresholds are required.

However, the method of adjustment has two major limitations. First, it leaves the definition of "seeing" up to the subject. That is, a liberal vs. conservative definition of seeing might well cause a difference in the measured threshold. One subject may set the threshold higher than another because the first subject will only say she "sees" the stimulus when it is clearly visible, while the second subject requires it to be only fleetingly so. Second, the subject can turn the intensity of the light up or down at will over any range she chooses. That is, the method of adjustment does not allow the experimenter to control the order of presentation of different stimulus intensities. If the immediate history of stimulation influences the detection threshold – and it does, as will be discussed in Chapter 10 – these variations will increase the variability of the measurements.

### 2.2.2 The Yes-No method of constant stimuli

A second approach, which allows the experimenter more control of the order of presentation of stimuli, is the *method of constant stimuli.* In this method, the experimenter pre-selects a set of stimulus intensities near where she thinks the subject's threshold will be. These stimuli are presented to the subject, one at a time, in random order, many times each. For example, the experimenter might decide to present the five stimuli 40 times each, for a total of 200 trials.

The experimenter's next decision concerns the choice of response measures and the subject's task. In what we will call the *Yes-No method of constant stimuli*, the experimenter asks the subject to use the responses "Yes" and "No". The subject's task is to say "Yes" (I saw the stimulus), or "No" (I didn't see it) on each trial. The use of many trials at each of several different intensities allows the experimenter to plot the percentage of "Yes" responses as a function of the intensity of the stimulus.

A hypothetical example of the kind of data one would obtain is shown in Figure 2.2B. A data set of this kind is called a *psychometric function.* In this example we have chosen our stimuli well, as the psychometric function spans the range from near zero to near 100% over the chosen stimulus range. By fitting an S-shaped curve to these data, we can estimate quantitatively the intensity at which the subject says "Yes" on, say, 50% of the trials, and define that intensity as the threshold value. As with the method of adjustment, the threshold value defines the location of the psychometric function along the intensity axis[4].

The Yes-No method of constant stimuli has the major advantage of giving the experimenter control over the stimuli used. In particular, in case the prior history of stimuli viewed influences the threshold – and it often does – the method of constant stimuli brings this history under control. It is not without its own problems, however. If each trial takes, say, 3 seconds, and 200 trials per threshold are required, measurement of a single threshold will take 10 minutes; and a set of ten thresholds will take two hours, compared to 20 minutes for the method of adjustment.

Moreover, even though the experimenter controls the order of stimuli, the subject is still in charge of deciding how he defines seeing. In fact, by changing the instructions to the subject, and thereby changing the subject's *criterion* of what "seeing" is, it is easy to move the psychometric function by small amounts along the intensity axis. If the instruction is "Be liberal – say "Yes" if

[4] Psychometric functions actually have four parameters: the threshold (or location along the abscissa), the slope (or steepness), and the upper and lower asymptotes (which are assumed to be 1 and 0 in Figure 2.2B, and 1 and 0.5 in Figure 2.2C). All of these parameters can be estimated if there are enough trials. Mathematically, psychometric functions have traditionally been fitted with a variety of S-shaped functions, including the cumulative normal (probit), logit, or Weibull functions.

there's even just a tiny flicker of something," the curve will shift toward lower intensities. If the instruction is "Be conservative – say "Yes" only if you see the stimulus very clearly," the curve will shift toward higher intensities. The experimenter can also shift the curve by providing different incentives or payoffs for saying "Yes", and in various other ways. In short, the experimenter controls the stimuli, but the experiment still confounds the sensory variable of threshold with the subject's cognitive criterion for saying "Yes" or "No".

### 2.2.3 The forced-choice method of constant stimuli

A third approach is the *forced-choice method of constant stimuli.* In this method, the experimenter again modifies the pattern of stimulus presentation. Rather than presenting only a single stimulus on each trial, the experimenter presents the stimulus in either of two alternative spatial or temporal positions. For example, the stimulus can be presented either on the left or on the right, or in a first or second time interval. This stimulus format allows us to change the subject's task: instead of asking him to say whether or not he sees the stimulus, thereby leaving the criterion for seeing up to him, we can require the subject to make a judgment as to whether the stimulus occurred in one place or another, or in one time interval or another. For example, in a spatial forced-choice technique, the subject is asked to respond "Left" if he judges that the stimulus was presented on the left, or "Right" if he judges that the stimulus was presented on the right.

A critical change of task is being made here. In this case, the task is not, "Did you see it?" – a question about the subject's perceptions – but rather, "In which location did it occur?" – a question about the state of the physical world. This change of task has two other intertwined and important consequences. First, there is a right and a wrong answer on each trial – the stimulus is always presented in one or the other position (or time interval). And second, since the judgments can be either correct or wrong, trial-by-trial feedback can be given to the subject. The subject's overall task is to maximize the percent of trials on which his answer is correct, and providing trial-by-trial feedback allows him to learn, over time, to be correct on the maximum number of trials.

Tasks that involve judging the state of the physical world can be called *objective*, or *externally-referred*, or *physically-referred* tasks. Our example of judging whether a light is on the left or right of fixation is such a task. Tasks that involve judging one's own perceptions can be called *subjective*, or *internally-referred*, or *perceptually-referred* tasks. An example we will discuss in the following chapter is to present lights of different wavelengths on either the left or the right of fixation and judge which of the lights is brighter. Because the lights are different wavelengths, there is no correct answer to this question. We will use the terms physically-referred and perceptually-referred for this important contrast between tasks.

Hypothetical results of a forced-choice experiment are shown in Figure 2.2C. As in Figure 2.2B, psychometric functions are plotted, but this time with the subject's *percent correct* plotted on the ordinate. Since the subject can get 50% correct by guessing, the psychometric function spans the range from 50% to 100%. The threshold in such an experiment is typically defined as 75% correct – halfway between chance (50%) and 100%.

Among the three methods discussed here, the forced-choice method of constant stimuli is the most logically elegant, in the sense that it brings the stimuli, the subject's task and the subject's criterion under the tightest possible experimental control. However, it is also the least efficient, as many more trials (and therefore more time) must be invested to estimate the location of the psychometric function. For statistical reasons, it takes two to three times as many trials to locate the threshold to a given degree of accuracy when the lower asymptote is at 50% than when it is at zero. So a set of ten thresholds, held to the same criterion of accuracy, would take perhaps six hours, compared to 20 minutes for the method of adjustment and two hours for the Yes-No method of constant stimuli.

A final note on terminology: unfortunately, the term *forced-choice* is not consistently used in the psychophysics literature. Sometimes the term is used very broadly, to refer to any experiment in which the subject is allowed only two responses (e.g. "Yes" and "No"). With this definition, the Yes-No method of constant stimuli – the second method defined above – is also a forced-choice method. Other sources use "forced-choice" less inclusively, to refer only to experiments in which the response is to specify which of two stimuli was presented: Stimulus A is present on some trials and Stimulus B is present on others (e.g. leftward vs. rightward direction of motion). In the most stringent usage (and particularly in Signal Detection Theory, as discussed below), the term "forced-choice" is reserved for experiments in which, within each trial, the stimulus is presented in either of two spatial positions or temporal intervals *and* the subject's task is physically-referred. This usage has the advantage of making it most likely that the specific models used in Signal Detection Theory will apply to the task. In this book we mostly follow the most stringent usage. But when an experimenter says that a forced-choice technique was used, the only way to find out what was actually done is to study the details of the paper.

### 2.2.4 Staircases and other adaptive techniques

There's one more trick worth knowing. With the method of constant stimuli, we choose the set of stimuli before starting the experiment, and we're stuck with it. If it turns out not to sample the actual psychometric function well, we will do a poor job of estimating the threshold; we may even have to start over. There is another set of psychophysical methods, in which the previous trials are used to determine the stimulus on the next trial. The earliest adaptive methods were called *staircase methods*, for reasons that become obvious below.

Suppose an experimenter is carrying out a Yes-No staircase experiment. An example sequence of stimuli are illustrated in Figure 2.3. The experimenter starts with a high intensity stimulus on the first trial. If the subject says "Yes", the experimenter *decreases* the intensity for the second trial. In this example, this pattern was repeated until on trial 5 the subject says "No". At this point the experimenter *increases* the intensity for trial 6. Depending on the subject's responses, the pattern of intensities goes up and down – hence the name staircase. [Could you use a staircase for a physically-referred forced-choice experiment? Why or why not?]

In more sophisticated versions of *adaptive techniques*, all of the trials up to trial $n-1$ can be collapsed into a psychometric function; the psychometric function can be fitted with a theoretical curve; and statistical rules can be used to select the optimal stimulus – the one whose use would yield the maximal information – to use on trial $n$. These adaptive techniques increase the efficiency of estimating the threshold, in comparison to the corresponding method of constant stimuli. There is also a cost in using these techniques. For example, if you use a technique optimized to estimate the threshold, the resulting measurements will be good for estimating the threshold but say less about other aspects of the measurement such as the shape of the psychometric function.

In sum, the choice of a psychophysical method depends upon one's needs for balancing speed and accuracy. For a "quick and dirty" estimate, the method of adjustment is the obvious choice; it's

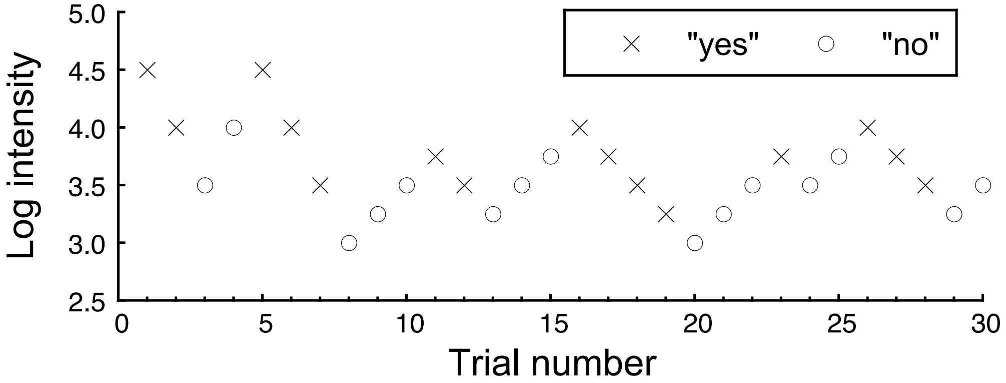


Figure 2.3: A "staircase". In this graph, the log intensity of the stimulus is shown for each trial (identified by its trial number). The experimenter starts on trial 1 by presenting a stimulus she judges the subject will always be able to see. The subject replies "Yes", marked by the X over trial #1. The experimenter then presents an 0.5 log unit dimmer stimulus. This pattern continues until trial #3, when the subject says "No", whereupon the experimenter reverses direction and presents (say) an 0.5 log unit higher intensity stimulus on trial #4. When the subject again says "Yes" on trial #5, the experimenter again an 0.5 log unit dimmer stimulus on trial #6.

fast but open to serious stimulus artifacts and potential criterion problems. For a more controlled but still criterion-limited method, the Yes-No method of constant stimuli may be the best choice. For maximum control, choose the forced-choice method of constant stimuli, but expect to spend a lot of time in the laboratory! And, if you know exactly what you want to measure, efficiency can be improved by the use of adaptive techniques.

## 2.3 Discrimination thresholds

Up until now, we have confined our discussion to detection thresholds. Is a spot of light present? In which of two positions is it located? *Discrimination thresholds* are a closely related concept. To measure discrimination, two stimuli are presented, and the subject is asked to tell them apart. Is one of the spots of light of higher intensity than the other? In which of two locations is the higher intensity spot located?

Each of the methods used for detection experiments (adjustment, Yes-No method of constant stimuli, forced-choice method of constant stimuli, and staircase methods) can also be used for measuring discrimination thresholds. For example, one could use the method of adjustment by setting one of the two stimuli to a fixed value, and having the subject adjust the intensity of the second stimulus until it looks just barely different from the first. [How would you measure a discrimination threshold with the Yes-No method of constant stimuli? With the forced-choice method of constant stimuli? With a staircase?] We will return to both detection and discrimination experiments below.

## 2.4 Explaining thresholds using signal detection theory

In 1966, David Green and J. A. Swets published a book describing their comprehensive theoretical analysis of detection thresholds, called *signal detection theory*. Their analysis suggests that our previous discussion of the concept of threshold is incomplete, and that a full account of thresholds requires two parameters. The first parameter, which Green and Swets called *d'* (d-prime), represents the *sensitivity* of the observer (or equivalently the *detectability* of the stimulus. This is the sensory parameter. The second parameter, which Green and Swets called $c$, represents the subject's *criterion* – a more cognitive variable. Green and Swets argued that a traditional Yes-No experiment, as described above in Figure 2.2B, confounds the effects of these two variables. That is, when the subject chooses to be liberal or conservative, his psychometric function shifts. Is this to be considered a change of the sensory threshold, or just a change in the subject's criterion? How can we tease the two apart?

To do so, let's modify our Yes-No method of constant stimuli experiment in two ways. First, instead of presenting five different stimuli interleaved, let's present just one of these stimuli – say, the middle one of the former five. And second, let's present the stimulus on only half of the trials (*stimulus trials*), and not on the other half (*no stimulus trials*). The subject's task is to judge whether or not the stimulus was presented on each trial. (Notice that this is a physically-referred, Yes-No task – a different combination of experimental design factors than was used in any of our previous three examples.)

The fundamental argument made by Green and Swets is that a detection task should be viewed as a *signal-noise discrimination.* Even on trials on which no stimulus is presented, Green and Swets proposed that the internal perceptual variable will have a non-zero value due to sensory *noise.* The noise arises from many internal and external sources, and the value of the noise fluctuates randomly over time.

The essence of signal detection theory is shown in Figure 2.4. Green and Swets began by assuming that the subject has access to an internal perceptual variable whose strength varies with the intensity of the stimulus. This variable is the evidence upon which the subject must make their decision. For the sake of concreteness, in the present case we can think of this variable as the perceived brightness of a spot of light. The key idea is that the value of this variable varies from trial-to-trial resulting in a distribution of values for a set of trials.

The hypothetical distribution of the perceptual variable for trials with no stimulus is shown by the graph in Panel A of Figure 2.4. These are trials in which only noise contributes to the distribution. For trials on which a stimulus is presented, the value of the perceptual variable is increased – say, a constant value will be added to the noise. The perceptual variable for trials with the stimulus present are shown by Panel B in Figure 2.4. These trials have both the signal and the noise contributing to the distribution.

The abscissa of Figure 2.4 represents the possible values of this internal variable, and the ordinate represents the probability of occurrence of each of the possible values which for a continuous distribution is called the *probability density*[5].

[5] For continuous distributions, rather than plot probability one must plot the probability density. This is the probability of an event falling into an interval of values relative to the size of that interval. For example, suppose the probability of the perceptual variable having a value between 1.0 and 1.1 was 0.05, then the probability density for that interval would be 0.5 (0.05/0.1). To get the probability density at a value of 1, one would effectively calculate the probability density for a very small interval near 1.

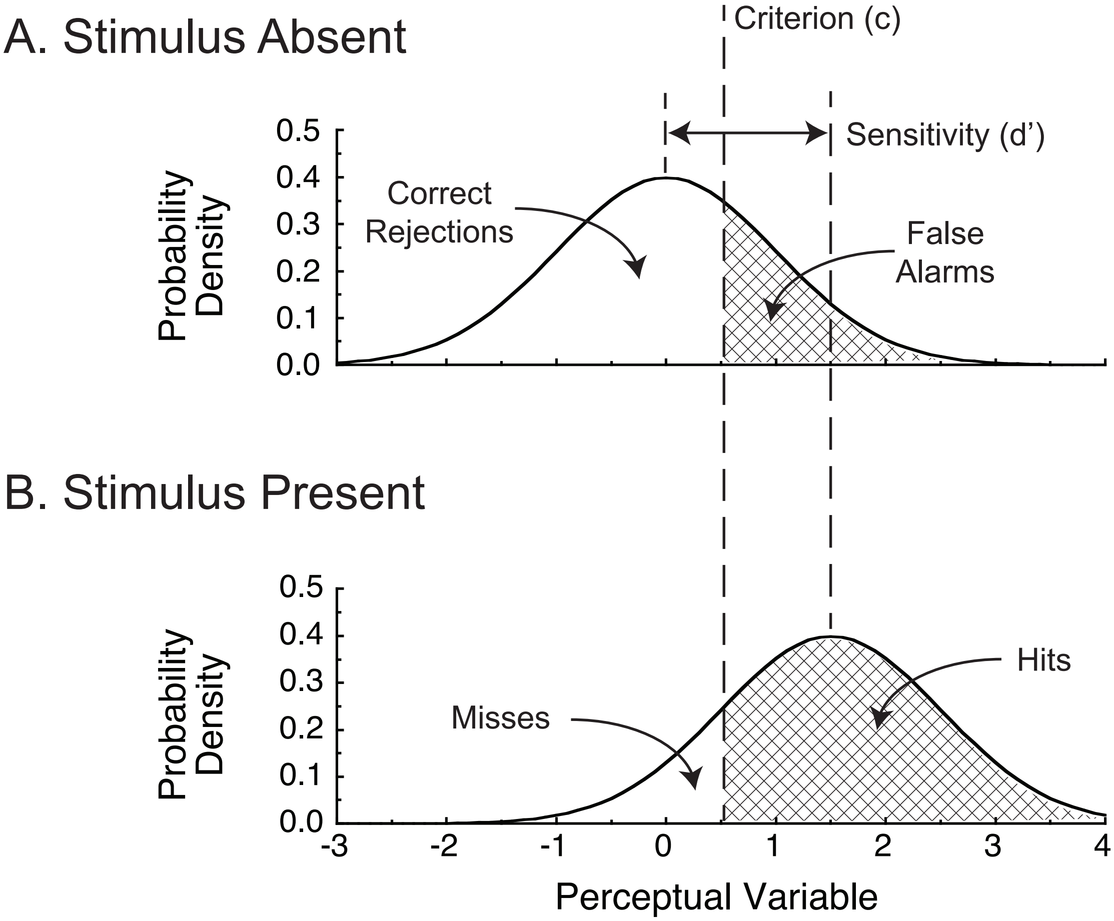


Figure 2.4: An illustration of signal detection theory. A. A graph of the underlying perceptual distribution when a stimulus is absent. This distribution is divided into correct rejections and false alarms. B. A graph of the underlying perceptual distribution when a near-threshold stimulus is present. This distribution is divided into misses and hits. For each graph, the abscissa shows the value of the relevant perceptual variable upon which the task depends and the ordinate shows the probability of each of the possible values of the perceptual variable (the area under these *probability density* distributions sum to 1). The figure illustrates the sensitivity and criterion parameters of signal detection theory. The sensitivity parameter defines the difference in location of the peaks of the two distributions. The criterion parameter defines what values of the perceptual variable results in a "no" or a"yes" response. For example, given the stimulus was present, values below the criterion result in misses and values above the criterion result in hits.

| | RESPONSE | |
|---|---|---|
| STIMULUS | "Yes" | "No" |
| Present | Hit | Miss |
| Absent | False alarm | Correct rejection |

Table 2.1: Contingency table for outcomes of a trial in a signal detection experiment.

The key assumption is that these two distributions lie on a single perceptual dimension: noise and stimulus both contribute to the value of the same internal variable and this is the variable upon which the response is based. Thus, there is no way for the subject to know whether a given value arises from noise alone or from the presentation of the stimulus in the midst of the noise. Nonetheless, the subject's task is to judge whether or not the stimulus was presented on each trial by saying, "Yes", the stimulus was presented, or "No", the stimulus was not presented. By the theory, the subject chooses a *criterion* value, shown by the arrow on the abscissa in Figure 2.4, and to say "Yes" if the value of the internal variable is above the criterion value, and "No" if it is below the criterion value.

The four possible outcomes of each of the trials of this experiment are shown in Table 2.1. On each trial the stimulus is either present or absent, and the subject's response is either "Yes" or "No". On trials on which the stimulus was presented, a "Yes" response yields a *hit*, and a "No" response yields a *miss*. On trials in which no stimulus was presented, a "Yes" response yields a *false alarm*, and a "No" response yields a *correct rejection*. The probabilities of these four outcomes are related to areas under the two curves in Figure 2.4. The subject's criterion forms the boundary between hits and misses on signal trials, and between false alarms and correct rejections on noise-alone trials. As the criterion is shifted leftward, the subject will say "Yes" on an increasing percentage of the trials, and generate more hits, but of necessity he will also generate more false alarms. By comparing the percentage of hits to the percentage of false alarms, and applying established theoretical formulas, the experimenter can estimate both *d'*, the subject's sensitivity to the signal, and *c*, the subject's criterion.

What if we now go back to a more classical method of constant stimuli, and think about stimuli of several different intensities rather than just one? The result is shown in Figure 2.5. The location of the stimulus-present distribution will vary along the abscissa with the intensity of the stimulus. As shown in Figure 2.5A, the lowest intensity of our three stimuli yields a stimulus-present distribution that overlaps nearly completely with the stimulus-absent distribution, so that the percentage of hits and the percentage of false alarms be nearly equal no matter what the criterion. But as shown in Figures 2.5B-C, the higher the intensity of the stimulus, the more the stimulus-present distribution will shift to the right, and the more the percentage of hits can exceed the percentage of false alarms. The parameter d' corresponds to the distance between the peaks of the stimulus-absent and stimulus-present distributions. The threshold region we first encountered in Figure 2.1 can now be seen to be the range of stimulus values that yield substantial (but not complete) overlap between the stimulus-absent and stimulus-present distributions.

In short, signal detection theory is a useful mathematical model because it provides a theoretical

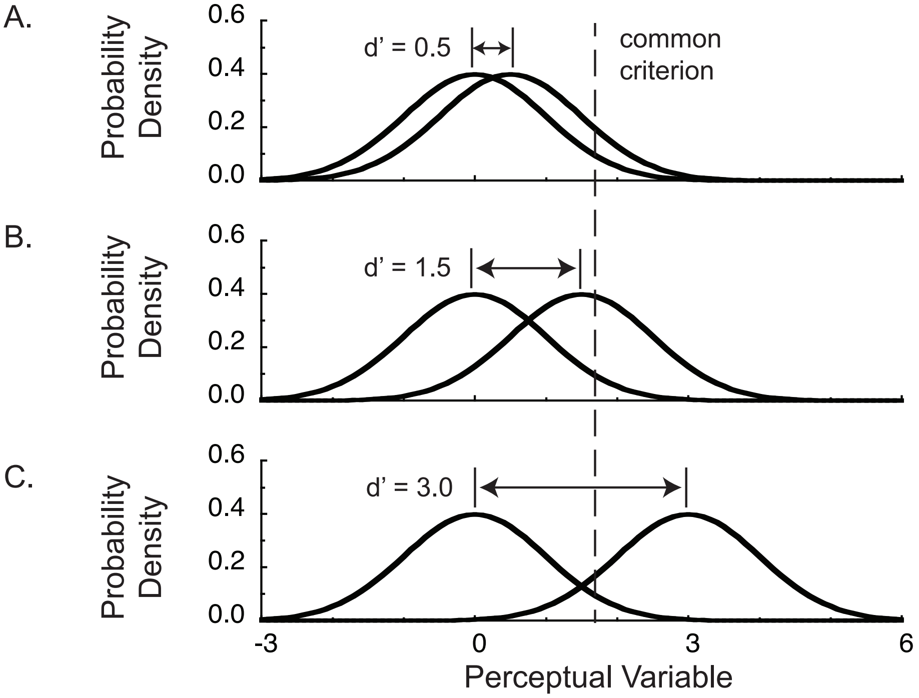


Figure 2.5: The effect of stimulus intensity on *d'*. The three panels show the effect of increasing the intensity of the stimulus from a low (A) to a high (C) value within the threshold range. The stimulus-absent distribution remains constant in all three panels. As the stimulus intensity increases, the stimulus-present distribution shifts rightward, and *d'* increases. One possible criteria is also shown. For this example, the subject has chosen to hold the false alarm rate constant at about 5%. The percent of hits and false alarms vary with $d'$ and $c$. If one assumes the shape of the two distributions (e.g. Gaussian), the value of hits and false alarms allows one to estimate of $d'$ and $c$.

account of several aspects of detection thresholds. It suggests a reason why observed psychometric functions are gradual rather than completely abrupt – because of the fluctuating value of noise from trial to trial. It also allows us to separate the sensory variable *d'* from the cognitive variable *c*, and provides us with an honorable way of arguing that the subject's adoption of a liberal vs. conservative criterion does not change the incoming sensory signal.

Since Green and Swets landmark book in 1966, signal detection theory has provided the basic conceptual foundation for our understanding of detection and discrimination thresholds. Many quantitative accounts of phenomena we will see later have been formulated in terms of signal-noise models, with the noise arising from sources both outside and within the subject's visual system. The basic model presented here has spawned a wide range of alternative models that make slightly different assumptions and can be applied to a wide variety of tasks. One can use these models to address a variety of issues in perceptual decision making. Due to space considerations, we will not spend much time on such models, but it's important to know that this theory is one of the critical foundations of psychophysical theory. For further reading, see the still very relevant original text of Green and Swets (1966) and the concise update provided by Macmillan and Creelman (2004).

## 2.5 Identity matching

Our third example of an identity experiment is *identity matching.* The situation is similar to a discrimination experiment in which one compares two stimuli. But now, the goal is to determine the pair of stimuli that are identical rather than just barely discriminable. For example, using the method of adjustment, one might adjust one or more attributes of a first stimulus until it appears identical to a second standard stimulus. At match, the stimuli need not be physically identical, but must be perceptually identical in all respects. Identity matching differs from discrimination in the question asked of the observer. For a discrimination task, the observer is asked to adjust the first stimulus until it looks just different than the second stimulus. For an identity matching task, the the observer is asked to adjust the first stimulus until it looks identical to the second stimulus. We return to identity matches when considering color matching in Chapter 7.

## 2.6 Animal psychophysics

In Chapter 1, when we discussed three types of questions, you may have noticed a potentially major limitation. That is psychophysical questions concern the system properties of visual perception: What and how well do we see? Obviously, psychophysical questions are usually studied by testing human subjects. In contrast, physical-physiological questions concern the properties of the visual substrate: the optics, photochemistry, anatomy, and physiology of the visual system. Since we can't do invasive experiments on human subjects, we usually study the substrate – particularly the physiology of single cells – in animals. And physiological-perceptual questions concern trying to explain system properties on the basis of substrate properties. But how can we assume that animals see as we do, or that our physiology works like that of the experimental animal we study? And if not, won't our theories always be potentially flawed?

Part of the answer to this question is yes. To some extent we have to live with this problem. But the other part is no. True, we can't do invasive experiments on human subjects; but we can in fact do psychophysics on animals. Many different species of animals have been tested successfully, including cats, fish, fruit flies, and (most importantly for our purposes) non-human primates.

Although other approaches are possible, the most straightforward approach to animal psychophysics is to focus on identity experiments rather than appearance experiments[6]. Begin by using a physically-referred task – a task in which there is a right and a wrong answer. Suppose you train a monkey subject to sit in a primate chair facing a stimulus screen and two response levers. On each trial of the experiment, the stimulus occurs either on the right or on the left; the monkey pulls the right or left lever, and he gets feedback – he is rewarded with a drop of water – for pulling the correct lever. If he pulls the wrong lever, he gets no reward. He may even get a "time out" – say a ten-second period during which no trials are run. If he's thirsty, this matters to him.

Initially the experimenter presents high-intensity, easily visible stimuli, and the monkey is trained to use the levers to respond right or left. Once the monkey gets, say, 90% correct or more over a series of trials, one uses lower and lower intensity stimuli, backing up to higher intensities if his error rate goes up, and progressing toward lower intensities again when he does well. Eventually his performance will stabilize, and he will generate a forced-choice psychometric function of the kind shown in Figure 2.2C, just as does a human subject.

[6] Appearance experiments with animals require a more indirect approach and will not be tackled in this book.

| | | | |
|---|---|---|---|
| 1. Initial proposition | A | → | B |
| *2. Contrapositive | not B | → | not A |
| *3. Converse | B | → | A |
| 4. Converse contrapositive | not A | → | not B |

Table 2.2: The general family structure of linking propositions.

When old world monkey subjects are tested, the resulting psychophysical data often resemble the data from humans very closely. The parts of vision science that involve relating physiology to human perception rely heavily on this similarity.

## 2.7 Linking propositions for identity experiments

We now return to the topic of linking propositions. In the same chapter in which Brindley (1960) introduced the Class A versus B terminology to refer to differences in psychophysical tasks, he also introduced the concept of what he called a *linking hypothesis*. Brindley was at the time a visual physiologist, and his question was, what is the role of psychophysical data in elucidating the physiology of vision? He argued that mental terms (perceptual terms, hence psychophysical data) and physiological terms were from different realms of discourse, and could not be used in the same sentence without introducing some form of special statements by means of which their meanings could be linked. Rather than referring to these statements as axioms, Brindley called them linking hypotheses.

Brindley further argued that most of the linking hypotheses in use in the vision science of his day were quite arbitrary, and that a careful physiologist would not want to pay attention to arguments that depended on them. Hence, he was ready to exclude most of psychophysics and perception from vision science. However, he found one linking hypothesis which he felt was rigorous enough and safe enough to include among the premises of his science, and by allowing its use he allowed detection and discrimination experiments (Class A experiments) to slip into the science. The linking hypothesis Brindley found acceptable was as follows:

"Whenever two stimuli cause physically indistinguishable signals to be sent from the sense organs to the brain, the sensations produced by these stimuli...must be indistinguishable" Brindley (1960, p. 144). In other words, identical incoming physiological signals must yield identical perceptions.

Teller (1984) elaborated on the concept of such linking statements. Since such statements are not usually hypotheses, she suggested they be called linking *propositions* – that is, statements that can potentially play many different roles in scientific argument, and whose truths and uses need to be individually evaluated. A *linking proposition* is *a claim that a particular mapping occurs, or a particular mapping principle applies, between perceptual and physiological states.*

### 2.7.1 Family structure

Teller (1984) also argued that relational linking propositions come in families, with different family members building on different experimental outcomes and allowing different directions of inference

| | | | |
|---|---|---|---|
| 1. Initial identity proposition | Identical $\Phi$ | → | Identical $\Psi$ |
| $*$2. Contrapositive identity | Non-identical $\Psi$ | → | Non-identical $\Phi$ |
| $*$3. Converse identity | Identical $\Psi$ | → | Identical $\Phi$ |
| 4. Converse contrapositive identity | Non-identical $\Phi$ | → | Non-identical $\Psi$ |

Table 2.3: The identity family of linking propositions.

between psychophysics and neurophysiology. Specifically, a family of relational linking propositions has four members that relate to each other as shown in Table 2.2. The family is composed of (1) an *initial proposition*; (2) its *contrapositive*; (3) its *converse*, and (4) its *converse contrapositive.* In abstract terms, the initial proposition (1) is: A implies B (if A is true, then B is true). The contrapositive (2) is: not-B implies not-A (if B is not true, then A is not true). The converse (3) is: B implies A (if B is true, then A is true). And the converse contrapositive (4) is: not-A implies not-B (if A is not true, then B is not true).

What are the relationships among these four statements? Students who have taken a course in logic will remember that the initial statement and its contrapositive – (1) and (2) –can be inferred from each other: if one is true, the other is also true. An example: if the initial statement is, *if it rains the sidewalk will be wet*, the contrapositive is, *if the sidewalk isn't wet, it didn't rain.* The same is true of the converse (3) and the converse contrapositive (4). Importantly, however, the converse (3) does not follow from the original statement (1): *if the sidewalk is wet, it rained* is not a logical inference, as the sidewalk could be wet for other reasons (perhaps the sprinklers are on)[7].

### 2.7.2 The identity family

Following this format and the inspiration provided by Mueller's laws, Teller (1984) then formulated several families of relational linking propositions. The first family, with which we are concerned in this chapter, is called the *identity family*, and is shown in Table 2.3. In order to use symbols that are more mnemonic for physiological vs. perceptual states, we will substitute the Greek letter $\phi$ (phi) for physiology and the Greek letter $\psi$ (psi) for psychology. Given that notation, we can use *Identical* $\phi$ for the A's in Table 2.2, and *Identical* $\psi$ for the B's.

With these substitutions, the initial identity proposition (1) becomes: *Identical physiological states imply identical perceptual states.* The contrapositive identity proposition (2) is: *Non-identical perceptual states imply non-identical physiological states.* The converse identity proposition (3) is: *Identical perceptual states imply identical physiological states.* And the converse contrapositive identity proposition (4) is: *non-identical physiological states imply non-identical perceptual states.* As before, the initial proposition and the contrapositive can each be inferred from the other, as can the converse and the converse contrapositive; but the initial proposition and the converse are logically independent.

[7] Teller's all time favorite example of the fact that an initial statement doesn't imply its converse arose when she was a graduate student. A fellow graduate student remarked one day that he didn't mind being misunderstood, because to be great is to be misunderstood. She undertook the delicious responsibility of pointing out to him that unfortunately, converses being what they are, to be misunderstood was not necessarily to be great.

Are the identity propositions analytically true or just likely? Are they relatively safe speculations, or risky ones? The initial identity proposition has the same content as the only linking hypothesis that Brindley found acceptable; that is, "Whenever two stimuli cause physiologically indistinguishable signals to be sent from the sense organs to the brain, the sensations produced by those two stimuli must be indistinguishable." Moreover, Brindley argued that his acceptable linking hypothesis is probably analytically true and tautological (it follows from the definitions of the other concepts involved). The contrapositive, being logically identical to the initial proposition, would also have the same logical status – true and tautological[8].

The converse and converse contrapositive, however, are not necessarily true. For example, the converse of Brindley's linking hypothesis would be something like this: Whenever the sensations produced by two stimuli are indistinguishable, these two stimuli must be sending identical signals from the sense organs to the brain. This statement is not necessarily true, because the signals could start out different in the retinal image but become identical at some later stage of processing; and because even if the neural states remain distinguishable, two different neural states could in principle map to the same perceptual state (a many:one mapping between neural and perceptual states could occur). In fact, with only one possible exception (see later), initial and contrapositive Identity propositions seem to be the only general linking propositions that are analytically true in simple systems, and therefore relatively safe to adopt as premises. All the rest are more risky tools.

We now turn to how linking propositions can be used in a theory. Notice that the second (2) and third (3) identity linking propositions allow physiological conclusions to be drawn from psychophysical experiments. These two propositions are starred in Table 2.3. Detection and discrimination experiments determine whether the sensations produced by two different stimuli are discriminable (non-identical), or not discriminable (identical). If two stimuli are discriminable in a psychophysical experiment, contrapositive identity (2) insists that these two stimuli have sent non-identical signals from the sense organ to the brain. Most people find this proposition difficult to doubt, as Brindley did. On the other hand, if two stimuli are not discriminable, the converse identity proposition (3) suggests that they have sent identical signals, or more sensibly, that the signals differed initially but were rendered identical at some later stage of processing. Thus, any identity psychophysical experiment allows us to draw a conclusion about the probable identity or non-identity of physiological states, depending on only the outcome of the experiment – which stimuli are discriminable and which are non-discriminable – and our willingness to employ either the contrapositive (2) or the converse (3) Identity proposition in our argument.

Finally, notice that these identity arguments are the same as some arguments we introduced less formally in Chapter 1. We argued initially that if a subject can discriminate between a grating and a homogeneous field, information about the spatial structure of the grating must be retained through all levels of the visual system. There is a contrapositive identity proposition embedded in the premises of this argument. Similarly, we argued that if the subject could *not* discriminate between the grating and the homogeneous field, information about the spatial structure of the grating must have been lost somewhere within the visual system. There is a converse identity proposition embedded here.

[8] While the identity proposition and contrapositive are logically consistent in a simple system, any specific application to physiology and perception may or may not be appropriate. For example, two physical stimuli may be identical, but the physiological system may add random noise to them which makes them not identical.

### 2.7.3 Application of identity propositions to identity experiments

How, exactly, do we apply identity propositions to the data from detection experiments, in which there is only one stimulus? Terminology can make this question confusing. The trick is that in this context, vision scientists think of the background alone as one "stimulus", and the background plus the test stimulus as the other "stimulus". Some intensities of the test stimulus are below the threshold region; these are marked NOT SEEN in Figure 2.1. In terms of signal detection theory, this means that the noise distribution is indistinguishable from the signal-plus-noise distribution. When such discrimination failures occur –when we are below threshold – a converse identity proposition will be included as a premise in any argument from perceptual data to physiological conclusions. We will conclude, slightly speculatively (as is always the case with converse propositions) that the two physiological states arising from background and background-plus-test-stimulus are indistinguishable.

Other stimuli are above the threshold region; they are marked SEEN in Figure 2.1. In this intensity region, the noise distribution is very different from the signal-plus-noise distribution. In this case, a Contrapositive Identity proposition will be included as a premise in our arguments. Since the perceptual states are distinguishable, we can conclude with confidence that the physiological states are distinguishable. Thus, just as the psychometric function marks the transition from not seeing to seeing, it marks the transition from using converse to using contrapositive identity propositions in drawing physiological conclusions from perceptual data.

For discrimination experiments, very similar arguments hold. When two suprathreshold stimuli are physically different but indiscriminable, we use converse identity any time we argue that the physiological states are indiscriminable. When the two stimuli are discriminable (as most pairs of stimuli are!) we use contrapositive identity in arguing that the physiological states are discriminable. The former is speculative, whereas the latter is (as Brindley said) more difficult to doubt.

As a final comment on linking propositions, consider again the definition of identity experiments. Typical identity experiments require comparisons between very similar stimuli. Often called near-threshold judgments, they also commonly involve judgments that have a correct answer. But these are not the defining characteristics. Instead, the defining characteristic of identity experiments is that they require the judgment of whether two perceptions are identical or not identical. Thus, the defining characteristic of an identity experiment is the identity relation at the heart of the identity proposition. Brindley's goal in defining Class A experiments was to distinguish the experiments that allow the application of the Identity linking proposition.

## 2.8 A psychophysical theory of scotopic vision

You have now been introduced to psychophysical methods for measuring thresholds and a logical analysis of the role of thresholds in identity propositions. The next question is what kinds of substantive questions can be investigated with measurements of thresholds? In this chapter we introduce psychophysical theories that use thresholds. In future chapters, these purely psychophysical theories will be elaborated with physiology into linking theories.

As it turns out, an individual threshold value usually has little in the way of theoretical implications beyond the general one of information retention and information loss. However, experiments in which *sets* of thresholds are measured often give important hints about physiological processes. In such experiments thresholds are measured as a function of some stimulus parameter or parameters. As an example, we now turn to an important set of thresholds: detection thresholds as a function of the wavelength[9] of the stimulus.

What we experience as "white" light usually contains a large range of wavelengths, the component wavelengths of which we experience as colors. As a student at Cambridge in the 1660s, Isaac Newton noticed a beam of sunlight coming through the shutters of his room. He passed the beam through a prism, and saw that the light now produced a rainbow of colors. That is, he had shown that sunlight can be broken down into its component wavelengths by passing it through the prism, which bends or *refracts* light differentially according to its wavelength. In the case of natural rainbows, internal reflections in water droplets act as the prism did for Newton.

As Newton showed, at moderate and higher light levels we can discriminate among lights of different wavelengths – different wavelengths map to different perceived colors. But you also may have noticed that if the light is sufficiently dim the colors fade away, and all that is left is shades of gray. [If you have not noticed this phenomenon, toss some shirts or towels of different colors around your room tonight, and see whether you can discern their colors when the room is nearly dark and your eyes have adjusted to darkness.]

In fact, vision turns out to have different properties at low and high light levels. The term *scotopic vision* refers to vision at low light levels, at which no colors are perceived, and the term *photopic* refers to vision at higher light levels, at which colors are perceived. We will spend the remainder of this chapter discussing two of the major properties of scotopic vision and end with a psychophysical theory of scotopic vision.

### 2.8.1 The scotopic spectral sensitivity curve

The *absolute threshold* is the smallest amount of light a subject can detect when an subject's eyes are fully adjusted to the dark. To begin our exploration of the effects of wavelength, we ask, does a person's absolute threshold vary with the wavelength of light?

The quickest way to answer this question is to use the method of adjustment with a series of stimuli that vary in wavelength. We use a calibrated light source so that the physical energy at any given wavelength is known. We put the subject in a dark room for an hour before we start. Then, we present each of the wavelengths in turn, and ask the subject to adjust the intensity of the light until she just barely sees the stimulus. The subject does this, say, ten times for each wavelength, and we take the mean of these ten settings. We then plot this threshold value as a function of the wavelength at which it was measured. Of course, more elegant psychophysical techniques could also be used.

The results of the experiment, plotted in terms of thresholds, are shown in Figure 2.6A. This data set forms what is called a *scotopic spectral threshold* curve. Alternatively (and more commonly), the data are plotted as *scotopic spectral sensitivity*, where sensitivity is defined as 1/threshold. With a logarithmically spaced ordinate, the conversion is particularly simple, because the threshold curve can simply be inverted to get the sensitivity curve. The same data plotted as sensitivity are shown in Figure 2.6C. We will use sensitivities rather than thresholds from now on.

This data set has several interesting features. First, the highest sensitivity is always around 500 nm (closer to 490 nm to be more exact). Second, sensitivity varies enormously with wavelength. Compared to the sensitivity at 500 nm, sensitivity declines by a factor of roughly 100 as we change

[9] The wavelength of light is usually specified in *nanometers* (nm). One nm is $10^{-9}$ meters. We will say more about the nature and specification of light in Chapter 4.

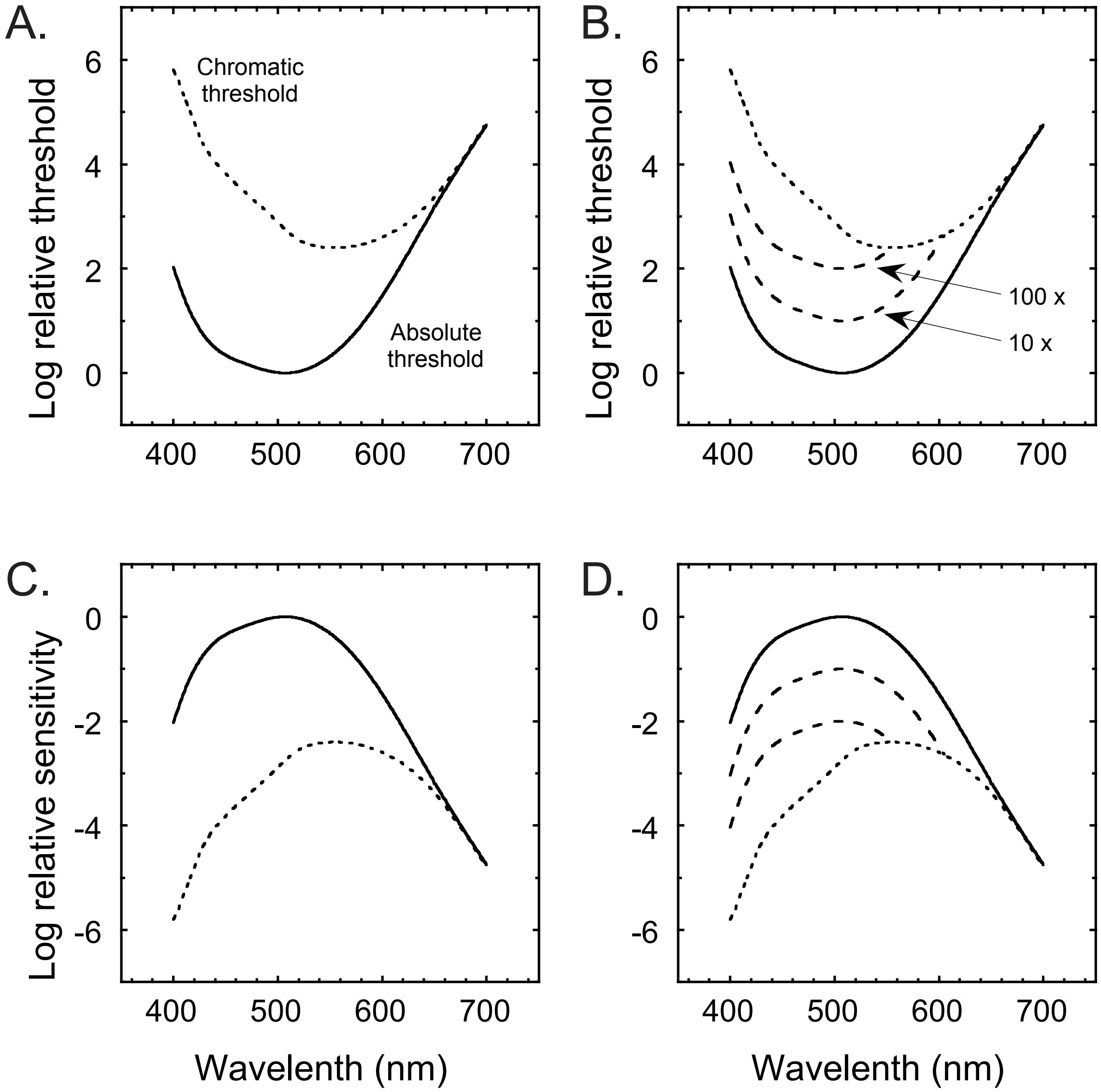


Figure 2.6: Spectral threshold (A, B) and spectral sensitivity (C, D) curves. A: Absolute thresholds and chromatic thresholds as a function of wavelength. The abscissa shows the wavelength of light; the ordinate shows the relative intensities of the lights of different wavelengths needed for absolute thresholds (solid curve), or for chromatic thresholds (dotted curve). B. Scotopic equivalence classes as equal multiples of thresholds. The dashed curves are shifted versions of the absolute threshold function. C. Absolute thresholds and chromatic thresholds plotted as spectral sensitivity curves. D. Scotopic equivalence classes plotted in terms of sensitivities. By convention, the maximum of a spectral sensitivity curve is labeled zero (0), with decreases in sensitivity away from the maximum labeled with negative log values (-1, -2, etc.).

the wavelength from 500 to 400 or to 600 nm; and by another factor of roughly 1000 as we change the wavelength to 700 nm. So the change in sensitivity for lights of 500 vs. 700 nm encompasses five orders of magnitude.

Third, the scotopic spectral sensitivity curve is a relatively simple U-shaped curve. And fourth, its shape is extremely stable. For example, it doesn't matter what psychophysical technique we use. All of the results reveal the same simple curve, possibly shifted up and down the ordinate, but of exactly the same shape. Similarly, backgrounds of various intensities and wavelengths shift the curve up and down; but over a broad range of conditions, the shape of the curve remains unchanged.

In sum, we have discovered a new system property of human vision. Absolute threshold experiments for lights of different wavelengths reveal a smooth, stable spectral sensitivity curve, with its maximum at about 500 nm, and with large and characteristic losses of sensitivity with changes in wavelength.

### 2.8.2 Failures of wavelength discrimination: Metamer sets

Here's a second set of system properties for scotopic vision. Suppose that you set up a row of test lights of different wavelengths, each one set to its own threshold, and ask the subject to discriminate among them. The subject cannot do the task at absolute threshold, where the lights are barely visible, but this seems hardly fair. So let's set the intensity of each stimulus to twice its detection threshold. The new stimuli are indicated by the lowest dashed line, marked 10x for "ten times threshold", in Figure 2.6B. The stimuli all look brighter than they did at absolute threshold, but they all still look whitish, and remarkably, the subject still cannot discriminate among them. Similarly, the stimuli along the second dashed line (100 times the absolute threshold), or any similar line in between, are indiscriminable, until we reach a limit (about to be described) for each wavelength of light[10]. (Stimuli from any two different dotted lines are discriminable because they vary in brightness.) In short, at scotopic light levels, wavelength information is lost. [How would one use identity matches to measure these indiscriminable lights?]

Vision scientists are so impressed with the fact that very different physical stimuli can be indiscriminable, that we use several special terms to describe this phenomenon. Such sets of stimuli have been called *equivalence classes*, emphasizing the idea that the signals arising from them must be rendered equivalent within the visual system. They are also called *silent substitution sets*, emphasizing the idea that one could be substituted for the other with no change of the neural signal. In addition, the term *metamers* is also used to describe lights of different wavelength compositions that are indiscriminable. The sets of lights indicated by each of the dashed lines of Figure 2.6 are *metamer sets.* The challenge, of course, is to explain why these losses of information occur.

### 2.8.3 A theory of scotopic vision: The funnel analogy

We next build a theory of how the system properties of scotopic vision arise from the physiological properties of the visual system. How to proceed? As psychophysicists, we are entitled to use the

[10] In the early days of science, what a challenge it must have been to sort out the effects of physical variables from the effects of our own sensory systems. What visual sensations would Newton have seen if he had used his prism in moonlight instead of sunlight? (Moonlight, of course, is reflected sunlight.) If the whole spectrum looked white, what conclusion would he have drawn about the nature of light? Might he have decided that light has different properties when it is dim, or comes from the moon? Or would he have placed the cause correctly, within his own visual system?

system properties of vision as a basis of theory and speculate about how the visual physiology might work.

Here's a theory of how the scotopic spectral sensitivity curve might come about. Suppose that there were an anatomical stage of the visual system composed solely of a set of identical elements (soon to be introduced as photoreceptors), and that each element had a spectral sensitivity curve matching that of the psychophysically defined scotopic curve in Figure 2.6. Under these assumptions, the system as a whole would necessarily show a spectral sensitivity curve that matches the scotopic curve. So we may choose to adopt the hypothesis that such elements, and such a processing stage, exist, and decide to go look for them.

But how shall we explain both the scotopic spectral sensitivity curve and the loss of wavelength information at the same time? It seems kind of odd – as though the system is influenced by wavelength yet loses wavelength information. But a gadget that would have the right properties can be created by combining a funnel with a counter (Figure 2.7). This analogy may seem a bit silly to those with some physical sophistication, but it will come in handy when things get more complicated later.

Let us assume we have an ordinary kitchen funnel. The funnel is a bit misshapen, being widest at the middle and narrowest at the bent corners. We add a counter to its output spout. To make the analogy, we metaphorically place the funnel under the wavelength scale of the scotopic spectral sensitivity curve, with the widest part at about 500 nm. Along the wavelength scale, we think of curtains of marbles of different colors, perpendicular to the page, raining down on the funnel at a specified rate. The probability that a marble of any given color will be caught is determined by to the width of the funnel at each particular wavelength. But once a marble is caught, it just rolls down to the spout of the funnel, and gets counted by the counter.

This analogy is also easy to express mathematically. Let $R$ be the total number of marbles caught by the funnel – the count on the counter at the base of the funnel. For each wavelength $\lambda$, let $Q_\lambda$ be the number of marbles that arrive per unit time in the curtain of marbles incident at the mouth of the funnel, and let $r_\lambda$ be the width of the funnel at that wavelength. The catch of marbles at any wavelength, then, is just the number of incident marbles of the color corresponding to $\lambda$, multiplied by the probability $r_\lambda$ that a marble of that color will be caught.

$$R = Q_\lambda r_\lambda$$

For several wavelengths – say, 450, 500, and 550 nm – presented together, we just add up the individual catches of marbles:

$$R = Q_{450}r_{450} + Q_{500}r_{500} + Q_{550}r_{550}$$

For the whole spectrum of wavelengths, we keep adding to get a final expression:
$R = \sum Q_\lambda r_\lambda$ , or for the continuous case, $R = \int Q_\lambda r_\lambda d\lambda$

The conditions that lead to metamerism in scotopic vision can now be formally stated. Suppose we have two patches of light, A and B, as shown in Figure 2.8. If the catches of marbles from the two patches are identical, the lights must be metamers. Why? Because the model system only has signals corresponding to $R_A$ and $R_B$. There's no way for the system to encode the information that the two physical stimuli are different. So by necessity:

$$\textit{If } R_A = R_B, \textit{ then } A \equiv B,$$

where $\equiv$ is the symbol for a metameric match.

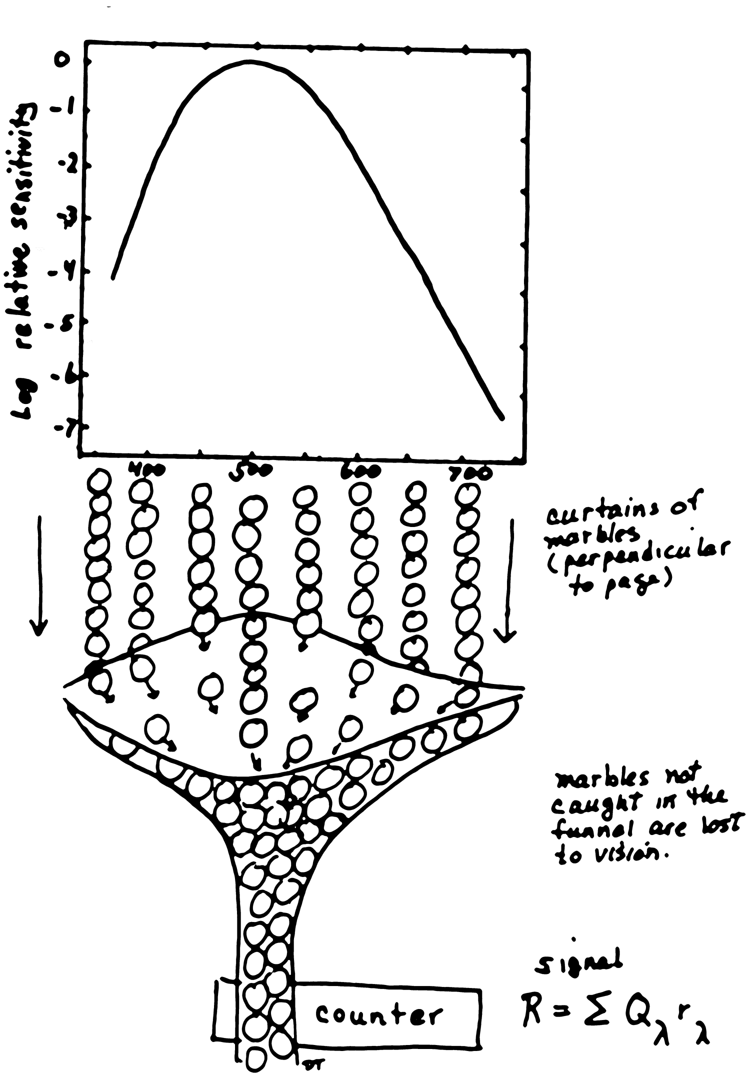

Log relative sensitivity
0
-1
-2
-3
-4
-5
-6
-7
400
500
600
700
curtains of marbles (perpendicular to page)
marbles not caught in the funnel are lost to vision.
signal
$R = \Sigma Q_\lambda r_\lambda$
counter
out

Figure 2.7: The funnel analogy. In this analogy the visual system is modeled by a funnel with a counter attached. Curtains of marbles, perpendicular to the page, rain down upon the funnel. The funnel varies in width. When the funnel is wide, it catches most of the marbles; when the funnel is narrow it catches only a small fraction of the marbles. This provides an analogy for the fact that sensitivity varies with wavelength. To complete the story, the counter only counts the marbles, providing an analogy for the loss of wavelength information that creates scotopic metamers.

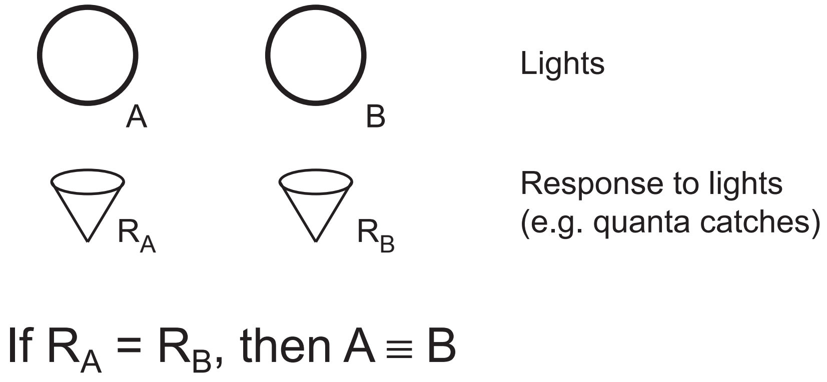


Figure 2.8: Scotopic metamers. A and B are lights of two different spectral compositions and intensities. If the intensities of A and B can be adjusted to make the two stimuli indiscriminable, A and B are called metamers. In the funnel analogy, lights A and B are indiscriminable because they make equal counts, $R_A$ and $R_B$, on the counter.

### 2.8.4 Univariance: A single value for both intensity and wavelength

This analysis introduces an important assumption about our theory of the scotopic sensitivity function. These physiological processes that respond to light are univariant. By *univariant*, we mean that they can be expressed by a single variable: one number. In the funnel and counter analogy, the only thing that mattered to the counter was the number of marbles. Wavelength information was lost. Similarly, in this mathematical model, the lights are represented by a single-valued function $R$. With a single univariant function, one cannot code for more than intensity. Wavelength information must be lost.

### 2.8.5 Next: Chromatic thresholds and photopic vision

Now, let's relax our testing techniques for a while, and just let the subject tell us what things look like (or look at them ourselves). Let's present two dim stimuli, a spot of 500 nm and a spot of mixed wavelengths that looks white, and set them both to intensities just above their absolute thresholds. The subject tells us they both look white, and equally bright. Now we double the intensities of both lights together, and keep doubling them, each time asking the subject to tell us what he sees. What will happen?

At first, the subject continues to say that both spots look white, with the brightness of both spots increasing together as intensity is increased. But after the intensities have been increased about a thousandfold (3 orders of magnitude), the subject suddenly says, "OH! The one on the left just turned green!" The intensity at which the 500 nm light changes from looking white to looking green defines the subject's *chromatic threshold* – the threshold for the onset of color sensation – for 500 nm light. We have passed from the scotopic to the photopic realm.

This experiment can be repeated for each wavelength. Interestingly, the intensity range between the absolute threshold and the chromatic threshold varies with wavelength. The lowest absolute threshold occurs at about 500 nm. But, as shown by the dotted curve in Figure 2.6, the lowest chromatic threshold occurs at about 555 nm, and chromatic thresholds increase as the wavelength increases or decreases from this value. Moreover, the shape of the chromatic threshold curve is not a simple U – it is asymmetrical, with a shallower rise at shorter wavelengths.

In sum, as we turn up the intensity of each light past a critical level, the system properties change. The *photopic spectral sensitivity curve* has its maximum at about 555 nm, and is more complex in shape than was the scotopic curve. It is also labile – its shape is affected by the method and conditions of measurement, as we will see in the next chapter. And above all, we become able to discriminate among different wavelengths of light – scotopic vision gives way to photopic vision: wavelength information is preserved, and color vision occurs.

## 2.9 Summary: Identity experiments and scotopic vision

In this chapter we have concentrated on the kinds of identity experiments that Brindley called Class A – detection, discrimination and identity matching. We described four examples of methods for measuring thresholds, and reviewed some of the advantages and limitations of each. We also introduced signal detection theory which provides a mathematical model that allows the separation of sensory from cognitive parameters.

Next, we returned to the concept of a linking proposition. We reiterated the claim that arguments from psychophysics to physiology, or vice versa, will always involve linking propositions. We elaborated on a family of relational linking propositions – the identity family – that deal with information retention and information loss. We argued that identity propositions enter into physiological conclusions based on detection and/or discrimination data. The properties of the identity family set the stage for examination of other, less intuitively obvious linking propositions in future chapters.

Finally, we used sets of threshold measurements to define two system properties of scotopic vision. Sets of detection thresholds were used to define a scotopic spectral sensitivity curve, and sets of discrimination failures were used to define metamer sets among supra-threshold but still scotopically detected stimuli. These two system properties may be seen as intuitively contradictory – the scotopic visual system is influenced by the wavelength of light, yet loses all information about it. In any case, these system properties are in need of physiological explanation.

We then considered a possible psychophysical theory of scotopic vision. We made use of our psychophysicist's speculation license to design a gadget, described by the funnel analogy, that would mimic scotopic vision. In fact, neural elements with the right characteristics will emerge without warning within the next several chapters. They will provide the additional ingredient to build a linking theory out of this psychological theory. Keep the system properties of scotopic vision in mind, and when you think you spot the neural elements that explain them, make a note of them (but unless you are Archimedes, do not jump out of the bathtub and run up the street yelling "eureka!").

In the next chapter we turn to psychophysical techniques for studying the appearance of visual stimuli using examples from photopic vision.

# Chapter 3

# Psychophysics: Appearance Experiments

## Contents

Chapter 2 begins the discussion of psychophysics and psychophysical techniques. We introduced the distinction between identity and appearance experiments, and discussed identity experiments at length. Examples of such experiments include detection, discrimination and identity matching. Our examples emphasized measuring detection and discrimination thresholds for various wavelengths and intensities of light. In these experiments the perceived colors of the lights might or might not have been changing with light level and wavelength. But in order to stay within the domain of identity experiments, we didn't ask the subject about these qualitative variations – a sort of "don't ask – don't tell" mentality applied to visual perception. As you may have noticed, one appearance observation did creep into the discussion, when the subject exclaimed, "Oh, the one on the left just turned green!" as we reached the chromatic threshold for a 500 nm light. In a strictly identity experiment, the most we could do would be to show that above the chromatic threshold the subject could discriminate the 500 nm light from lights of other wavelength compositions, but we wouldn't ask him what color it looked, and he wouldn't tell.

But suppose we are specifically interested in the perceptual qualities of lights that are clearly detectable, and discriminable from each other. Suppose we want to map wavelength to perceived color, or set lights of different wavelengths to be perceived as equally bright, or characterize the supra-threshold similarities and differences among perceived colors. These questions simply cannot be directly addressed with identity experiments such as detection. We will have to bite the bullet, and start asking subjects what they actually see. Thus we turn to appearance experiments.

One way of highlighting the distinction between identity and appearance experiments is to break down the difference between them into interrelated parts. As discussed in Chapter 2, many identity experiments can be seen as externally referred – the subject's task concerns a judgment about the physical stimuli so that the subject's judgments can be either correct or wrong, depending on the state of the physical world. And, since the judgments can be either correct or wrong, trial by trial feedback can be given to the subject. It can be argued that feedback provides a mechanism whereby the subject can learn more and more about exactly how the experimenter is defining the task.

In contrast, in an appearance experiment the judgments are internally referred – the subject's task concerns a judgment about how things appear to her. What color do lights of 450, 500, 550, and 600 nm look? Second, therefore, there are no wrong answers; provided that the subject is telling the truth, we have to take her responses as correct at face value. And third, since her judgments cannot be wrong, no meaningful feedback can be given. It wouldn't make sense to say "correct" after every trial. We can give the subject descriptions and instructions, but we can't use feedback to further specify the task.

Viewed in this light, we can see that identity and appearance experiments have tradeoffs of advantages and disadvantages. The most logically elegant techniques for identity experiments – forced-choice techniques with feedback – embody questions about the physical world, and have an objectivity to which appearance experiments can never aspire. But all that such identity experiments can reveal is the discriminability of stimuli – for some of us, an impoverished topic at best! On the other hand, appearance experiments can be attacked for their subjectivity. But we get to set aside thresholds, and study what many of us came to study in the first place – the qualities of the subject's perceptions. Of course, given that we choose to do appearance experiments, we will want to develop and use the most rigorous possible techniques, as one would in any branch of science, and some examples of these techniques will be introduced in this chapter.

In the present chapter, we discuss two major cases of appearance phenomena in the context of color vision. Case 1 deals with the brightness aspects of the appearance of different wavelengths, and Case 2 with the color aspects of the appearance of different wavelengths.

## 3.1 Case 1: The brightness of different wavelengths

In physics, the intensity of a spot of light of a single wavelength is specified in terms of its radiant energy, $E$, in *watts*. But suppose that the physicist wants to specify the total physical energy in a spot of light made up of many different wavelengths. He can specify the radiant energy of each individual wavelength, $E(\lambda)$, but how do they combine? If things were complicated, different wavelengths could be like apples and oranges, and the energies might not follow a simple combination rule. But in fact the physicist is lucky, because the measurement system works if the total energy, $E$, in a light of mixed wavelengths is specified to be just the sum of the energies at each of the individual wavelengths; that is,

$$E = \sum E(\lambda).$$

In other words, radiant energy is *additive* across wavelength. Additivity, of course, is a necessary

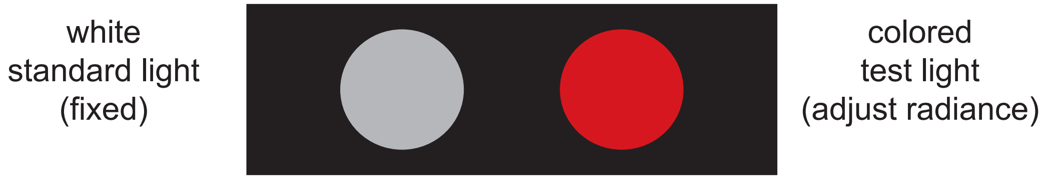


Figure 3.1: Heterochromatic brightness matching. The experimenter sets the radiance of the white standard light (appears grey in the context of this figure), and sets the colored test light to each of a series of different wavelengths in turn (appears red in this example). The subject's task is to vary the radiance of the colored test light, to match its perceived brightness to the fixed white standard.

condition for well-behaved units of measurement – if 2 inches plus 3 inches doesn't yield 5 inches, we are all in trouble.

But now suppose we want to specify the amount of light coming from a light source in terms of its *effectiveness for human vision.* For example, as a practical problem, suppose we want to design an airport runway system, using lights of different colors to mark different runways. What intensities shall we use? A physical specification, such as the energy of each bulb in watts, is not useful, because radiant energy near the middle of the visible spectrum is orders of magnitude more effective for vision than is radiant energy at the spectral extremes (see Figure 2.6). And radiant energy outside the visible spectrum, by definition, has no visual effectiveness at all.

In such cases it is useful to specify stimuli in *quasi-physical* units; that is, in units based on their effectiveness for human vision. The appropriate quasi-physical units would weight the radiant energies of the different wavelengths, $E(\lambda)$ by some psychophysical measure of the "visibility" – call it $V(\lambda)$ – of the different wavelengths for the human eye. And ideally, the units would be additive: the total visibility, $V$, of a light of mixed wavelengths would be just the sum of the energy at each wavelength weighted by the visibility of light at that wavelength:

$$V = \sum V(\lambda)E(\lambda).$$

Can $V(\lambda)$ specify the "visibilities" of lights of different wavelengths, to allow additivity to prevail? Or are different wavelengths of light like apples and oranges when human vision is involved?

### 3.1.1 Heterochromatic brightness matching

The most obvious approach to the problem is simply to ask the subject to look at lights of different wavelength compositions, and match them in brightness. This is the idea of *heterochromatic brightness matching.* The experimental set-up is shown schematically in Figure 3.1. We set up two spots of light: a *standard* light – say, a patch of white light at a fixed radiance – and a *test* light of the same size. To use the method of adjustment, we would set the test light to each of a series of different wavelengths, and for each wavelength have the subject vary the radiance of the test light to make brightness matches between the two. This is an appearance experiment because we are asking the subject to judge one perceptual property – brightness – while ignoring the other property – color. An identity experiment requires that all perceptual aspects of a stimulus are matched.

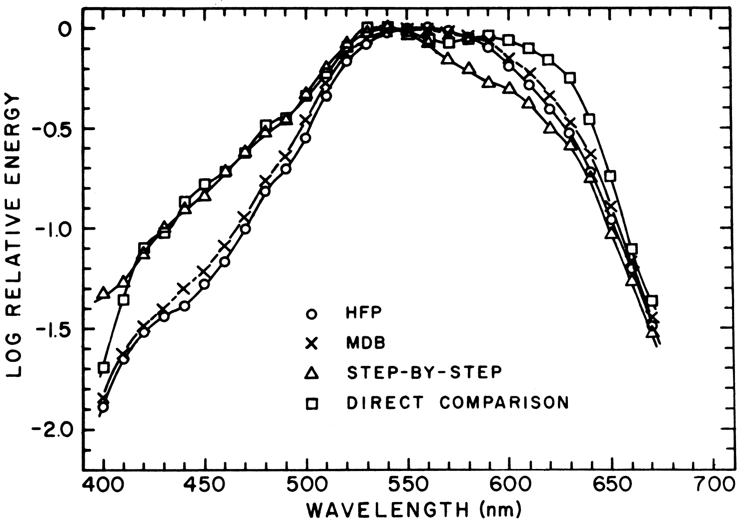


Figure 3.2: Photopic spectral sensitivity curves. The ordinate shows the relative radiance (relative physical energy in log units) required for a given perceptual match. For example, all of the curves show that matching a 650 nm light to a 550 nm light requires about 10 times as much energy (1 log unit). The results of heterochromatic brightness matching to a white standard are shown by the open squares (called "direct comparison" in the figure). Also shown is a variation on brightness matching called the "step-by-step" method. Of primary interest are the open circles and the crosses that show results from two photometric techniques: heterochromatic flicker photometry (HFP) and minimally distinct border judgments (MDB). The data from flicker photometry and minimally distinct border judgments agree well, whereas the data from brightness matching yield a broader curve. The data are from Wagner and Boynton (1972) and the graph is from Pokorny, Smith, Verriest, and Pinckers (1979, p. 25, Fig. 2.1).

A spectral sensitivity curve resulting from a heterochromatic brightness matching experiment is shown by the circles in Figure 3.2. Notice that the maximum sensitivity – the minimum energy required for a brightness match to the fixed white standard – is no longer near 500 nm, the sensitivity maximum we saw for scotopic vision. Instead, it shifts to about 555 nm, and the curve falls off sharply toward both shorter and longer wavelengths.

### 3.1.2 Problems with brightness matches: Variability and non-additivity

Unfortunately, heterochromatic brightness matching is plagued by two problems. First, subjects find the task difficult, because the two lights that are to be matched in brightness differ so much in color. There is considerable variability from trial to trial within a session and between sessions for a single subject, and also considerable variability among subjects. We interpret these problems to mean that there is no *natural perceptual criterion* – no perceptually striking event that happens right at the brightness match – to guide the subject's performance. Lights of higher radiance look too bright; lights of lower radiance look too dim; nothing looks quite right; and the subject is on her own to interpret the instructions and do the best she can.

A second and more devastating problem is that the resulting values are non-additive. A test of

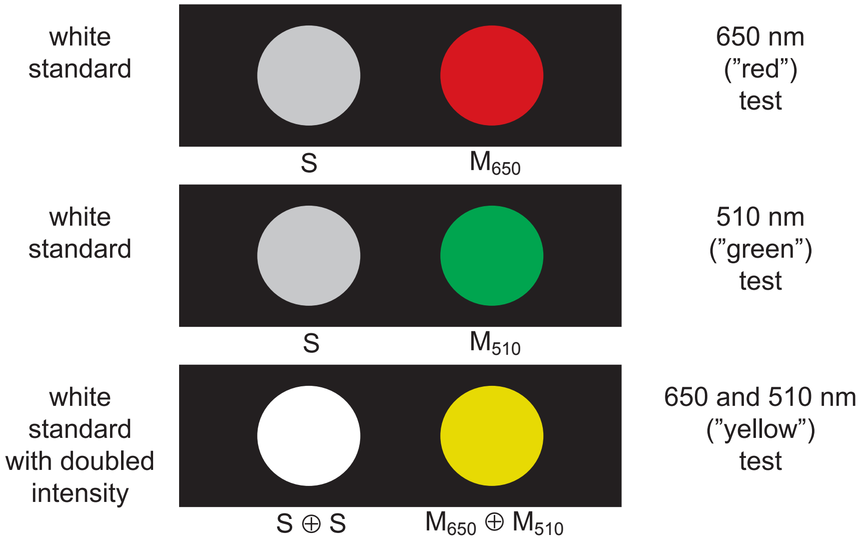


Figure 3.3: A test of the additivity of heterochromatic brightness matches. The subject first matches the brightness of a 650 nm light to that of a white standard light, to yield a matching value – call it $M_{650}$. He then matches the brightness of a 510 nm light to the same white standard light to yield a matching value $M_{510}$. The experimenter then superimposes $M_{650}$ and $M_{510}$, and asks the subject if this combination matches the white standard with twice the intensity. (The symbol ⊕ stands for superposition.) The subject's answer is no – the combined test field looks dimmer than the doubled standard. Brightness is not additive. (The colors in this figure are a crude simulation of the intended wavelengths.)

additivity is shown schematically in Figure 3.3. Suppose that the subject sets the intensity of a 650 nm light and a 510 nm light to match the brightness of a white standard. Now, we combine the 650 and 510 nm lights, and compare them to the white standard with the doubled intensity. Subjects report that now the test and standard lights no longer match in brightness; the superimposed 510 and 650 nm lights look dimmer than the doubled white standard. These failures of additivity can be as much as a factor of five (e.g. Kaiser and Wyszecki, 1978; Burns, Smith, Pokorny, and Elsner, 1982). The radiant energies of lights of different wavelengths sum additively, but their perceived brightness do not. Crazy physics indeed!

The non-additivity of heterochromatic brightness matching data makes this technique unacceptable as a basis for equating lights for visual effectiveness. We are left with our original practical problem: how to develop a set of quasi-physical units that specify lights in terms of their effectiveness in human photopic vision, yet are additive across wavelength.

## 3.2 Photometric methods

In about 1900, vision scientists had the insight that perhaps a solution might come from varying the psychophysical *task*. That is, perhaps a different task would yield additive values. In pursuit of this goal, several methods, called *photometric methods*, have been developed. We will discuss three of these – heterochromatic flicker photometry, minimally distinct borders, and motion photometry

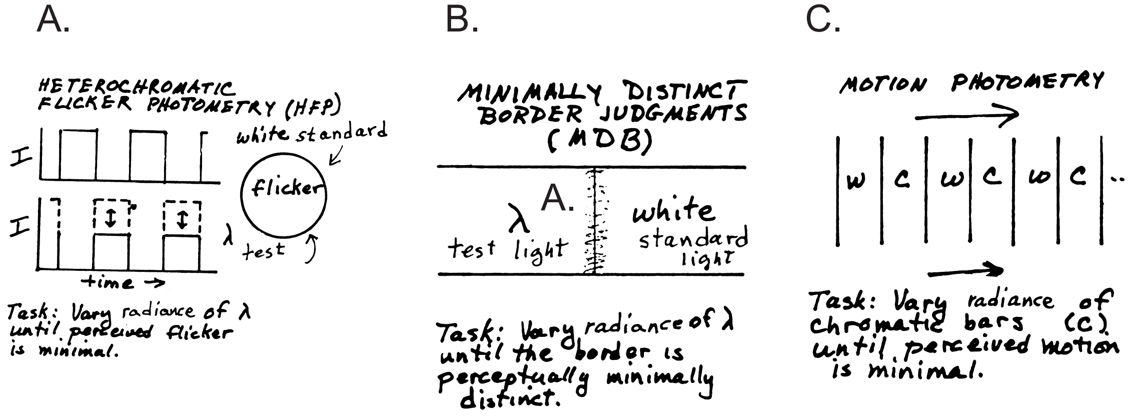

A.
HETEROCHROMATIC FLICKER PHOTOMETRY (HFP)
white standard
flicker
I
I
λ
test
time ->
Task: Vary radiance of λ until perceived flicker is minimal.
B.
MINIMALLY DISTINCT BORDER JUDGMENTS (MDB)
A.
λ
test light
white
standard light
Task: Vary radiance of λ until the border is perceptually minimally distinct.
C.
MOTION PHOTOMETRY
w c w c w c ..
Task: Vary radiance of chromatic bars (c) until perceived motion is minimal.

Figure 3.4: Three photometric methods. A. Flicker photometry. B. Minimally distinct border judgments. C. Motion photometry. In each case, the subject's task is to vary the radiance of the test light to set some aspect of her perception – perceived flicker, or the perceived distinctness of a border, or perceived motion – to a minimum.

– because they provide interesting examples of variations of the subject's task; and because, as it turns out, they each yield an additive system. Notice that in each case, the subject's task is to set two intensities to yield a perceptual minimum – in perceived flicker, in the perceived distinctness of a border, or in the amount of perceived motion – at a pair of relative intensities at which there is no physical reason to expect one.

### 3.2.1 Heterochromatic flicker photometry

The oldest of the photometric methods is called *heterochromatic flicker photometry*, or flicker photometry for short. As shown in Figure 3.4A, we again choose a white light as the standard, and use each of a series of test lights of different wavelengths. We arrange to alternate the stimulus in time between the standard and test lights, at a rate of, say, 15 cycles per second (15 Hertz, or Hz). Perceptually, the light appears to flicker between white and the color of the test light. But as we vary the radiance of the test light over an appropriate range, remarkably, the sensation of flicker diminishes, passes through a minimum, and then increases again. The subject's task in flicker photometry is to vary the radiance of the test light until the sensation of perceived flicker is minimal. We repeat the experiment for each different wavelength in turn.

It turns out that with a little practice, subjects can do this task remarkably consistently, both from one day to the next and from one subject to the next. We take this consistency to suggest the presence of a natural perceptual criterion in this task. A distinct perceptual event – a minimum in perceived flicker – is seen reliably at a particular radiance of the test light, and guides the judgments, allowing a high degree of consistency both within and between subjects.

The results of using the flicker photometry technique are shown along with the heterochromatic brightness matching data in Figure 3.2. As was the case with heterochromatic matching, the maximum of the flicker photometry curve falls near 555 nm. The two curves are similar in shape, but they differ in detail, with sensitivity at both extremes of the spectrum being higher with heterochromatic brightness matching than with flicker photometry. And the matches made with flicker photometry are additive.

### 3.2.2 Minimally distinct border judgments

A second method of photometry is called the *minimally distinct border (MDB)* method. The minimally distinct border paradigm is shown in Figure 3.4B. The stimuli are a white standard and a series of test lights of different wavelengths, juxtaposed at a sharp border. For each wavelength, subjects report that at a particular narrow range of relative radiances, the border between the two fields loses its perceptual sharpness – it becomes relatively indistinct. For some wavelengths, the border even appears to melt completely away, so that perceptually the colors of the two fields blend or smear together across the center of the stimulus field. The subject's task is to set the radiance of each test field to yield a minimally distinct border against the fixed standard light.

A spectral sensitivity curve generated by these judgments is also shown in Figure 3.2. The curve generated by minimally distinct border is highly similar to the curve found with flicker photometry. In addition the results are highly consistent, suggesting that a distinctive natural criterion – the minimal sharpness of the border – occurs reliably at one particular radiance of each wavelength. And, most importantly, the matches are found to be additive

### 3.2.3 Motion photometry

A third method of photometry, called *motion photometry*, is shown in Figure 3.4C. In motion photometry, a subject is shown a set of stripes of two different wavelength compositions (say, stripes of a white standard alternating with red, blue or green stripes produced by one of the three phosphors on a video monitor)[1]. The grating is set in motion across the face of the monitor. The radiance of the white bars is fixed, while the radiance of the chromatic bars is varied. As it varies, surprisingly, the perceived velocity of motion slows or even stops, and then speeds up again. The subject's task is to vary the radiance of the chromatic bars until the perception of motion is minimized. Again, the perceptual event is very distinct, and again the relative radiances required for motion minimization resemble those of flicker photometry and minimally distinct border methods. And, again, the values found are additive.

In summary, the problem of equating photopic lights of different wavelength compositions in visual effectiveness has been attacked with several tasks: match the brightness, minimize the perceived flicker, minimize the perceived distinctness of a border, or minimize the perception of motion. The last three tasks produce estimates of visual effectiveness that are very similar. Moreover, all three are additive. These results are important because they have allowed vision scientists to develop a quasi-physical stimulus specification system that is both reasonably similar to the results of heterochromatic brightness judgments, but, at the same time, are additive.

From a purely physical point of view, the reason these tasks work is mysterious. There is no physical reason why perceived flicker, or perceived border sharpness, or perceived velocity of motion, should go through a distinctive minimum at some particular relative intensities of the test

[1] Motion photometry is usually done on a video monitor because this is an easy way to create the pattern of moving stripes. Since the phosphors of a video monitor produce only broadband stimuli, spectral sensitivity curves cannot be measured directly. However, the motion minima are predictable from flicker or minimally distinct border data, by adding up the visual effectiveness of the different wavelengths emitted by the phosphor, as determined with flicker or minimally distinct border methods.

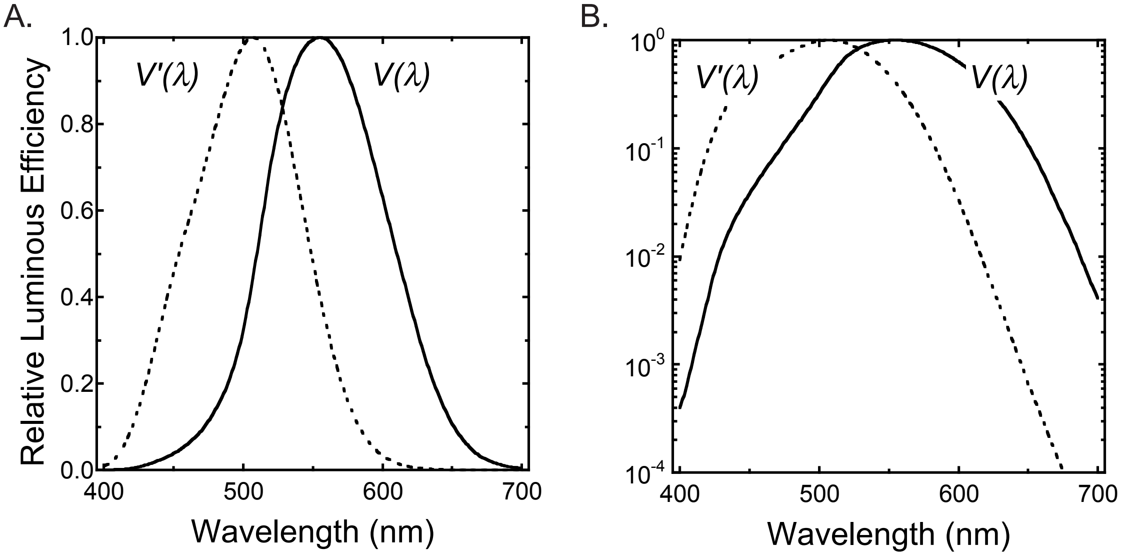


Figure 3.5: Standard photopic and scotopic spectral luminosity functions. The photopic function $V(\lambda)$ is plotted as a solid curve and the scotopic function $V'(\lambda)$ is plotted as a dotted curve. A. $V(\lambda)$ and $V'(\lambda)$ are plotted with a linear ordinate. B. $V(\lambda)$ and $V'(\lambda)$ are plotted with a logarithmically scaled ordinate. Notice how the linear ordinate shows the detail near the maximum of each curve, but obscures it in the tails of the function; whereas the logarithmically scaled ordinate compresses the values near the sensitivity maximum, but displays the sensitivity falloff in the tails. [Graphs based on the tables in Wyszecki and Stiles (1982).]

and standard lights. Moreover, there is no obvious reason why the quasi-physical specification systems they generate should be additive. More system properties in search of explanation.

## 3.3 The specification of light: Physical versus quasi-physical units

### 3.3.1 The standard spectral luminous efficiency functions $V(\lambda)$ and $V'(\lambda)$

The additivity of photometric matches has allowed the development of an additive quasi-physical stimulus specification system for photopic vision. Figure 3.5 shows the now-standard photopic spectral sensitivity curve, or more formally the *photopic spectral luminous efficiency curve*, $V(\lambda)$, together with the corresponding *scotopic spectral luminous efficiency curve* $V'(\lambda)$. These curves were adopted as standards in 1924 and 1951, respectively, by the International Committee on Illumination, or more properly the Committee Internationale de L'Eclairage (CIE). They are widely used in industry, to approximate the radiances of lights of different wavelengths needed to produce lights of equal effectiveness for human vision.

Since the measurements that enter into the CIE photopic spectral luminosity curve are additive, the total photopic effectiveness of a light of mixed wavelengths can be specified in this system by multiplying the energy at each wavelength $E(\lambda)$ by the visibility at that wavelength $V(\lambda)$, and adding up the visibilities of its components, as we had hoped. That is, for photopic vision, we can now legitimately define the *photopic luminance*, $L$, of a light of mixed wavelength composition by:

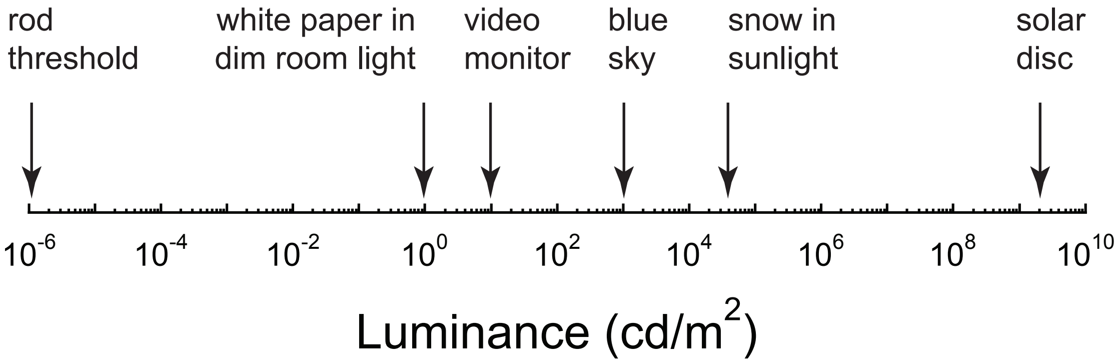


Figure 3.6: Approximate luminances of some familiar objects, specified in candelas per meter squared (cd/$m^2$). Luminance is a quasi-physical specification of light level, in which the radiance at each wavelength is multiplied by the sensitivity of the human visual system at that wavelength. [Following Rodieck (1998, p. 152).]

$$L = \sum V(\lambda)E(\lambda).$$

The *luminance* of a light, then, is its intensity specified in units of its effectiveness for human photopic vision. A similar equation can be written for scotopic vision.

### 3.3.2 Radiometric versus photometric units

Implicit in the above discussions is a distinction between two different ways of specifying the intensities of lights. *Radiometric units* are purely physical units – they specify light of any specific wavelength or wavelength mixture in units of physical energy (watts). In contrast, *photometric units* are quasi-physical units that weight the energy at each wavelength by the sensitivity of the human eye at that wavelength. Terminologically, words with the root *rad* (*radiance, irradiance, radiant flux*, etc.) are used in specifying light levels in radiometric units. Terms with the root *lum* (*luminance, illuminance, luminous flux*, etc) are used in specifying light levels in photometric units. The curve $V(\lambda)$ shows the multiplicative factors needed at each wavelength to convert radiometric to photometric units at photopic light levels.

Over the years many different sets of photometric units, with different names, have been used, but these need not concern us here. Suffice it to say that luminance is now usually specified in units of *candelas per meter squared (cd/$m^2$)*. The typical luminances of a few familiar surfaces in cd/$m^2$ are shown for reference in Figure 3.6.

A brief mention of one other aspect of the specification of lights. It is important for various reasons to distinguish between the amount of light *falling on* a surface, and the amount of light *emanating* from the surface; these are specified in different kinds of units. Terms with the prefix *ir* or *il* (*irradiance, illuminance*) refer to the former, whereas terms without this prefix (*radiance, luminance*) refer to the latter. If the surface reflects, say, half of the light falling on it, then its luminance will be half of the illuminance falling on it, assuming that properly corresponding units are used.

| | Wavelength max | Smooth over wavelength? | Stable over tasks? | Discrimination among wavelength's? |
|---|---|---|---|---|
| Scotopic vision | 500 nm | Yes | Yes | No |
| Photopic vision | 555 nm | No | No | Yes |

Table 3.1: The properties of scotopic and photopic vision.

We can now return to another practical problem – that of specifying the visual effectiveness of a light bulb. Light bulbs are traditionally specified in *watts* – a physical unit. That is, a 500 watt light bulb is being specified in terms of the energy it uses, but not directly in terms of its effectiveness for human vision. If it is specified in *lumens* it is being specified in photometric units, and its effectiveness for human vision is built into this specification. [Would it make sense to specify light in photometric units when testing the visual sensitivity of a goldfish? Why or why not? What would you do instead?]

### 3.3.3 Estimating retinal illuminance with Trolands

Our final comment on light measurement is to consider how much light reaches the retina. This is not necessarily proportional to the light in the corresponding part of the world because the size of the pupil varies. The common approach is to estimate the retinal illuminance of a region of the retina by taking the product of the luminance for the corresponding part of the visual field and the area of the pupil. The unit of retinal illuminance is called the *Troland* after its originator Leonard Troland. A Troland is defined as the conventional retinal illuminance when a surface with luminance of 1 $cd/m^2$ is viewed through a pupil with an area of 1 $mm^2$. This unit of light is particularly relevant to the study of light and dark adaptation (see Chapter 10). For a more technical treatment of the units of light and photopic spectral luminosity functions, see Pokorny and Smith (1986). For the definitive handbook on the topic, see Wyszecki and Stiles (1982).

### 3.3.4 Summary: Scotopic versus photopic vision

With our photopic spectral sensitivity curve in hand, we are ready to compare the system properties of scotopic and photopic vision. These properties are summarized in Table 3.1. The main features are these: the scotopic spectral sensitivity curve has a maximum sensitivity at about 500 nm, whereas the photopic maximum is at about 555 nm. The scotopic curve is a simple, smooth and relatively symmetrical U, and is remarkably stable over variations of measurement techniques and conditions, whereas the photopic curve is more complex and more labile as will become clear in future chapters. In addition, wavelength information is lost at scotopic light levels (no color), but preserved at photopic levels (color!). Finally, with spectral sensitivities specified in accord with $V(\lambda)$ and $V'(\lambda)$, the effectiveness of lights can be measured in an additive way for both scotopic and photopic vision.

In Chapter 2, we argued that the funnel analogy captures the essence of two system properties of scotopic vision: sensitivity varies with wavelength, yet wavelength information is lost. In a similar vein, what can we infer, or what speculations might we be drawn to, by the differences in properties between scotopic and photopic vision summarized in Table 3.1?

## 3.4 Linking theories of photometry

### 3.4.1 Theories linking physiology and luminance

In G. E. Mueller's time, before neurobiological techniques were available, one of the goals of psychophysics was to use the system properties to deduce the nature of the underlying physiological processes. Having read only the first few sections of this book so far, you are in the unique position of being like the early psychophysicists – you know quite a lot about the system properties of scotopic and photopic vision, but little about the physiological processing that gives rise to them. So think about the following "bumblebees can fly" arguments through your soon-no-longer-to-be naive eyes. Take the following paragraphs as exercises, and try to sort out strong from weak theories, and speculations from guesses.

The differences in spectral sensitivity lead to some new constraints on theory. For example, the difference in the wavelength for maximum sensitivity for $V(\lambda)$ and $V'(\lambda)$ implies that scotopic and photopic vision arise from at least partially different physiological processes with different spectral characteristics. The simplicity and stability of the scotopic curve $V'(\lambda)$ encourages the speculation that the neural basis of scotopic vision consists of a single process with a fixed spectral sensitivity curve that is not affected by task variables. In contrast, the relative complexity and lability of photopic spectral sensitivity suggests that it might be produced by combining inputs from several different physiological processes with different spectral sensitivity curves.

The differences in the loss versus preservation of wavelength information also give us clues about physiological processes. The failure of wavelength discrimination in scotopic vision suggested a neural process that doesn't preserve wavelength information, like the counter on the funnel in the funnel analogy. The success of wavelength discrimination in photopic vision implies a set of neural processes that do preserve wavelength information – it's a strong inference that a single funnel with a counter will not do. What mechanisms might be involved? [Might several funnels do the trick? Think about this.]

What about additivity? We did not mention it previously, but scotopic vision is additive, as is implicitly assumed as a feature of the counter in the funnel analogy. But we know that the single funnel model must be discarded for the photopic system, since wavelength information is preserved. So how do photometric methods remain additive? One can speculate that both the inputs to each of the putative processes underlying the photopic spectral sensitivity curve (see above), and the combination of signals across these processes, are additive. In contrast, the non-additivity of heterochromatic brightness matching is consistent with a more complicated signal combination rule.

Finally, it is a interesting question why the tasks that underlie the three photometric techniques – flicker, minimally distinct border, and motion photometry – are feasible psychophysical tasks at all. The sources of flicker, border, and motion minima do not lie in the physics of the stimuli, but rather within the human visual system. It all seems odd, but we might speculate that the physiological processes that support our perception of flicker, the perceived sharpness of borders, and the perception of motion, all somehow go through internally generated minima at particular relative intensities of lights of different wavelength compositions. The fact that these intensities are the same for all three tasks suggests that the three tasks all depend on neural processes with the same spectral sensitivity curve. Another way of saying this is that a single putative spectral process *leaves its signature* on our perception of flicker, of borders, and of motion.

As a general rule, when the same curve – the same signature – turns up repeatedly, vision scientists tend to speculate that this signature has its origins in the responses of individual neurons. That is, we speculate that there is a set of individual neurons within the visual system that also have this same spectral sensitivity curve. If we take these speculations seriously, we would predict that there will be neurons in the visual system with spectral sensitivities that correspond to $V'(\lambda)$ at scotopic light levels and to $V(\lambda)$ at photopic light levels, and we might choose to go and look for them.

### 3.4.2 Linking propositions

Each of the attempts at theory given above will have at least one linking proposition, implicit or explicit, among its premises. For example, arguments concerning simplicity and stability depend on premises concerning simplicity and stability, such as "Simple perceptual facts imply simple neural correlate", and "stable perceptual phenomena imply a stable underlying neural process". Arguments concerning additivity depend on an additivity proposition: "Additive perceptual properties imply additive neural properties".

The argument on the tasks of photometry seems to rely on an analogy linking proposition such as, "Perceptual minima imply corresponding physiological minima". That is, a minimum of perceived flicker, or border sharpness, or motion, suggests a corresponding minimum in the signal that codes flicker, or borders, or motion respectively, somewhere within the visual system. And arguments from "signatures" to neurons seem to involve a proposition to the effect that "A recurring perceptual "signature" implies the presence of neurons with an analogous signature".

These assumptions are all superficially innocuous, and are easy to ignore, or leave implicit in an argument. But they are probably not all true. Watch for processes that conform to the inferences and speculations that depend on them, or that reject them, in later chapters of the book.

## 3.5 Case 2: The color of different wavelengths

We now turn to a second example of the study of appearance. At photopic light levels, colors are perceived. How do psychophysicists specify and quantify the perceived colors of different wavelengths of light, and the relationships among the colors?

### 3.5.1 Physics versus perception: The appearance of the spectrum

Imagine that you are looking at a rainbow, or at a spectrum made with a prism. The physical spectrum is a set of lights of different wavelengths, ordered by wavelength. If we ask a subject what he sees, he will say that perceptually the physical spectrum looks like an array of colors, with neighboring wavelengths taking on similar colors. If we give him color names and ask him to point to the location of each color within the spectrum, he will point approximately as given in Table 3.2. The intermediate colors – blue-violets, blue-greens, yellow-greens, and oranges – falling in between the other color names. In short, there is an orderly mapping between wavelength and perceived color[2].

[2] It is dangerous to emphasize the mapping between wavelength and color, because doing so tends to reinforce the common but erroneous belief that perceived color depends only on wavelength. In fact, perceived color is influenced by many other factors. Except when we are viewing a physical spectrum in isolation, the perceived color of a light tells us remarkably little about its wavelength composition. We revisit this topic in Chapter 7.

| | |
|---|---|
| Violets | 400-450 nm |
| Blues | 460-480 nm |
| Greens | 500-530 nm |
| Yellows | 560-580 nm |
| Reds | 620-700 nm |

Table 3.2: The typical appearance of selected wavelengths of light in an otherwise dark field.

Beyond the orderly mapping of wavelength to color, there are a variety of striking differences between physical and perceptual realms in color vision. In particular, the physical spectrum varies continuously in wavelength, and short and long wavelength light have no special physical commonality. Yet most subjects report a perceived similarity – a common reddishness – between short wavelength lights (which appear violet) and long wavelength lights (which appear red). Moreover, if we mix long and short wavelength lights in varying proportions, we can generate a continuous variation in color, from violet through reddish violet to red – colors that never arise from individual wavelengths of light. these colors are often called the *extraspectral purples.* Wavelength just varies from short to long, but the set of perceived colors makes a circle, as shown in Figure 3.7. And other mixtures of wavelengths are seen as whites (grey in context) and pastel (desaturated) colors.

Now imagine setting up a display with lights of many different wavelength compositions, arranged in haphazard order, but matched in brightness, and asking your subjects to arrange them in order of perceived similarity. If your subjects are color-normal, they will probably tell you that they cannot arrange the colors by similarity along a single line, but that all of the lights fit together naturally on a two-dimensional surface, with saturated colors around the outside (as in Figure 3.7) and whites and desaturated colors in the center. If we now add variations in intensity to the display lights, the subjects will tell us that they require a third dimension, with the higher intensity lights occupying spaces above the lower intensity ones.

Notice that again, the regularities of the three-dimensional color solid correspond only partially to the physical similarities among the stimuli. Nothing in the physical stimulus explains why a straight line from red to green goes through white (or mid-grey) in the color circle; and nothing explains why red (associated with long wavelengths) and violet (associated with short wavelengths) are similar in our perceptions.

### 3.5.2 Unique versus binary hues and mutually exclusive hue pairs

In 1878, the German scientist Ewald Hering drew attention to another set of facts about color appearances. First, he argued that perceived hues come in two distinctly different kinds: *unique* versus *binary.* That is, some hues seem to be made up of perceptual combinations of other hues. For example, if we ask you what color is reddish yellow, you would probably readily come up with orange; purple can be described as a reddish blue, and so on. These are the binary hues. Other hues – red, yellow, green, and blue – seem to be perceptually simple, or unitary; they are not perceptually analyzable into components. For example, if you were asked to imagine a purplish orange, it would take you a while to figure out what hue fall between purple and orange, and then you would probably reject the original request. No, you would probably say, red is not a purplish orange. Red is red, and that's all there is to it! You have just cast a vote for red as a unique hue, and orange and purple as binary hues.

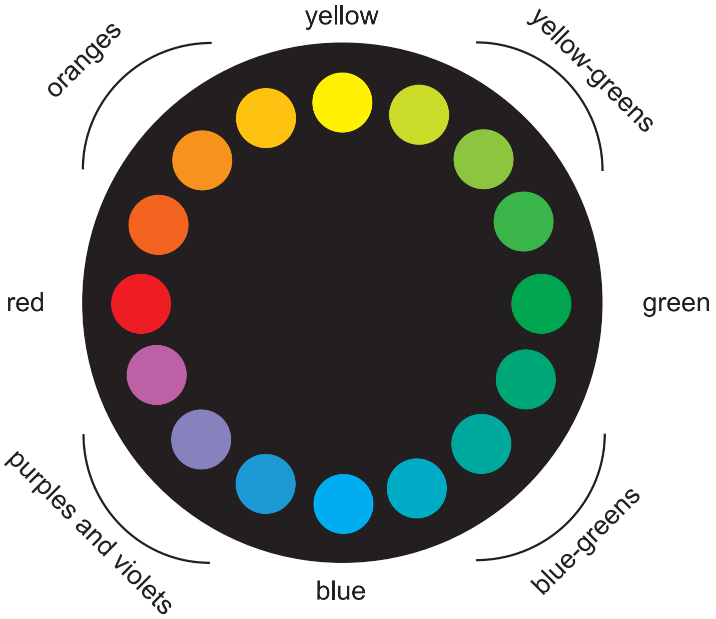


Figure 3.7: Saturated colors placed in a hue circle. The colors from red to yellow to blue to violet can be produced using lights of a single wavelength: the colors of the spectrum. In contrast, the purples require mixtures of short and long wavelength.

Hering further observed that the unique hues come in two pairs – red versus green and yellow versus blue – such that the colors in each pair are perceptually *mutually exclusive*. The claim is, perceptually there is no such thing as a reddish-green or a yellowish-blue. The mutually exclusive hue pairs were also called *opponent hue pairs*, and color theories that arise from these observations are called *opponent process theories* of color vision. A third dimension, lightness, was also included, with white and black as the defining sensations. This dimension isn't like the two color dimensions, particularly in that subjects usually find it possible to imagine blackish whites (as grays).

But shall we believe Hering's observations? Are there really unique and binary hues, and mutually exclusive hue pairs? Not everyone agrees. Some people say that for them green is perceptually "made up of" yellow and blue. We suspect that this common report comes from a knowledge of mixing paints; yellow paint mixed with blue paint often does yield a paint that looks greenish. But this is not what we're asking. Rather, we're asking, what do the colors themselves *look like*? To address this question, we need to develop new psychophysical techniques.

## 3.6 Quantitative color naming techniques

A more quantitative technique for describing the appearances of colors is called *hue naming* or *color naming*. In a color naming experiment, a subject is given a set of color words to use, and asked to use them quantitatively to describe the perceived hues of spectral lights. For example, the subject might be given the names corresponding to Hering's four unique hues – red, yellow, green, and blue – and asked to name the color of each wavelength. If more than one color is perceived, the subject is asked to assign a percentage of the perceived hue to each of the color names. He might describe

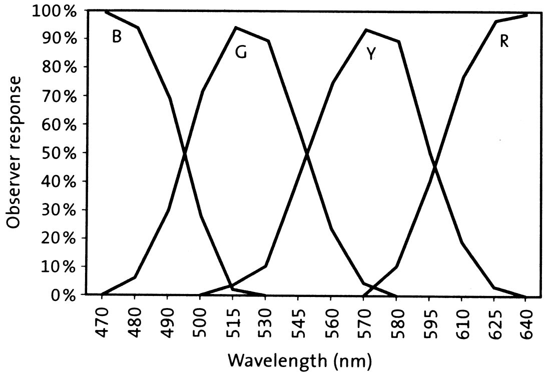


Figure 3.8: Results of a color-naming experiment over a large spectral range, from 470 to 640 nm. Subjects were allowed to use the color names blue, green, yellow, and red. [From Wooten and Miller (1997, Fig. 3.7, p. 77).]

a 610 nm light (which typically looks orange) as 40% yellow, 60% red; and a 575 nm light as 100% (unique) yellow. The percentages are averaged across presentations to yield an overall percentage score for each color name at each wavelength. Remarkably, subjects do this task very consistently, and agreement among subjects is strong, especially given the seemingly subjective nature of the task. [Try color naming on yourself and a friend with the colors in Figure 3.7. Make sure neither of you is color-blind!]

The results of a color-naming experiment using the four color names red, yellow, green, and blue, over the wavelength range 450 to 660 nm, are shown in Figure 3.8. Under the conditions of this particular experiment, the subjects consistently used the color names blue, green and yellow over the appropriate wavelength ranges. The consistency of use of the color names verifies the orderly mapping of wavelengths to colors. Unfortunately wavelengths below 470 nm were not used in this experiment; but other experiments show that in addition to the color name blue, the color name red is used at wavelengths below 450 nm. This result validates the claim (Figure 3.7) that perceptually the colors form a circle rather than just an ordered line.

Experiments like these provide some vindication of Hering's original claims. At 470, 515, and 570 nm, the appropriate unique hue names were used, almost to the exclusion of the other color names (e.g. the subject reported 98% yellow at 570 nm); and the color name red predominated beyond 625 nm. Moreover, the intermediate wavelengths were readily described with pairs of names of the neighboring unique hues. And the color names of the mutually exclusive hue pairs – red and green, and yellow and blue – were virtually never used to describe the same wavelength, confirming the mutual exclusivity of particular hue pairs. This experiment shows that subjects perform reliably, and that these four color names are *sufficient* to describe all of the hues; and it bolsters the claim of uniqueness versus binariness as well as the perceptual mutual exclusivity of the members of the two mutually exclusive hue pairs.

But to what extent are the results determined just by the experimenter's original choice of color names? What would subjects do if we gave them fewer color names to use, or a different set? Charles Sternheim and Robert Boynton (1966) tested this question in the wavelength range 530-620 nm, asking subjects to use several different sets of color names on different runs of the experiment. Subjects were told that their response categories need not add up to 100% – if the available color names were insufficient to describe the perceived hues, the subject could just leave out some of the percentage points. These leftover percentage points are given in what they called the *computed function* labeled CF that is plotted with each data set. Consequently, the CF curve reveals the degree to which the available color terms in a particular condition were insufficient to describe the color.

Some of Sternheim and Boynton's data are shown in Figure 3.9. As we would expect by now, the three color names green, yellow, and red were sufficient for the color-naming task across this wavelength range (Figure 3.9A) – virtually no percentage points went unused. But the two color names green and red were not sufficient (Figure 3.9B), nor were the three color names green, orange and red (Figure 3.9C). In both cases, the CF curve closely resembled the curve generated by the term yellow when it was allowed (Figure 3.9C). Finally, all four color names – green, yellow, orange and red – were also sufficient to describe the hues in this wavelength range (Figure 3.9D) (no CF curve at all).

The implication of Sternheim and Boynton's experiment is that the color name yellow is necessary for describing the colors in this wavelength range, whereas the color name orange is not necessary. Orange can be described satisfactorily as a reddish yellow, but yellow cannot be described as an orangish green. The results thus support the idea that yellow is a unique hue, whereas orange is a binary hue. Similar experiments have been done in other regions of the spectrum and among the purples. In each case, the data support the uniqueness of Hering's four unique hues, and the binariness of the hues in between. These data thus provide a modern, quantitative vindication of Hering's original observations.

## 3.7 Linking theories of color appearance

As was the case with photometry, the system properties associated with color appearance led color theorists to propose several physiological inferences and speculations.

### 3.7.1 The appearance of the spectrum

The fact that different wavelengths of light look different colors suggests that different wavelengths of light set up different neural codes. The fact that neighboring wavelengths look similar in color suggests that neighboring wavelengths cause similar values in the neural code for color. The perceived similarity between short and long wavelength lights (with violet having a reddish tinge) suggests that, surprisingly, the neural codes arising from short and long wavelength lights have some internally generated feature in common. The existence of the purples suggests that some mixtures of lights set up novel neural codes that are not set up by any individual wavelength. And the circularity of the perceptual hue circle suggests that the values of the neural code form a continuous variation from short to long wavelengths, through the purples, and back to the code for short wavelengths again.

### 3.7.2 A neo-Heringian opponent process theory

What about unique versus binary hues, and mutually exclusive hue pairs? Ewald Hering not only called attention to these system properties, but also proposed a theoretical account of why they

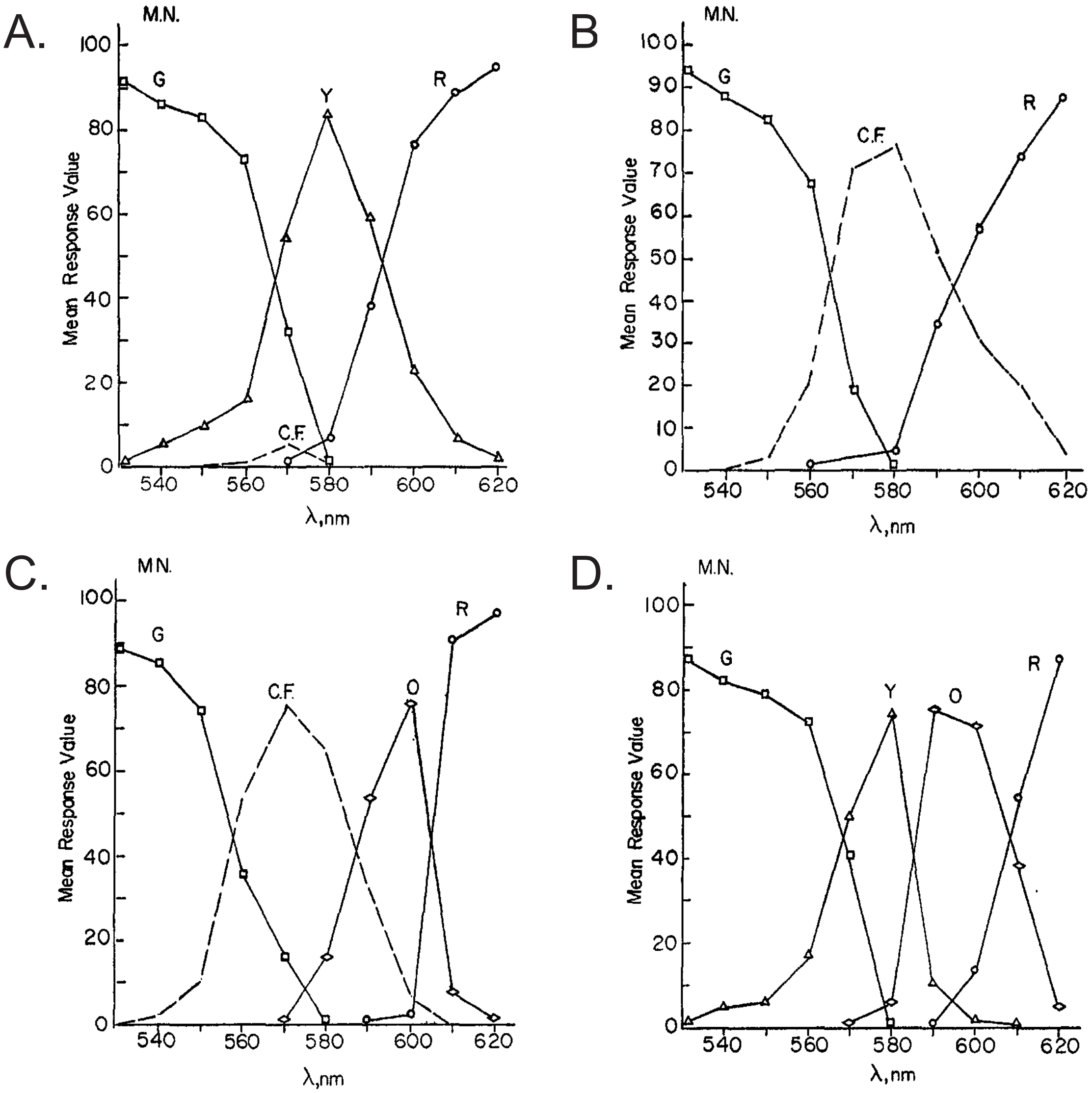


Figure 3.9: Color naming for one subject with four different sets of color names. Wavelengths over the range 530-620 nm were used. Subjects were allowed to use four different combinations of color names. The subjects were told that if there was a color name they needed, but were not allowed to use, they should assign less than the sum of 100 points to that wavelength. When points were left out, the authors calculated and plotted the missing points, labeled *C.F.* for *computed function.* The color names allowed were: A. Green, yellow, and red; B. Green and red; C. Green, orange, and red; D. Green, orange, yellow and red. The computed functions in B and C resemble the use of the color term yellow in A. The authors argue that the color term yellow is necessary to describe the colors in this wavelength region, whereas the color term orange is not necessary. [Adapted from Sternheim and Boynton (1966, pp. 772-773).]

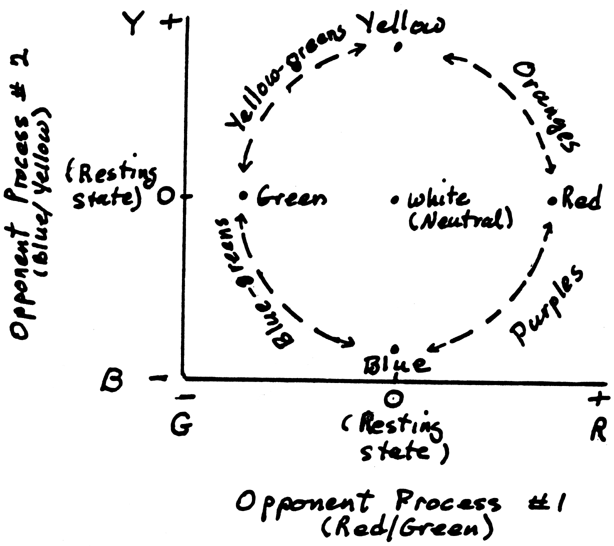


Figure 3.10: A neo-Heringian opponent process theory. According to this theory, the unique hues come about when one of the two physiological opponent processes deviates from its resting state in a particular direction, and the other is at its resting state. The binary hues come about when both channels deviate from their resting states. The perceptual mutual exclusivity of the mutually exclusive hue pairs comes about because of the literal mutual exclusivity of the signals: a single physiological process cannot both increase and decrease from its resting state at the same time.

should occur. His suggestion was that we could use the perceptually mutually exclusive properties to infer the existence of physiologically mutually exclusive processes. In fact, he argued that the same physiological process – a so-called opponent process, which changes in two opposite ways from a neutral state – provides a neural code for a pair of opponent colors.

Of course Hering lived before recordings had been made from single neurons, and his opponent processes could not have been framed in modern terms. But today, we can invent a neo-Heringian theory by thinking of a neural mechanism that (say) increases its output to signal redness, and decreases its output to signal greenness. Opponent process theory posits that the two perceived colors are mutually exclusive precisely because the two states of the neural mechanism are necessarily mutually exclusive: the physiological opponent process can't both increase and decrease its output at the same time. To complete the theory, we would add a second kind of opponent process to code the perception of yellow and blue; say, increasing its output to signal yellowness, and decreasing its output to signal blueness. (The polarity on each dimension is assigned arbitrarily in both cases).

This neo-Heringian theory is shown schematically in Figure 3.10. The theory posits the existence of two opponent processes, each of which can be at its resting state, or deviate from its resting state in either of two directions. The states of the two processes are plotted on the two axes. The unique hues come about when one of the opponent processes is active in a particular direction, while the other channel is at its resting level. The binary hues come about when both channels are active. For example, the unique hue yellow occurs when opponent process #1 is at its resting state

| | | | |
|---|---|---|---|
| 1. Initial ME proposition | ME Φ | → | ME Ψ |
| 2. Contrapositive ME | Non-ME Ψ | → | Non-MEΦ |
| 3. Converse ME | ME Ψ | → | ME Φ |
| 4. Converse contrapositive ME | Non-ME Φ | → | Non-ME Ψ |

Table 3.3: The mutual exclusiveness family of linking propositions. As with others, the mutual exclusiveness (ME) family has four members: the initial proposition; its contrapositive; its converse; and its converse contrapositive. The initial proposition is: Mutually exclusive physiological states imply mutually exclusive perceptual states. Its contrapositive is: non-mutually-exclusive (mutually compatible) perceptual states imply non-mutually-exclusive (mutually compatible) physiological states. The converse is: mutually exclusive perceptual states imply mutually exclusive physiological states; and the converse contrapositive is: mutually compatible physiological states imply mutually compatible perceptual states. The contrapositive and the converse are used in reasoning from psychophysical data to neurophysiological conclusions. [Adapted from Teller (1984).]

and opponent process #2 deviates from its resting state in the positive direction. The non-unique hue orange occurs when both the redness/greenness process and the yellowness/blueness process deviate from their resting states in the positive direction; and so on.

## 3.8 Mutual exclusiveness as a linking proposition

The neo-Heringian model uses system properties – the perceived mutual exclusiveness or mutual compatibility of particular hue pairs – to make predictions about visual physiology. In fact, it brings us to another family of relational linking propositions, which we can call the mutual exclusivity family. This family is shown in Table 3.3.

Since the original observations were psychophysical, Hering's reasoning had to be from psychophysics to physiology, and had to start from the contrapositive and converse propositions. Hering's argument was that mutually exclusive perceptual states imply mutually exclusive physiological states – hence the axes with their mutually exclusive ends in Figure 3.10. And mutually compatible perceptual states imply mutually compatible physiological states – hence the off-axis regions and the binary hues. But of course, logically the contrapositive implies the initial proposition, and the converse implies the converse contrapositive, so an opponent process theorist must be prepared to endorse the whole mutual exclusiveness family of linking propositions.

## 3.9 Summary: Appearance experiments and color vision

In this chapter we have explored appearance experiments. In the course of the chapter, we addressed three goals: to develop two examples of the use of appearance experiments in color vision; to examine linking theories between color appearance and neural coding; and to ferret out some novel linking propositions implicit in these arguments.

The first set of appearance experiments centered on photometry. Photometry arose from the need for a set of quasi-physical units for the intensity of light. In searching for an additive system, physicists and psychophysicists tried several tasks, including brightness matching, as well as the minimization of perceived flicker, border distinctness, or motion. Of these, brightness matching has the greatest face validity, but the task is difficult and the measured values are not additive across the spectrum. The three minimization techniques yield more consistent and less variable data, and the values obtained obey additivity.

The second set of experiments dealt with color appearance: the perceived hues of lights of different wavelengths. The psychophysics of color appearance reveals some properties not present in the physical nature of light. For example, ordered by similarity, the hues of the perceptual spectrum converge at the spectral extremes, and form a circle when the extraspectral purples are included. We also examined claims that the color circle has an additional finer structure, in that certain hues (red, yellow, green and blue) are perceptually unique, whereas others are perceptually binary; and that the unique hues come in mutually exclusive pairs.

In terms of linking theory, we argued that vision scientists have tried to explain system properties in terms of inferences or speculations about the properties of the underlying neural codes. We examined several such arguments, and tried to develop an intuitive feel for the strength of the logic involved, from tight inferences down to cruder speculations. In regard to photometry, perhaps the most interesting argument is that the internally generated flicker, border distinctness, and motion minima that allow photometry to succeed, point to the existence of corresponding internally generated minima in the neural codes for flicker, border distinctness, and motion. In regard to color appearance, the most interesting argument is that the existence of unique and binary hues implies an opponent hue code at the neural level.

And what of linking propositions? In the case of photometry, the most novel linking proposition is that psychophysical minima suggest corresponding neural minima. In the color appearance cases, the most novel is that mutually exclusive hues imply the existence of mutually exclusive neural states.

Again, how do the system properties of vision come about? Why do we see as we do? What is it about the physiology of the visual system that creates the observed variations of perceived brightness with wavelength? Why are minimization matches additive, while brightness matches are not? Why are some colors unique and others binary, and some pairs mutually exclusive? All of these questions have been food for speculation and theory. We will return to them all more concretely in later chapters. For now, they wait on our list of system properties in search of explanations.

# Chapter 4

# Optics of the Eye

## Contents

In Chapter 4, we leave the province of psychophysics and enter the province of optics, one of the classical branches of physics. Optics is the study of light and how it interacts with matter. Our topic also includes physiological optics – the optics of eyes built by biological systems. Students with backgrounds in physics and biology, of course, will be more comfortable with these topics than they were with psychophysics, whereas students with backgrounds in perception will be less so.

In Chapter 1, we raised the question of spatial resolution: what limits the finest grating that one can see? We laid out four possibilities: the optics of the eye; the photoreceptor matrix; the spatial convergence of signals within the retina; and other factors at higher levels of the visual system. In the present chapter we return to the first of these options. The goal of the present chapter is to fill in the background needed to evaluate the possibility that the optics of the eye are the major factor that limits grating acuity.

To begin, we first define the units used to specify the spatial frequency of square wave gratings, both in physical and in quasi-physical terms. We then introduce the basic properties of light, including its dual nature as both waves and quanta. We outline four ways that light interacts with matter – reflection, refraction, diffraction, and absorption. We expand on the property of refraction in order to explain how lenses make images, and on the property of diffraction because of the remarkable role that interference fringes produced by diffraction patterns play in defining optical quality.

We then turn to physiological optics, and examine the human eye as an optical system. The optical elements of the eye – the cornea, pupil and lens – form an image of the physical world inside the eyeball, at the back, on the sheet of neural tissue called the retina. We spell out the consequences for vision of optical errors within the eye. We describe the possible sources of information loss within the human optical system, and outline how to specify its quality. Finally, we introduce adaptive optics, a technique with which, remarkably, it is possible to improve the quality of the retinal image within the living human eye.

In sum, the eye is a remarkable physical and physiological structure. It captures rays of light coming from the three-dimensional world, and focuses the rays to make the retinal image. The retina in turn provides the initial encoding of the information contained in the retinal image, that eventually allows us to judge the shapes, colors, motions, and distances of physical objects. But how much do the eye's optics affect the information available in the retinal image? The background provided in this chapter will allow us to attack this question again at greater depth in Chapter 5, in which psychophysical, optical, and neural themes combine.

## 4.1 Square wave gratings

We begin with a digression on the question of units. In Chapter 1, we specified the acuity gratings in Figure 1.2 only by letters: A, B, C, and so forth. Obviously, we need to adopt more formal units. Vision scientists specify gratings in two different kinds of units – one physical, and the other quasi-physical.

### 4.1.1 Spatial frequency in cycles per centimeter

Figure 4.1 shows luminance profiles of three square wave gratings that correspond to gratings from Figure 1.2. Each white stripe of the grating gives a relatively high luminance, and each black stripe gives a relatively low luminance. The name *square wave grating* comes about because the transitions between black and white are abrupt thus producing right-angle corners in Figure 4.1 like the corners of a square.

Physically, these gratings alternate between black and white stripes at regular intervals across space. Such cyclical patterns can be specified in terms of the number of *cycles* per unit distance. By this convention, one black and one white stripe of the grating constitute a cycle. The number of cycles per unit distance is called the *spatial frequency* of the grating (we will refine this definition later). Thus, the gratings in Figures 1.2 and 4.1 can be specified in terms of *cycles per centimeter*[1].

### 4.1.2 Spatial frequency in cycles per degree

As we saw in the case of luminance, vision scientists often develop quasi-physical units; that is, special units that specify the properties of the physical stimulus in terms of its probable effectiveness for human vision. Figure 4.2 shows another schematic view of a human eyeball. As stated previously, the eye forms an optical image of the physical objects in the visual field. For an object of a fixed size at a fixed distance, we can estimate the size of the image to a first approximation by drawing lines from the edges of the object, crossing (as it turns out) just behind the lens, and diverging again to hit the retina. For a fixed distance, the larger the physical object, the larger

[1] Specifying the grating in cycles per unit distance yields units that are counterintuitive for some people, since coarser stripes are designated by smaller numbers. Just remember – the *finer* the stripes, the *more* of them will fit in a unit distance – so the *higher* the spatial frequency.

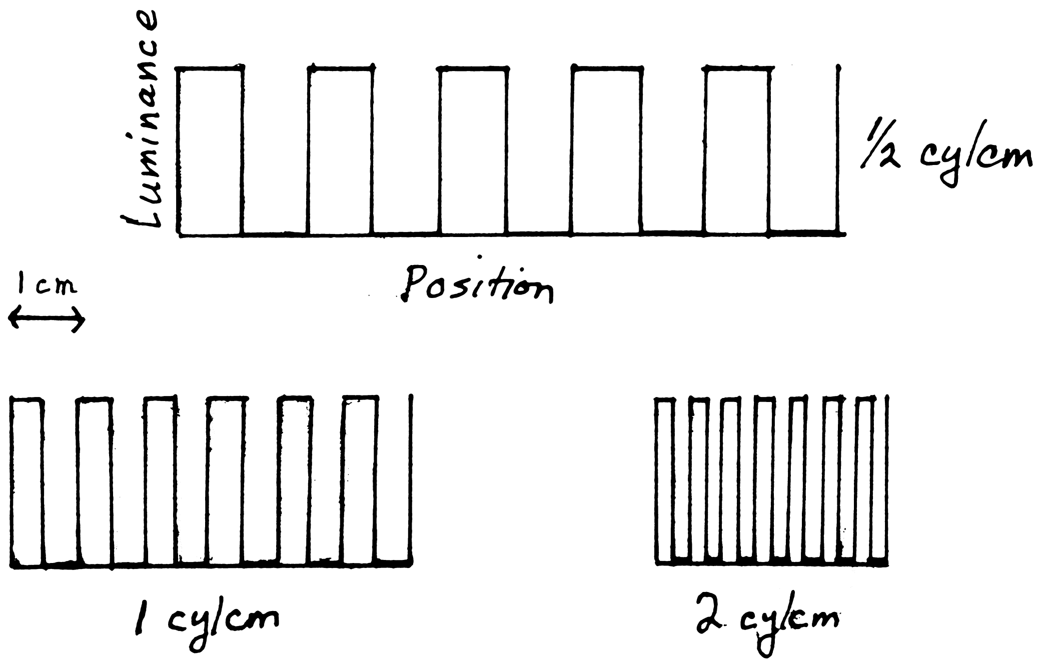


Figure 4.1: Physical specification of a square wave grating. The sketches show variations in luminance across spatial position. One cycle of a grating consists of one high-luminance and one low-luminance region. The spatial frequency of each grating is specified by the number of cycles per centimeter (cy/cm).

the retinal image. Similarly, the higher the spatial frequency of a physical grating, the higher the spatial frequency in its retinal image.

Things become more complicated when we vary the distance of the object. For an object of a fixed size, if we double the distance we will cut the image size in half; or, to keep the image the same size, we will have to double the size of the object. Similarly, for a square wave grating with a fixed spatial frequency, if we double the distance we will cut the width of each stripe in half, and so double the number of stripes per unit distance. To keep the same spatial frequency in the retinal image, as we double the distance we will have to double the widths of the stripes in the grating (*decrease* its spatial frequency).

How do we derive units for the sizes or spatial frequencies of objects in retinal image terms? We can think of the eye as occupying the center of a 360 degree circle. So, we can specify a stimulus in units of its *angular size at the eye*; that is, in terms of the *visual angle* it occupies (*subtends*). For example, if three cycles of a grating fit into a single degree of visual angle, it is a three cycles/degree (cy/deg) grating. (Other common abbreviations for cycles per degree are *c/deg* and *c/* °.)

We call these measurements quasi-physical because they depend on how the subject views the stimuli rather than on the stimulus in the physical world. Specifically, rather than specifying the size with respect to that physical grating (physical size), we specify the size of the image in the eye

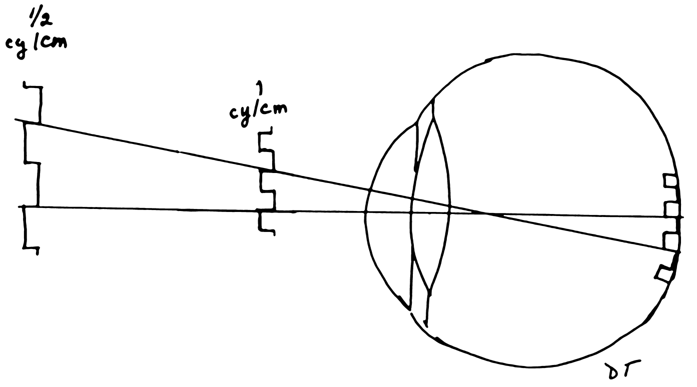


Figure 4.2: Quasi-physical specification of a square wave grating. The spatial frequency of the grating is specified in units of the angle subtended by one cycle at the eye. The two gratings pictured differ in spatial frequency, specified in cy/cm. However, the 1/2 cy/cm grating is twice as far from the eye as the 1 cy/cm grating, such that a single cycle of each grating occupies the same angular size at the eye. At their respective distances, these two gratings produce nearly identical retinal images of the same spatial frequency when specified in cycles per degree.

made by the grating (visual angle). This is analogous to the case of measuring intensity at different wavelengths by either the energy in the world (radiance) or its the effect on vision (luminance).

There are two rules of thumb that will give you a better intuitive feel for specifying stimuli in terms of visual angle. The first is wonderfully literal: your thumbnail at arm's length subtends about 1° of visual angle. The second is that the sun and the moon, at their respective distances, each subtend about 1/2°. You can check that these two rules of thumb are consistent by measuring the angular size of the moon with your thumbnail held at arm's length. (Do not try this with the sun!) You will find that in terms of visual angle, the moon is about half as large as your thumbnail at arm's length.

These quasi-physical units are useful in specifying spatial resolution, for both empirical and theoretical reasons. Empirically you already have evidence, from viewing Figure 1.2 at different distances, that the physical grating you could just barely resolve varied with distance. When the grating is specified in physical units, every doubling of distance requires approximately a halving of spatial frequency for resolution. But specified in quasi-physical units, spatial resolution turns out to be virtually constant across changes in the distance of the grating, allowing us to separate the parameters of spatial resolution and distance.

More theoretically, it makes sense to guess that the spatial resolution capacities of the retina will be related to size or spatial frequency in the retinal image rather than in the world. A retinal image of a fixed spatial frequency will always make the same pattern on a fixed patch of retina. This pattern will be processed by the same patch of photoreceptors, interneurons, and ganglion cells both times, and it makes some sense to assume that the same retinal image will be processed the same way each time. Thus, from this point on, gratings will always be specified in terms of cy/deg.

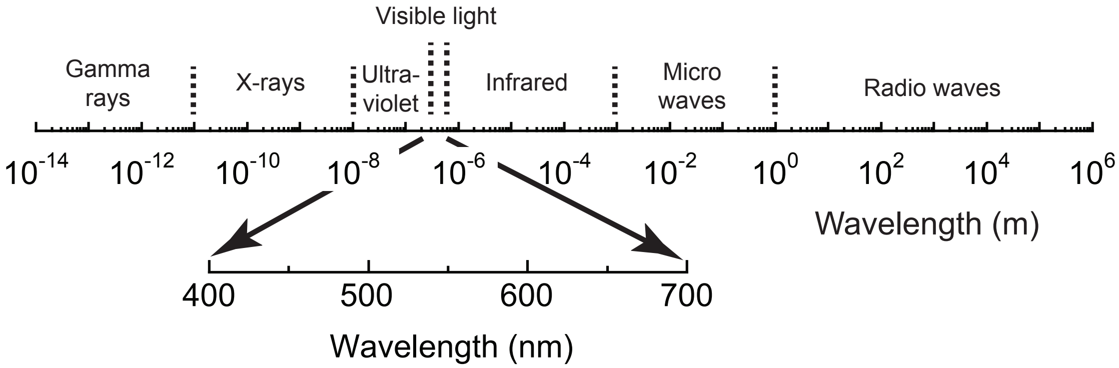


Figure 4.3: The electromagnetic spectrum. The electromagnetic spectrum is shown with wavelength logarithmically scaled and specified in meters. The visible portion of the electromagnetic spectrum, which we call light, occupies only a narrow range of wavelengths, from about 400 to about 700 nanometers (1 nm = $10^{-9}$ meters). The insert shows this narrow range with a linear scale and labeled in nm. Under photopic conditions and with a black surround, the wavelengths 400 to 700 appear the colors of the rainbow: violet, blue, green, yellow, orange, and red. Under scotopic conditions, all wavelengths of light look whitish, and no colors are seen.

### 4.1.3 How good is grating acuity?

Armed with units of measurement, we can now ask, how good is grating acuity? The answer is that under the best conditions, in the best young eyes, grating acuity is just *about 60 cy/deg.* Alternatively, since there are 60 minutes of arc in one degree, the best visual acuity can also be stated as about *1 cycle per minute of arc.* That is, the finest grating you can resolve contains about 60 black and 60 white stripes across your thumbnail, at arms length. In Figure 1.2, this is the approximate spatial frequency of grating $D$ at 4 meters or grating $E$ at 2 meters. Use Figure 1.2 to recheck your visual acuity with these numbers in mind. We will come back to these numbers several times in the next few chapters. Grating acuity of 60 cy/deg joins the scotopic and photopic spectral sensitivity curves as system properties in search of explanations.

## 4.2 Light

### 4.2.1 The electromagnetic spectrum and the visible spectrum

*Electromagnetic energy* (also called *radiant energy*) is one of the basic forms of energy in the universe. Electromagnetic energy varies in its *wavelength* which has already been introduced in our discussion of spectral sensitivity. As shown in Figure 4.3, the electromagnetic spectrum encompasses many orders of magnitude of variation in wavelength. This includes the phenomena of gamma radiation, x-rays, light, microwaves and radio waves.

But what is *light*? The term light is used to refer to the part of the electromagnetic spectrum to which human eyes are sensitive. As you already learned in Chapter 2 and 3, the visible spectrum covers a range of wavelengths of just less than a factor of two, from about 400 to about 700 nm. These limits are not absolute, but represent practical extremes based on the rapid fall-off of the eye's sensitivity at the ends of this range (as shown, for example, in Figure 3.5B).

The psychophysical measurements such as those involved in defining the spectral sensitivity curves you saw in Chapter 2 and 3 enter into the fundamental definition of light. In fact, the distinction between electromagnetic energy and light provides the quintessential example of quasi-physical specification. If it's a wavelength humans can see, it's light; if not, it isn't.

The other immediately striking thing about light is that at photopic levels, the different wavelengths of light take on characteristic colors, as we saw in the color naming experiment in Chapter 3. This fact is so universally appreciated that expectations about the perceived colors of lights of different wavelengths are often included in diagrams like that in Figure 4.3. Vision scientists, however, usually avoid this practice, in order to avoid conflating physical with perceptual entities. Instead, we reserve one set of terms (say, wavelength) to describe the physical characteristics of the stimulus, and another set (say, color names) to describe the perceptual characteristics of the stimulus.

There are two important reasons for making such terminological distinctions. First, if we are initially clear in separating physical and perceptual realms, we are set up to ask how one realm maps to the other, without any initial presumptions. And second, the mapping between physical and perceptual realms is often complex, and (as we will see) many factors other than wavelength influence perceived colors.

In terms of design, why is the visible range restricted to between 400 and 700 nm? Part of the answer is that it makes evolutionary sense to match the visual system to the wavelengths that are available at the earth's surface and are therefore available to our eyes in our natural environment. Electromagnetic radiation from the sun is blocked by the earth's atmosphere for all but two ranges of wavelength. One range encompasses radio waves which makes possible radio astronomy. The other range is for visible light. Thus, visible light is available on the earth's surface to support vision[2].

A second part of the answer is that vision in either the ultraviolet or the infrared has practical disadvantages. Ultraviolet light can destroy biological structures in the eye – remember what it can do to the skin! Ultraviolet radiation also encourages the yellowing of the lens and the formation of cataracts; and there is some evidence that it can even damage the short-wavelength-sensitive photoreceptors – the cells that capture short wavelengths of light within the retina.

Infrared wavelengths, at the opposite end of the spectrum, are similar to our bodily radiations due to heat. Snakes can sense infrared radiation, to help them locate prey. But for us, seeing our own body heat within our own eyes would add noise that would tend to mask the images of objects in the real world. Detection of our own body heat would be especially deleterious at absolute threshold, where every bit of energy in the visible range counts.

### 4.2.2 Quanta versus waves

Electromagnetic energy sometimes behaves like waves and sometimes behaves like particles, or discrete packets of energy. For a long time physicists argued about whether electromagnetic energy was "really" waves or "really" particles. We now know that both the wave-like and the particle-like properties of light, or any electromagnetic radiation, can be fully and consistently described mathematically. For the non-mathematician, the conceptual problem is that there is no single

[2] On this view it is not surprising that different animals have different spectral ranges for vision, depending on their environments. Different species of fishes, for example, tend to match their spectral sensitivity curves to the wavelength composition of the light that penetrates water to the particular depth at which they live.

entity at the level of things we can observe directly that has both kinds of properties, so it's hard to imagine light as having them both. The way around this conceptual blockade is to be willing to use different analogies to elucidate different properties of light. We will do this below.

In fact, the wave-like and particle-like properties of light manifest themselves under different conditions. Light behaves like waves when traveling through air or another transparent substance (*medium*) – for example, from the sun to the earth, or from a physical object to your eye, or within the eyeball. It behaves like particles of energy when interacting with matter – for example, when it is absorbed by a physical object, or by your retinal cells to start the visual process. Both the particle-like and the wave-like properties of light are important to understanding vision, and vision scientists use the two different concepts interchangeably at different times, depending on what the light is doing.

Physicists use the term *quantum* (plural *quanta*) to refer to a particle of light when its particle-like properties are being emphasized. Quanta of light (within the visible range) are sometimes called *photons*. However, in other contexts the terms quantum and photon are used interchangeably.

### 4.2.3 Quantal fluctuations

When considered as particles, light has another important property for vision. It turns out that the emission of a quantum is a probabilistic occurrence. Thus, the output of any given source of light is not precisely constant in terms of quanta/sec, but varies over time. Moreover, the quantal fluctuations increase (the magnitude of the noise increases) as the intensity of the light increases. Thus, the physical variability of the light source itself is one of the major factors that limits detection thresholds in human vision. This topic is elegantly discussed in Cornsweet (1970).

In Chapter 2, in the context of signal detection theory, we introduced the idea that a threshold can be considered as a signal/noise discrimination. We can now add that quantal fluctuations are a classic example of noise. In this case, the source of the noise is external to the observer, or *extrinsic*. *Intrinsic* noise – noise generated within the observer – will also influence visual thresholds.

## 4.3 Optics

### 4.3.1 Interactions of light with matter

A beam of light is traveling along as a wave on a straight path through the universe, minding its own business, when suddenly it encounters a bit of matter. What happens? There are four possibilities, as shown in Figure 4.4. First, *reflection* occurs when light, acting briefly in its particle mode, bounces off the surface of the matter. The wave now changes its direction of travel in a precise way. A useful analogy for reflection is a billiard ball bouncing off the side of the table. The angle at which the light hits the edge of the table (the *angle of incidence*) determines the angle at which it bounces off (the *angle of reflection*). In fact, all things being equal (e.g., no spin on the ball) the angle of reflection will be exactly equal to the angle of incidence.

The second possibility is *refraction*. Refraction occurs when light enters (but is not absorbed by) a new medium – for example, in passing from air to glass. If the medium is more dense (has a higher *index of refraction*) the wave is slowed down. As a result, it changes its direction of travel. A useful analogy here is a heavy vehicle going from asphalt to gravel at an angle (the asphalt is the air, and the gravel is the glass). As the first wheel (say the right front wheel) hits the gravel, it is

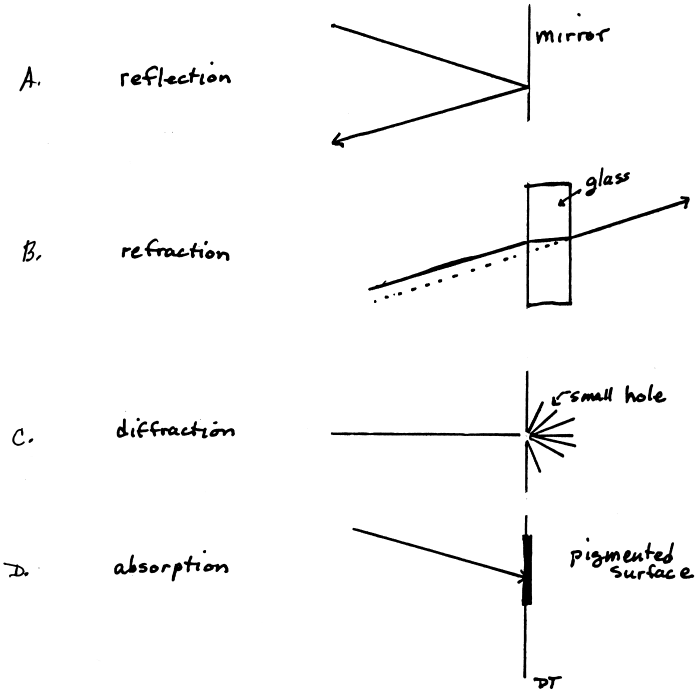


Figure 4.4: Four ways that light can interact with matter. A. reflection; B. refraction; C. diffraction, and D. absorption.

slowed down, and the vehicle tends to turn toward a line normal (perpendicular) to the boundary, changing its direction of travel. As it goes from gravel to asphalt again, the right front wheel hits the asphalt first, and speeds up again, and the vehicle turns the opposite way, once more changing its direction of travel. Any skier has experienced a similar phenomenon when going from ice to snow or vice versa. Note that by manipulating the boundaries between various transparent materials, we can manipulate the direction of travel of a beam of light. [What would happen in Figure 4.4B (refraction) if the pane of glass were triangular in cross-section?]

The third possibility is *diffraction.* Diffraction occurs when a ray of light passes very close to the edge of a piece of matter. The ray is bent in proportion to how close it is to the edge – the closer, the more bent. Think of water in a fast-moving stream as it courses around a rock. The bits of the stream that are close to the rock bend around it, while those sufficiently far away are not affected. Analogously, when light passes through a very small hole, it is bent outward in all directions. For slightly larger holes, the light is bent only at the edges of the hole and the result is a blurry spot (not a sharp one) on a piece of paper placed on the far side of the hole. As the hole gets larger, only the rays very near the edge are bent, so the light will make a concentrated spot with only a slightly blurry edge.

The final possibility is *absorption.* Acting as particles, individual quanta of light are absorbed by the individual molecules that make up the absorbing substance. They then cease to exist as electromagnetic energy, and become part of the energy state of the molecules that absorb them.

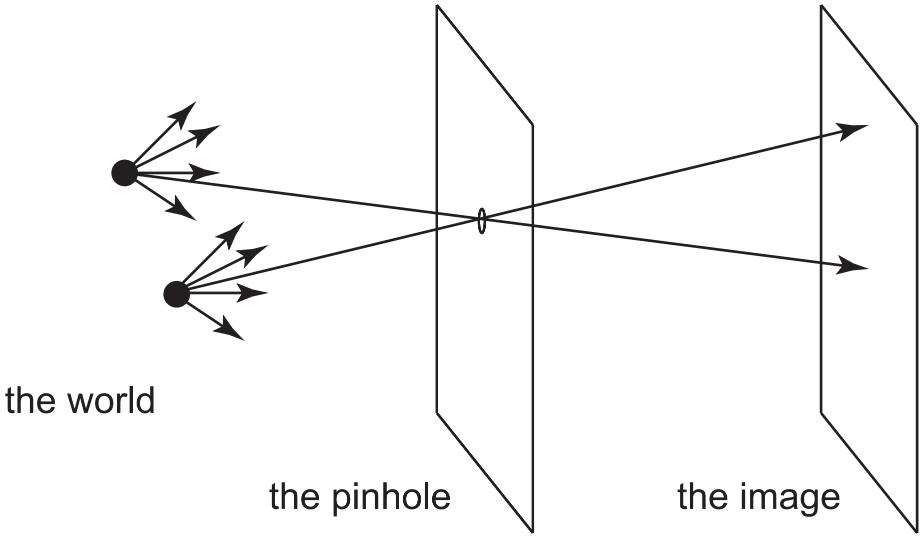


Figure 4.5: An illustration of a pinhole camera. The world with two light sources is on the left. On the right is a box with the front and rear surfaces shown. There is a pinhole in its front surface that restricts what rays of light can enter the box. As a result, there is an image of the light from the world on the back surface of the box.

### 4.3.2 Image formation

**Pinhole cameras**

An optical image is a two dimensional pattern of light that corresponds to the pattern of light coming to the image from a three dimensional world. Such images have a one-to-one relationship between points in the image and the points in the world that can send light rays to make the image. The principle of how such images are formed by eyes and cameras is illustrated by a pinhole camera in Figure 4.5. On the left is a "world" consisting of two points of light. On the right is a box with a small aperture on one side (the *pinhole*). A small fraction of the light from the points in the world enters the box through the pinhole. These rays of light fall on the back wall of the box and form an image. In this example, the image is two small regions of light that correspond to the two light sources in the world.

A key feature is that the box prevents light entering from anywhere but the pinhole. As a result, the pinhole constrains the light rays entering the box to create the one-to-one correspondence between the points in the world and the points of the image[3].

Consider now how the size of the aperture affects the image. A larger aperture will result in more light entering the box but at the cost of a blurred image due to the rays not quite aligning. A smaller aperture will result in less light entering the box but improve the sharpness of the image. Ultimately, if the aperture becomes small enough to cause diffraction, the image degrades from the "scattering" caused by diffusion. Thus, an optimal aperture for image sharpness is a compromise

[3] There are exceptions to the one-to-one correspondence due to phenomena such as partially reflective surfaces. For example, most windows allow one to see outside and to see your own reflection.

between limiting the rays that enter and avoiding diffraction.

**Lenses improve the pinhole camera**

The formation of an image by a lens takes advantage of refraction to improve on the idea of a pinhole camera. Imagine first that a lens is placed in the aperture of the pinhole camera. As shown in Figure 4.6A, light rays leaving a point on an object (or a point source of light) diverge from that point in straight lines in all directions. Suppose that a cone-shaped group of those rays encounters a glass lens. As the light passes from the air to the lens, it bends; and the greater the angle at which it strikes the air-glass interface, the more it will bend.

If we design the lens cleverly, with a surface that varies in curvature, we can bend each ray by a different amount; say, so that all of the rays are parallel to each other within the lens. If the second surface of the lens is equally cleverly shaped, we can bend each ray again, say just enough so that all of the original rays will converge to a single point on the far side of the lens. Rays from neighboring points on the object will converge at neighboring points in the image, and voila! – an optical image of the object. As shown in Figure 4.6B, the farther the source is from the lens, the closer to the lens the image will be.

Putting the lens in the aperture of a camera allows for much larger apertures without the loss of image quality that occurs with a pinhole camera. The result is that much more light can be gathered to create the image and that is what makes possible modern cameras and, as we will see, the human eye.

Now back to lenses. The power of a lens is defined by its *focal length.* Rays from one point on a very distant object[4] (say, a point on the surface of the sun) arrive at the lens virtually parallel to each other. This case is illustrated in Figure 4.6C. When these parallel rays pass through the lens, they will converge at a point on the far side of the lens. The distance from the lens at which the parallel rays converge is called the focal length, $f$, of the lens. The shorter the focal length, the greater the *power* of the lens. We express power in *diopters*, which are units of one over the focal length ($1/f$), where $f$ is in meters. So, if $f = 1$ meter, the power of the lens is 1 diopter. If f = 1/2 meter, the power is 2 diopters, and so on.

The lenses shown in Figure 4.6A-C are all convex, or *positive*, lenses – they *converge* the incoming rays of light and form an image. Concave, or *negative*, lenses, on the other hand, *diverge* the light and do not form images. The more the divergence, the higher the power of the lens. A negative lens is shown in Figure 4.6D. Both positive and negative lenses are used in fitting glasses to correct focusing errors of the eye, as will be discussed below.

### 4.3.3 Interference

The phenomenon of *interference* is illustrated in Figure 4.7. Interference patterns are a manifestation of the property of diffraction. If beams of light from the same source pass through two small neighboring slits in, say, a metal plate, each of the slits will diffract the light. The two diffracted beams will spread out, and can overlap beyond the slits. If the two overlapping beams then fall on a screen, they will form a set of fuzzy light and dark stripes called *interference fringes.* The analogy here is to the overlapping ripple patterns produced when you drop two stones into a pool

[4] The term *optical infinity* is used to refer to a distance beyond which further variations in distance have only negligible effects. For most purposes, objects more than 30 feet or so away are considered to be at optical infinity, and the rays from a point on an object at 30 feet or more are considered functionally parallel.

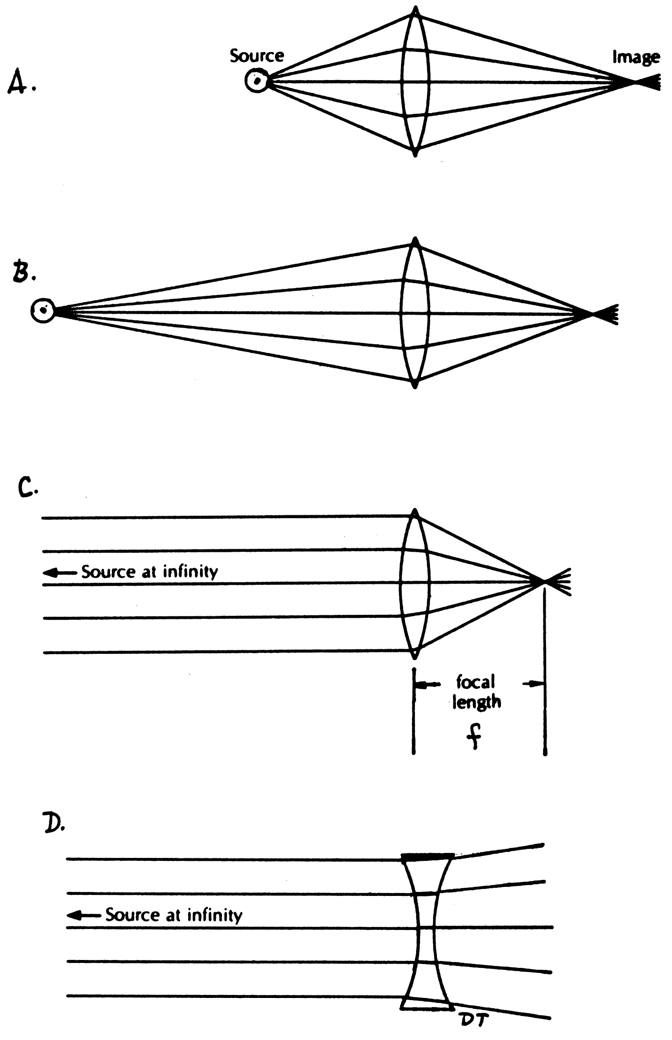


Figure 4.6: Lenses and image formation. A. A convex, or positive lens, forming an image of a point source. B. The farther the source is from the lens, the closer to the lens will be the image. C. When the source is at optical infinity, the rays from the source are parallel. The image is formed at a distance $f$ behind the lens. The distance $f$ is the focal length of the lens. D. A concave, or negative, lens diverges the light. [Modified from Cornsweet (1970, Fig. 3.9, p. 37).]

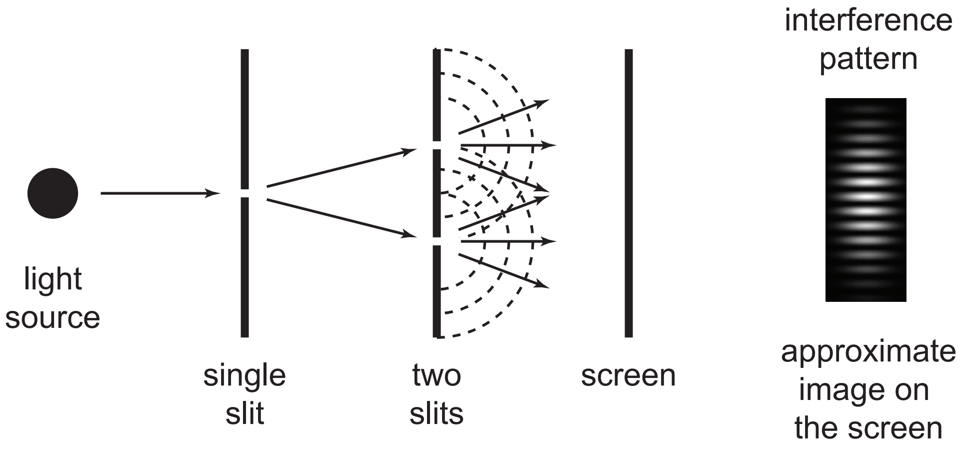


Figure 4.7: Interference. Light from a single source is diffracted by a first slit, and diffracted again by a pair of slits. The interference patterns are generated when the light from the two slits interact in the image on the screen.

of water – each set of waves produces high and low points, and where they overlap, the highs add to produce super highs and the lows add to produce super lows.

Here's a puzzle that illustrates the paradoxical nature of light. Suppose that you set up a double slit experiment. You shine a light from a laser onto the two slits, but you make the light so dim that on average, only one quantum per day will reach the two slits. Since a quantum is indivisible, one might think it would have to go through only one of the slits; and since interference is a property of the two beams together, one might think that no interference fringes could be made. Now you put a sensitive photographic film where the screen was, and go away for a year. The question is, when you come back and develop the film, will you see interference fringes? The answer is yes.

## 4.4 Physiological optics: The eye as an optical system

How do the properties of light and optics manifest themselves in the human eye? All of the interactions that light can have with matter are of importance to vision. Reflection is important in that some of the light reaching the front surface of the eye is reflected, never enters the eye, and therefore cannot contribute to vision. Refraction is important because the eye contains an optical system, and forms the retinal image. Diffraction is important because the pupil of the eye is a small hole, and all of the light that reaches the retina must pass through it. When the pupil is very small, many of the rays will be diffracted and will not reach the proper point on the retina. And, of course, absorption is critical, because the absorption of light by the photoreceptors within the eye changes electromagnetic energy into the first stage of the physiological signal, as we will see in Chapter 6.

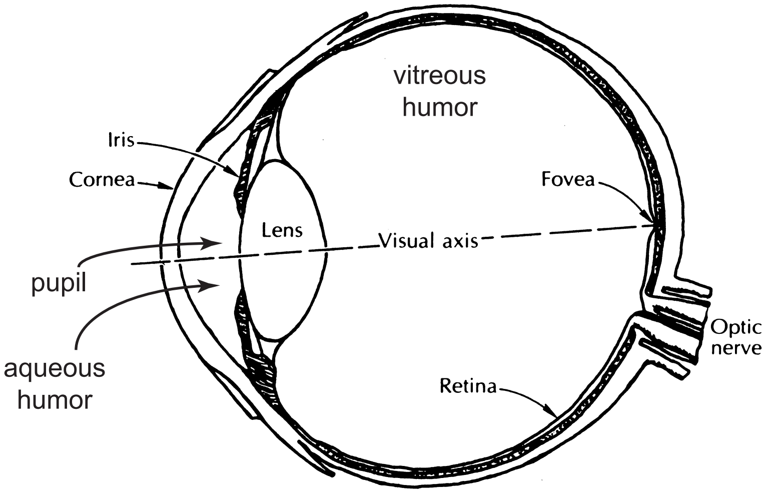


Figure 4.8: Optical elements of the eye. [Modified from Cornsweet (1970), Fig. 3.11, p. 40.]

### 4.4.1 Major optical elements: Cornea, lens, pupil

The optics of the human eye are shown in Figure 4.8. The light entering the eye passes through the transparent window, the *cornea*, which forms the external surface of the eye. It then passes through a thin liquid called the *aqueous humor*; then through the *pupil*, a small hole in the *iris* (the colored part of the eye); then through the *lens*; and finally through a viscous material called the *vitreous humor* and on to the retinal surface, where it passes through the inner retinal layers before being absorbed by the photoreceptors (see Figure 1.4).

The cornea and lens together serve to form an image of the physical world on the retina. Every time light encounters a change of refractive index, it changes its direction of travel. Interestingly, since the largest change in refractive index occurs between the air and the cornea, most of the focusing or bending of light rays is actually done by the cornea, and not by the lens. Then the lens fine tunes the focus onto the retina. The total refractive power of the eye is about 60 to 70 diopters; of the total, about 40 diopters is due to the cornea, and 20 to 30 diopters to the lens (see below).

You can demonstrate to yourself the importance of the cornea for focusing by opening your eyes underwater. The cornea's index of refraction is very close to that of water. Thus, if the light travels from water to the cornea, it will not be bent much at all, and the cornea is essentially ineffective. You can't change focus enough with the lens to compensate, so your vision is vastly degraded. However, if you put on a pair of goggles or a dive mask, vision is restored because you have provided the air interface required by the cornea.

### 4.4.2 Spectral transmissivity of the eye's optics

The optical elements of the eye do not transmit all wavelengths of light equally well. Some of the elements such as the lens absorbs some of the light. The absorption of light is specified quantita-

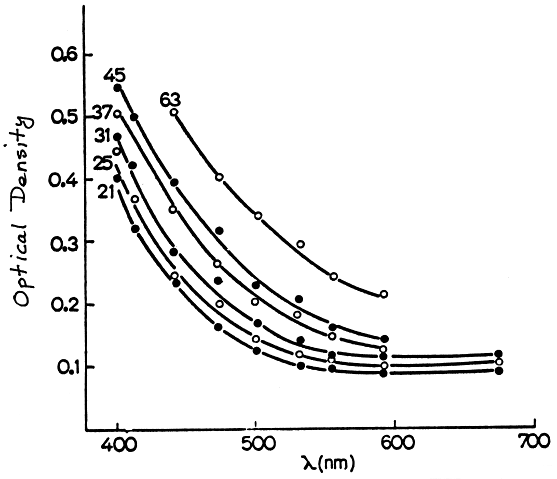


Figure 4.9: Optical density of the eye's optics as a function of wavelength and age. The density of the optics varies a great deal with wavelength with most of the absorption of light done by the lens. At 450 nm for an average 21-year-old, the optical density is about 0.22. This means about 40% of the light is absorbed by the optics. This amount increases with age. At the same wavelength for an average 63-year old, the optical density is about 0.52 which means about 70% of the light is absorbed by the optics. In words, the lens "yellows" with age, absorbing more and more of the short wavelength light. [Modified from Ruddock (1972, Fig. 3, p. 458). Original data from Said and Weale (1959).]

tively in terms of *optical density*, $D^5$. The most important contributor to this effect is the lens, which absorbs light more strongly at short than at middle or long wavelengths. The differential absorption of different wavelengths by the lens is shown in Figure 4.9. In addition, there is also the *macular pigment* which is an inert pigment in the foveal region of the eye that also absorbs at short wavelengths. The net result is that the optics of the eye act as a yellowish filter, letting through middle and long wavelengths but reducing the radiance of the light at short wavelengths that gets to the retina. This filter contributes to the falloff of light sensitivity at short wavelengths we have previously discussed for scotopic and photopic spectral sensitivity. [Why might the optics be designed this way?]

### 4.4.3 Accommodation

A glass lens such as that shown in Figure 4.6A-C has a fixed focal length. However, the lens of the eye can change its focal length in order to focus objects at different distances on the retina at different times. Changes in the focal length of the lens are called *accommodation.* To demonstrate accommodation, hold a finger as close to your eye as you can and still keep it in focus. Now concentrate on the finger and notice the blur of distant objects. Now reverse the process – look at the distant object and notice the blur of the finger.

[5] Optical density is defined as logarithm of the ratio of incident light $I$ to transmitted light $T$. That is, $D = log_{10}(I/T)$. A filter with a density of 1 transmits 1/10 of the incident light; with a density of 2, it transmits 1/100 of the light.

Accommodation is a change in focal distance brought about by changes in the shape of the lens. For far away objects a thin, flat lens is sufficient to bring the image to focus on the retina; whereas for near objects a thicker, more curved lens will be required. When you focus at a near distance, special muscles within your eye (the *ciliary muscles*) contract to make your lens thick – increase its power. When you focus far away, these muscles relax and allow your lens to become thin again – decrease its power. A young person with normal optics has a *range of accommodation* – the range of distances that can be brought into focus by accommodation – that covers 10 or more diopters, and goes from about optical infinity to within a few cm of the nose. [How close can you bring your finger to your nose and still keep it in focus?]

### 4.4.4 Common optical problems and their corrections

The human eye is susceptible to a variety of focusing problems, known collectively as *refractive errors*. An eye with a normal range of focus and no optical problems is called *emmetropic*. Common refractive errors include *myopia*, or nearsightedness, *hyperopia*, or farsightedness, and *presbyopia*, or "old eyes". Myopia, hyperopia, and presbyopia involve changes in the range of accommodation away from the normal range typical of emmetropia.

*Myopia* occurs when the whole accommodative range is moved in toward the eyeball. In consequence, close objects can still be focused readily, but distant objects cannot be brought into focus; the myopic child can't see the blackboard in class. Myopia often appears in adolescence and may be the result of long periods of focusing at near, or of continued growth of the eyeball while the eye socket (*orbit*) slows in its growth. This mismatch results in an eye that is too long for the available focusing power – the lens can't be made thin enough to focus objects at far distances. Myopia can be corrected by putting negative lenses – glasses or contact lenses – in front of the eyes. The negative lens diverges the incoming light, moving the accommodative range away from the eyes and back toward the normal range. (For example, a glasses prescription of -5.25 indicates that the myope needs a -5.25 diopter lens to move her accommodative range back to normal.)

*Hyperopia*, in contrast, results when the accommodative range is moved too far away from the eyeball, usually because the accommodative power of the lens is too limited. A person with hyperopia can focus far away objects fine, but cannot make his lens thick enough to focus close objects. A hyperopic child often has difficulty learning to read because she can't focus the type on the book page on her retina. Like myopia, hyperopia can be corrected with a properly chosen external lens. In this case a positive lens is needed to bring the accommodative range in closer to the eye. (For example, a prescription of +5 indicates the need for a 5 diopter positive lens to bring the hyperope's accommodative range back to normal.)

*Presbyopia* ("old eyes") refers to a condition that most people encounter after about age 40. Throughout the lifespan, from childhood on, the lens continues to grow, adding layers to its structure like an onion. As the lens grows it becomes less flexible, and harder for the muscle that controls accommodation to change its shape. So starting at about age 17, the range of accommodation narrows, and the near point of accommodation gradually moves out away from the nose. [Can you still focus as close as you used to?]

*Astigmatism* is another common refractive problem. People with astigmatism have an optical system that has one power for focusing lines or gratings at one orientation (say vertical) and a different power at the opposite orientation (say horizontal). In consequence lines at one orientation will be in focus with one level of accommodation, while lines at the opposite orientation will be in focus with a different level of accommodation. The astigmat can't ever focus both orientations in the visual scene at once, and always sees images with one kind of blur or the other. Astigmatism can be corrected by fitting the patient with an astigmatic (cylindrical) lens that compensates for the differential focusing power of the eye in the two orientations. (A prescription that looks like -5.25 +1.00 x 180 describes the correction needed by a myope who is also astigmatic. The first number gives the spherical correction; the second, the additional cylindrical correction to counter the astigmatism; and the third, the angle at which the cylinder axis is to be placed.)

Still other optical problems have to do with losses of transparency of the eye's optics. *Cataracts* are opacities in the lens which can greatly interfere with vision. They become increasingly common in old age. This compounds the fact that with normal aging the lens develops more and more of the yellow pigment that gives it its selective absorption of light at short wavelengths. Relatively routine surgical procedures have been developed to remove the cataractous lens and replace it with a plastic lens. A person who has just had a cataractous lens removed will often marvel at how the colors of things are restored – the blues look like they used to before the aging lens began to steal the short wavelengths of light away.

## 4.5 Optical information loss

Imperfections in the eye's optics limit the flow of information from the physical world to the visual neurons. Imperfections arise not only from refractive errors and opacities (discussed above), but also from additional factors we have yet to consider. How much degradation is there, what are the reasons for it, and how can we quantify it?

### 4.5.1 Why can't retinal images be perfect?

Even when the eye is in perfect focus, retinal images are not perfect, because the optics of the eye degrade the retinal image in several more complex ways. A point source in the world does not result in a point on the image. The cornea and lens refract the incoming light to form the retinal image. Insofar as their surfaces are not perfectly shaped, they will make blur circles instead of point images, leading to degradation of the overall image. The pupil, when it is very small, can degrade the image by diffraction. The vitreous too plays a role in image quality because it is not completely clear; the vitreous can contain imperfections including floaters: pieces of cellular debris that make a wash of scattered light and degrade the contrast of the image. Finally, internal reflections from various structures within the eye scatter the incoming light and further degrade the image. In the following sections we treat some of these problems in greater detail.

The refractive system of the cornea and lens produces four kinds of complex distortions: *chromatic, monochromatic, spherical* and *higher-order aberrations. Chromatic aberration* occurs because different wavelengths of light are refracted differently as they pass from one medium into another. When light passes through a traditional lens system the shorter wavelengths are bent more than the longer ones. The result is that if the long wavelengths are focused on the retina, the short wavelengths will be focused in front of the retina, and out of focus at the retina. Conversely, if the short wavelengths are focused on the retina, the long wavelengths will be aimed at a focus behind the retina, and out of focus at the retina. In short, it is impossible for the optical system of the eye to focus all wavelengths of light at the same time. In the laboratory, we often eliminate chromatic aberrations by using light of a single wavelength (*monochromatic light*).

*Monochromatic aberrations* occur because the surfaces of lenses are not perfect. If the curvature of a lens do not bend each ray of light by exactly the right amount, the image will not be perfect – it will be spread out due to the inaccuracies or irregularities in the lens surfaces. In *spherical aberration*, the center of the lens has one focal length and the periphery of the lens has another. With a large pupil both are used together, making a fuzzy image. More idiosyncratic irregularities in the lens and cornea can also produce *higher-order aberrations* that contribute to the imperfection of the image.

### 4.5.2 What pupil size makes the sharpest image?

The diameter of the human pupil ranges from about 1.5 mm to about 8 mm, as the person moves from bright to dim light. In general, optical quality degrades slightly from the center to the periphery of the lens. In addition, the power of the lens changes slightly from center to periphery, so optical degradation is worst overall when the pupil is largest and the whole lens contributes to the image. One might conjecture, then, that the way to optimize the image is to make the pupil small. But when the pupil is small, diffraction can make a significant contribution to imperfections in the image. How should one balance these two opposing demands? Given high light levels, is there an optimal pupil size?

## 4.6 Line-spread functions

How can the quality of the human retinal image be measured? Conceptually, one wants to form an image of a simple target – say, a point or a line – on the retina, with optimal focus. Then one can measure the diameter of the blur circle produced by the point, or the width of the image of the line, on the retina. These measurements would yield quantitative values for how far the image of the point or line was *spread* across the retina, due to optical errors of all kinds. That is, we can measure the *point spread function* or *line spread function* of the eye.

The classical measurements of the line spread function were made by Campbell and Gubisch in 1966. They used a special optical system similar to the system used in an ophthalmoscope – an instrument that allows another person to peer into your eye and examine your retina. This approach is known as the *double pass technique*, for reasons that will be clear below. The essential elements of their apparatus are shown in Figure 4.10.

First, Campbell and Gubisch used drugs to paralyze the subject's pupillary and accommodative systems temporarily – the pupil was dilated to its largest possible diameter, and the lens flattened to be focused at optical infinity. They set up a light source to project a line of light onto a subject's retina, to form the retinal image. The light was monochromatic in order to avoid the problem of chromatic aberration. Then they captured the light reflected back from the retina out of the eye, to make a *second image* (an image of the retinal image) in physical space outside the eye. Finally they optimized focus by placing glass lenses in front of the subject's eye until they produced the sharpest possible line spread function in the second image.

Campbell and Gubisch next scanned the second image with a photocell to quantify its luminance across space. They repeated the experiment with a series of artificial pupils – small holes in thin metal plates placed just in front of the eye – of different diameters, to simulate variations in pupil size. Of course, the second image has passed through the optics twice, not once as it would for normal vision. Campbell and Gubisch overcame this problem by factoring out the double pass

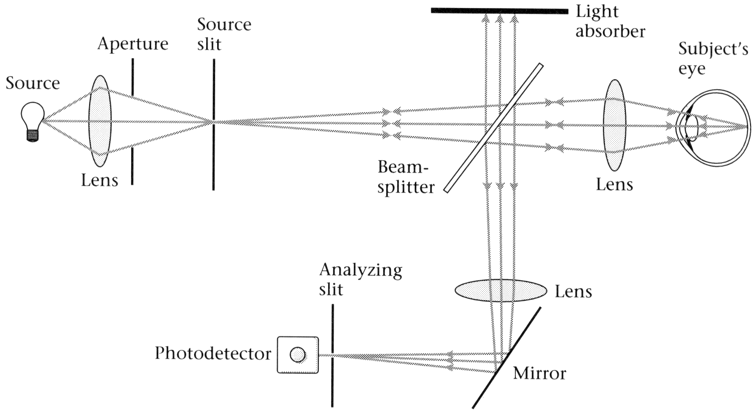


Figure 4.10: Double-pass apparatus. Light from the source (the light bulb at left) is imaged by a lens onto a slit. The slit makes the line that will be imaged on the retina. Light leaves the slit and passes through a beam splitter. The light reflected by the beam splitter is lost to the experiment at the light absorber at the top. The light transmitted by the beam splitter passes through another lens, enters the eye and is imaged on the retina (the *first* or *retinal image*). Some of the light from the retinal image is reflected back out of the eye. The returning light is again divided by the beam splitter, and the light transmitted is lost to the experiment. The light reflected by the beam splitter passes through another lens, is reflected by a mirror, and forms a *second image* in space. The second image is scanned by a photodetector, and the amount of light in the second image is plotted as a function of spatial position. The experimenter calculates from the second image to estimate the line spread function in the first (retinal) image. [From Wandell (1995, Fig. 2.3, p. 16).]

mathematically. The calculations produced an estimate of the line spread function in the retinal image[6].

Figure 4.11 shows both theoretical predictions and empirical measurements from Campbell and Gubisch's experiment. The thin lines show the theoretical predictions based on diffraction from the pupil alone. If diffraction were the only source of light dispersion in the retinal image, the line spread functions should follow these predictions. The image produced by diffraction varies with pupil diameter – the larger the pupil, the narrower the image. For large pupils the predictions based on diffraction have a width at half height of only about 0.2 minutes of arc – about five times narrower than the diffraction limit for small pupils.

The solid lines in Figure 4.11 show the measured line spread functions. For the small 1.5 mm pupil, the data come quite close to the limit set by diffraction. The width of the distribution at half height is about 1.3 minutes of arc. Thus, for small pupils, the optical quality of the retinal image is said to be *diffraction limited*, and not further degraded by the optics of the eye. Since diffraction is an absolute limit imposed by the laws of physics, the optics of our eyes do remarkably well in matching the physiological limit to the physical limit. The measured line spread function is actually narrowest at a pupil diameter of 2 to 2.4 mm. The distribution at its best is very tight

[6] Notice that the subject makes no judgments in this experiment. The measurements are entirely physical, not psychophysical. The subject needs only to hold still and fixate a fixation point.

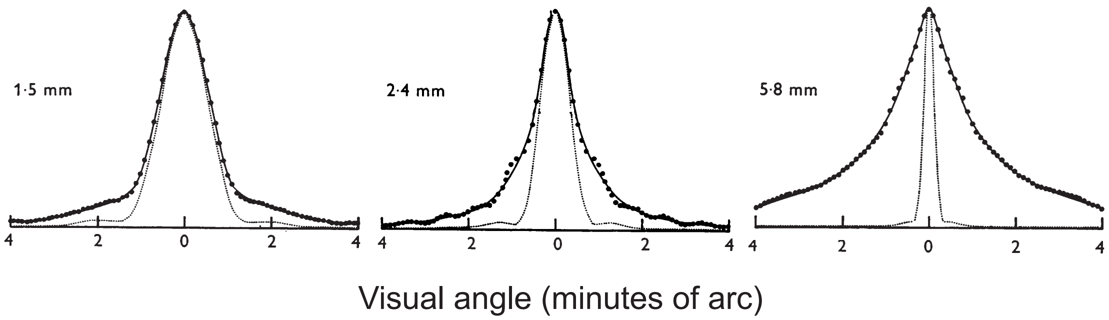


Figure 4.11: Campbell and Gubisch's results. The numbers beside each distribution show the diameter of the artificial pupil. The thin dotted lines show the expected diffraction limits if the pupil is the only factor contributing to the spread of light. The larger the pupil, the narrower the distribution predicted from diffraction. The solid lines show the measured line spread functions. The measured distributions approach their respective diffraction limits for small pupils, but not for large pupils. [Modified from Campbell and Gubisch (1966, Fig. 10, p. 570).]

– the width at half height is only about 1 minute of arc, and the central part of the distribution still approaches the diffraction limit. So for small and intermediate pupil sizes, what we have is as good as it gets.

But for large pupils, the observed line spread functions are much broader than the predictions based on diffraction, with a width at half height of about 2 minutes of arc. That is, for large pupils the quality of the retinal image is about ten times worse than the diffraction limit. When the pupil is large, other aspects of optical quality – spherical and higher-order aberrations, and scattered light – take over to limit the quality of the retinal image.

In short, given the properties of diffraction, the line spread function is potentially narrowest with a large pupil. For a large pupil, a more perfect optical system – one without such marked spherical and higher-order aberrations – could in principle yield a retinal image with a much narrower line spread function than we in fact have. But our visual systems do not take advantage of this opportunity. Viewed from this perspective, the optics of the human eye could be better. It is interesting to wonder why better optics haven't evolved for the human eye.

## 4.7 Adaptive optics: Improving on nature

Meantime, science is improving on nature. In the late 1990s an exciting new chapter was added to the story of optical quality: the use of adaptive optics. An adaptive optical system is one in which measurements of the optical quality of an individual eye can be made, and then fed back to correct the path of each ray in the incoming bundle of light rays, in such a way as to improve the quality of the retinal image for that particular eye. The system consists of two parts: a *wave front sensor* that allows measurement of the optical aberrations of a given eye, and a *deformable mirror* that allows correction of these aberrations.

In 1997, Liang and Williams (Liang and Williams, 1997; Liang, Williams, and Miller, 1997) built the first successful adaptive optical system for use with the human eye. Simplified optical

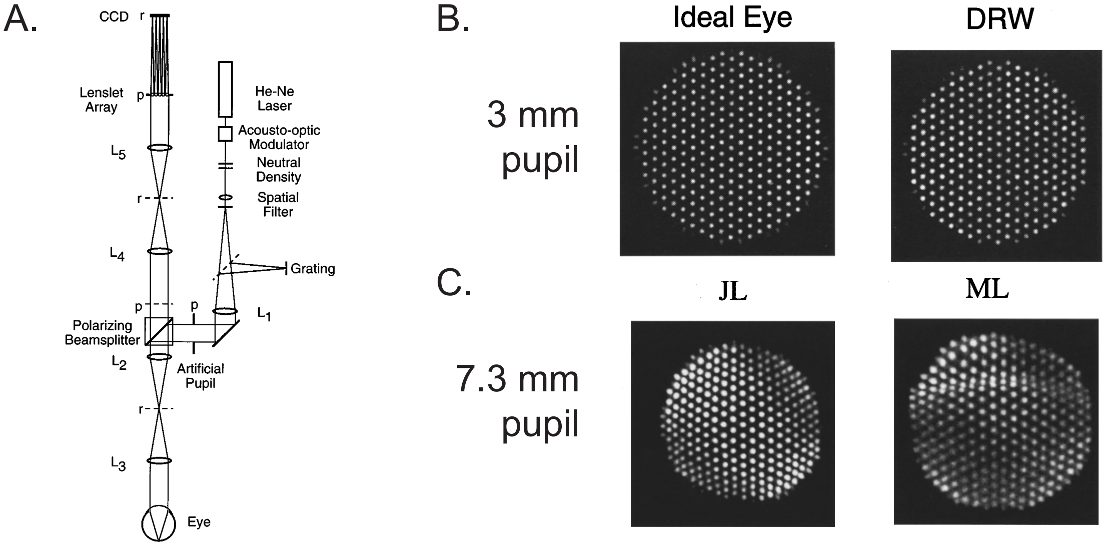


Figure 4.12: Adaptive optics: the wave front sensor. A. A tiny dot of light from the laser is imaged on the retina, and reflected back to the sensor array. Irregularities in the measured light in this array indicate aberrations in the optics of the eye. B. For a 3 mm pupil, the array is regular, both for an ideal eye (left) and for a real eye (right). C. For a 7.3 mm pupil, the array at the left shows a closer spacing of the dots at it's edge, indicating the presence of spherical aberration. The array at the right shows a set of irregular aberrations in the region where the eyelid normally rests against the cornea. [From Liang and Williams (1997, Fig. 1, p. 2874, Fig. 2, p. 2875 and Fig. 3, p. 2876).)]

diagrams of the system are shown in Figure 4.12 and 4.13. The first part of the adaptive optical system, the wave front sensor, is shown in Figure 4.12A. First, in a double-pass optical design similar to that used by Campbell and Gubisch, light from a laser is shined into the eye, and forms a tiny spot of light on the retina. Light from this spot is reflected back, and a cone of light – a wave front – emerges from the pupil. The trick is that different subparts of the wave front have passed through different parts of the eye's optics – different locations within the pupil, corresponding to different parts of the cornea and lens. By analyzing this beam of light, we can evaluate the eye's optics region by region. The light emerging from the eye eventually falls on a tightly packed array of light sensors. By analyzing the spatial irregularities in this array, a description of the overall irregularities in the optics can be derived.

Two dot arrays produced by the wave front sensor with a 3 mm artificial pupil are shown in Figure 4.12B. The first array was made by analyzing an "ideal" eye rather than a real one, and the regularity of the array is apparent. The second array was made by analyzing a real eye (subject DRW). When the pupil is small, as it is in this case, the array of dots remains regular – as we said earlier, the retinal image is usually diffraction limited for this pupil size, and spherical and other aberrations have little effect. In contrast, arrays produced with 7.3 mm pupils are shown in Figure 4.12C. In this case, there are obvious irregularities in the arrays, showing again that spherical and higher-order aberrations degrade the retinal image when the pupil is large.

The complete adaptive optics system, including the deformable mirror, is shown in Figure 4.13.

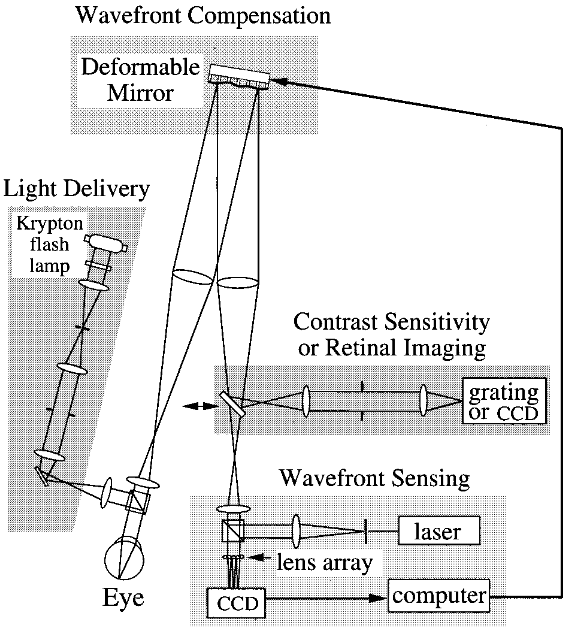


Figure 4.13: Adaptive optics: the complete optical system. In this figure the wave front sensor, including both the laser source and the array of sensors, has been folded up into the box at the lower right. The parts of the wave front sensor closest to the eye have been spread apart by adding lenses, so that the light from the laser can be bounced off the deformable mirror on the way to and from the eye. The deformable mirror is mounted on a set of tiny pistons. The surface of the mirror can be deformed by advancing some of the pistons and retracting others, in order to compensate for the particular aberrations of the eye being studied. [Adapted from Liang et al. (1997, Fig. 2, p. 2885).]

How does it work? A subject and his eye are aligned in the apparatus. Starting with the deformable mirror set to be flat, the apparatus is activated, and an array of measurements made. The computer makes an educated guess as to how the mirror should be deformed to make the array more nearly regular. This guess is implemented by deforming the mirror. New measurements are then taken, and a new deformation is tried. After 10 to 20 iterations, a highly regular array is usually produced. The deformations required to produce the regular array provide a description of the aberrations introduced by the particular eye being studied.

To what accuracy can the eye be corrected with adaptive optics? As of the late 1990s, with the pupil fully dilated, line spread functions could be made about a factor of two narrower than the best line spread function Campbell and Gubisch saw with a 3 mm pupil. The ultimate goal is to use a large pupil, for which the line spread function is potentially narrowest, and to reduce the line spread function to the diffraction limit calculated for the large pupil. If that goal were achieved, gratings of frequencies much higher than 60 cy/deg could be imaged on our retinas! [But could we see them? Think about it.]

## 4.8 Summary: Optical properties of the eye

We began this chapter by recalling our opening question of what limits spatial resolution as measured by grating acuity. One possible limit is the eye's optics and we use this chapter to introduce the optics of the eye. To fill in the needed background, we returned to square wave gratings, and introduced the specification of spatial frequency in both physical and quasi-physical units. We continued with a brief review of the nature of light and the properties of physical optical systems.

We then examined the properties of the human eye as a physiological optical system. The eye has two major optical components: the cornea and the lens. The cornea has about 40 diopters of optical power, and the lens a variable power between about 20 and 30 diopters. The variable power of the lens – accommodation – allows us to focus objects at different distances on the retina. Beyond questions of focus, the optical quality of the eye is limited by two major factors: the diameter of the pupil (which diffracts the light substantially when the pupil is small), and optical aberrations (which result in poor image quality when the pupil is large).

We then introduced the double-pass method used by Campbell and Gubisch to make *in vivo* measurements of the optical quality of the human eye. They showed that the optimal pupil diameter is in the range of 2 to 2.5 mm. At that pupil diameter the line spread function has a width at half height of about 1 minute of arc. But spreading the lines of a 60 cy/deg grating this much should lead to a major loss of contrast in the retinal image. Thus, optical quality could indeed be the major factor that limits grating acuity to about 60 cy/deg.

Finally, we introduced a more recent development in studies of the optics of the eye: adaptive optics. With adaptive optics we can measure the specific pattern of aberrations present in an individual eye *in vivo*, and use external instrumentation to shape the incoming light, in order to improve the quality of the individual's retinal image. Thus, we can potentially form images on the retina that are finer than those allowed by the eye's optics; and the question is, what will we see?

In the next chapter, we look at an alternative method of specifying the quality of an optical system: the modulation transfer function, or MTF. We also consider the effects of discrete sampling of the retinal image by the photoreceptors, and reconsider the question of whether the optics of the human eye limit our acuity.

# Chapter 5

# Optics and Vision

**Contents**

In Chapter 1 of this book, we introduced grating acuity as a fundamental psychophysical measure of spatial resolution. We then described three theories of how spatial resolution might be limited by the physiology: the quality of the optics; the discrete sampling imposed by the photoreceptor mosaic; or higher factors in the retina and/or the brain. We then posed the fundamental locus question that provides a unifying theme for the early chapters of this book: which of these stages actually limits grating acuity?

To answer this question, we need to re-address all of the alternatives introduced intuitively in Chapter 1. In Chapter 4 , we began this journey with the optics. We posed another of the most fundamental questions in vision science: how might one specify the physical quality of a lens or an optical system, and how good is our physiological optical system when such measurements are carried out?

Before we can continue, however, we need to introduce a considerable amount of background material, largely deriving from optical and electrical engineering. Toward that end, we begin the present chapter by addressing the concept of *linearity.* We then describe a kind of visual stimulus called a *sinusoidal* or *sine wave grating*, and try to explain why vision scientists use sinusoidal gratings as stimuli in vision experiments. We continue with an intuitive introduction to *linear systems theory.* In the next section, we introduce a new system property called the *contrast sensitivity function*, or *CSF.* The CSF, and its cousin the *optical modulation transfer function*, or *MTF*, extend the topic of spatial resolution to include sensitivity for large as well as small features of the visual scene.

Armed with the MTF and the CSF, we return to the goal of specifying optical quality as a physiological explanation of grating acuity. We describe three different techniques for measuring optical MTFs in the human eye. The first and third of these techniques depend on physical measurements. However, the second technique, *interferometry* – which is arguably the most accurate of the three – depends on psychophysical rather than on physical measurements. Moreover, interferometry serves our broader goals because, remarkably, it allows us to separate the limits of visual processing combined in the ordinary CSF into optical versus neural components.

We turn to the second of the factors that might limit grating acuity: discrete spatial sampling by the photoreceptor mosaic. We will see that discrete sampling also imposes some remarkable and unexpected system properties on our vision. Armed with this information, we return with increased sophistication to our original locus question – what limits grating acuity?

## 5.1 Some background: Why sinusoidal gratings?

Suppose you agree to be a subject in a vision laboratory. You walk in on the first day, and on a video screen you see a set of very fuzzy-looking stripes. Say hello! You have just been introduced to sinusoidal gratings, which are often used as stimuli in vision science. But why? It takes a while to explain.

### 5.1.1 Linearity

We begin with the fundamental mathematical concept of linearity (Wandell, 1995). Most generally, linearity has to do with the way two signals combine. At the simplest level, a linear system is one that does ordinary arithmetic: it just adds and subtracts, and multiplies and divides. In consequence, when a linear system is provided with two inputs simultaneously, the system's output (response) to one input is not affected by the level of response to another input. The output to the combination of inputs is just the sum of the outputs to the two separate inputs when each is provided alone. Figure 5.1 treats the topic of linearity, and Figure 5.1A shows an example of linearity operating in an optical system.

In symbols, suppose you have a system such that when you put in a signal $S$ you get out a response $R$. Further, let the signal $S_1$ yield the response $R_1$, and the signal $S_2$ yield the response $R_2$ when each is presented alone:

$$S_1 \rightarrow R_1$$

$$S_2 \rightarrow R_2$$

Then, if the system is linear, when you put in $S_1$ and $S_2$ simultaneously, you get out the sum of $R_1$ and $R_2$:

$$S_1 + S_2 \rightarrow R_1 + R_2.$$

From a vision science perspective, it is important that the stimuli and the responses can be highly dissimilar. Suppose that the input is a spot of light. The output can be the light distribution in the retinal image; or it can be a change in the state of some particular neuron along the visual pathway; or the perceptual report of the whole human subject. As long as the response to two

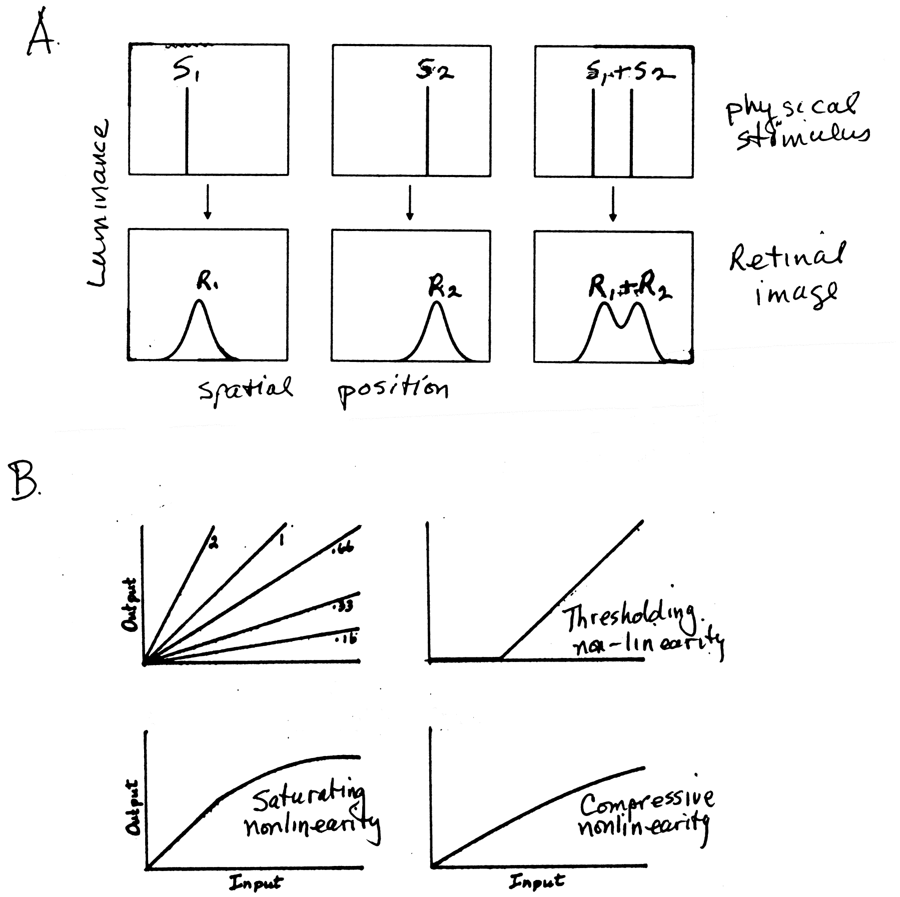


Figure 5.1: Linearity. A. The upper row shows three input stimuli – $S_1$, $S_2$, and the combined input $S_1$+ $S_2$. $S_1$ and $S_2$ are in different positions (e.g. on a video screen). The lower row shows the outputs (e.g. retinal images) $R_1$, $R_2$, and the combined response to $S_1$ and $S_2$. Because the response to $S_1$ + $S_2$ equals $R_1$ + $R_2$, the system is said to be linear. Any other output would reveal a non-linearity. B. Now imagine that $S_1$ and $S_2$ are spatially superimposed, for a stimulus of $2S_1$. In a linear system, multiplying the input by a constant, $k$, multiplies the output by the same constant. Thus, the output magnitude plotted against the input magnitude yields a straight line. Several linear input-output mappings with different gains (slopes) are shown in the upper left graph of B. The other three panels of B show examples of specific nonlinearities: thresholding, saturating and compressive.

stimuli is the sum of the responses to the individual stimuli, the system is linear (for the conditions tested).

An important special case of linearity arises when $S_1 = S_2$. In that case the input is $2S_1$; and if the system is linear, the output will be $2R_1$. Doubling the input doubles the output; and generalizing the argument, multiplying the input by any factor will also multiply the output by the same factor. In other words, in a linear system the relation between input magnitude and output magnitude is a straight line through the origin.

Several possible linear relationships between inputs and outputs are shown in Figure 5.1A. In a linear system, the slope of the line – the change in output per unit change in input – is constant across the whole range of inputs and outputs, and is called the *gain* of the system. The input-output relationships of several linear systems with different gains are shown.

Now that we've defined linearity, let's ask, what is a *non-linearity*? This question can't be answered in general, because there are many different ways in which linearity could fail[1]. Three examples of non-linearities are shown in Figure 5.1B. For a *thresholding non-linearity*, the system does not respond to very small inputs, and the output increases only after a threshold input level is reached. In the graph, this is shown by a function that is initially flat and then increases linearly. For a *saturating non-linearity*, the output approaches an asymptotic value at high levels of input. As shown in the graph, the output of the function cannot go beyond a certain value. For a *compressive non-linearity*, the output continues to grow with the input, but it grows more and more slowly at higher and higher input values. But it does not have an asymptote. In all three cases, the consequence of the non-linearity is that the same change in input, across different ranges of input, leads to different changes in output. For example, we will see saturating non-linearities in the photoreceptors in Chapter 6 (Figures 6.11 and 6.12).

The concept of linearity will prove useful for many issues in vision. The first is a quantitative evaluation of the human eye as an optical system. The optics are linear. But we need to further set the stage to appreciate this idea.

### 5.1.2 Sinusoidal gratings

We turn next to sinusoidal gratings. A set of four vertical sinusoidal gratings is shown in Figure 5.2. Figure 5.2A shows two graphs of luminance as a function of position across the face of the video monitor. Each of these graphs traces out an undulating mathematical function called a *sine wave* or *sinusoid*. The horizontal lines through the graphs show the average luminance values. Figure 5.2B shows three simulated pictures of the video monitor, each displaying a sinusoidal grating corresponding to the one graphed just above it. Speaking non-mathematically, it is immediately obvious that sinusoidal gratings are a lot like the square-wave gratings introduced in Chapter 1, but their transitions are smooth rather than sharp. Both of the gratings in Figure 5.2A and B have the same mean (space-average) luminance, and all have (nominally) 100% contrast, with the actual contrast of the pictures in B depending on the properties of the reproduction process used in printing the figure.

In addition to mean luminance, sinusoidal gratings have three parameters: *spatial frequency,*

[1] Teller was initially confused on this point. Every time a colleague spoke about non-linearity, he seemed to be talking about something different! This is because linearity and non-linearity are asymmetrical opposites. There's only one kind of linearity, but many different kinds of non-linearity. It's like a person being normal versus eccentric – there are fewer ways to be normal, and more ways to be eccentric!

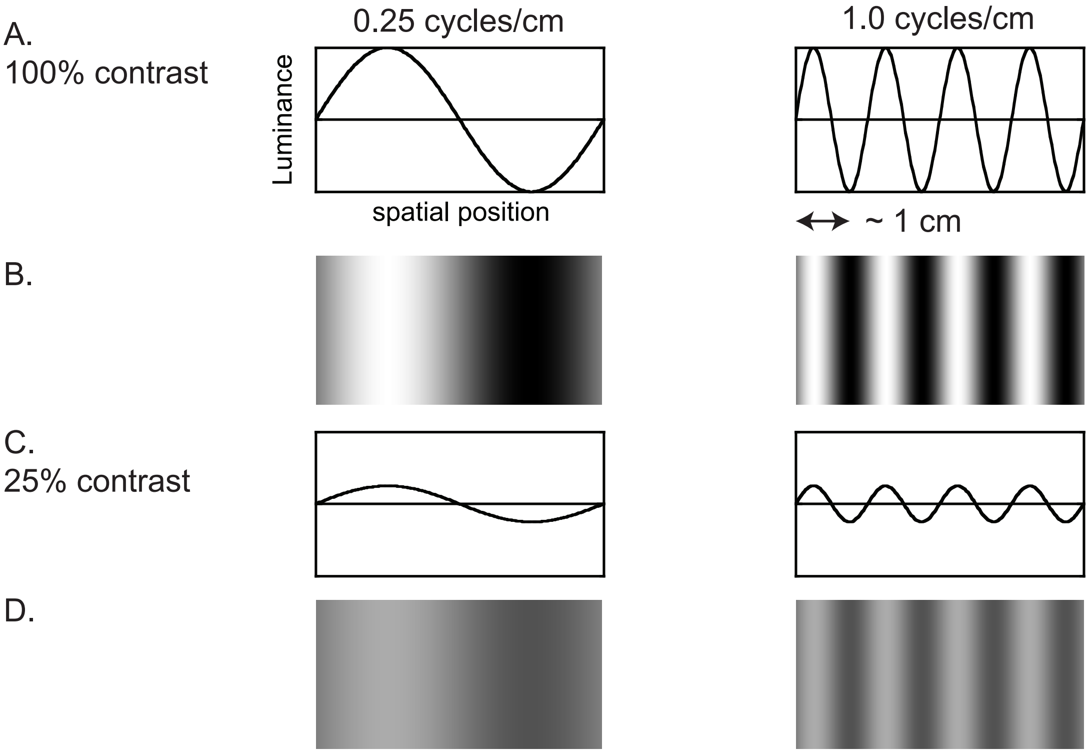


Figure 5.2: Sinusoidal gratings: variations of spatial frequency and contrast. A. Graphs of the variation of luminance with spatial position across the video screen, for sinusoidal gratings of two different spatial frequencies: 0.25 and 1.0 cy/cm, all at 100% contrast. B. Simulations of how these two sinusoidal gratings would appear. C. Graphs of the same two gratings at 25% contrast. D. Simulations of the appearance of these two 25% contrast gratings.

*contrast*, and *phase*. As in the case of square wave gratings, the *spatial frequency* of a sinusoidal grating describes the number of repeats of the spatial variation per unit distance. Spatial frequency can be expressed as the number of cycles of the sinusoidal grating per cm on the video screen (cy/cm) – here, 0.25 and 1.0 cy/cm – or more typically in vision science, in terms of cy/deg.

The *contrast*, or *modulation*, of a sine-wave grating describes the amount of variation of luminance around the mean luminance level. The peaks and troughs of a sine wave are symmetrical about its mean value, $L_o$. The minimum luminance that the troughs can take ($L_{min}$) is zero. Thus, by symmetry, if the mean luminance is fixed at($L_o$), the maximum luminance that the peaks of the sine wave can take ($L_{max}$) is $2L_o$. The contrast (often called Michelson contrast) of a grating is defined as

$$Contrast = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}$$

and it takes values between 0 and 1. Contrast is often expressed as a percent so that a contrast of 1 becomes 100%. The gratings in Figure 5.2A and B, whose luminances nominally vary from zero to $2L_o$, have nominal 100% contrast. Figure 5.2C and D show two sinusoidal gratings of the same spatial frequencies shown in Figure 5.2A and B, but at a lower nominal contrast level, 25%. Contrasts lower than 25% cannot be represented accurately in print. However, as we shall see,

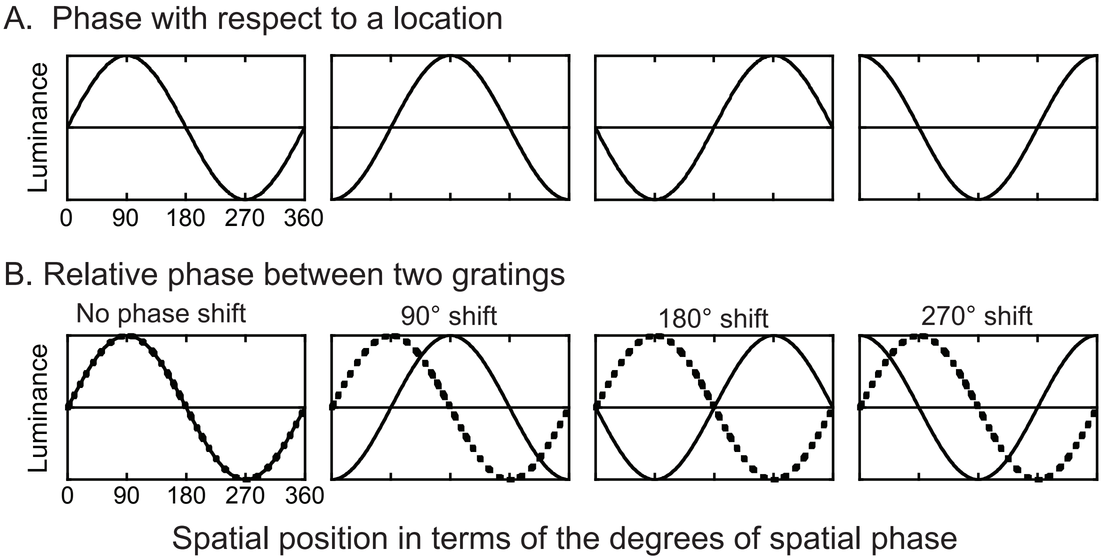


Figure 5.3: Sinusoidal gratings with variations of phase across the columns of the figure. A. A sinusoidal grating in various phases with respect to a fixed location (e.g 0 on the abscissa). B. Two gratings are shown to illustrate various phase relationships. In the leftmost panel, the two gratings are identical. In the others, the grating shown by the solid curve is shifted to the right while the grating shown by the dashed curve remains in the same location. The amount of the shift is quantified by the change in phase.

contrasts as low as 1%, and even less, are visible to human subjects under the right conditions.

The third parameter of the sinusoid is its *phase*. The phase of a sinusoidal grating describes its spatial displacement relative to a fixed reference point or to another grating. Figure 5.3A shows a series of sinusoidal gratings that are shifted in location with respect to a fixed reference point. Figure 5.3B illustrates two gratings that differ in their relative phase. The heavy dashed curve remains at the same location while the light solid curve shifts to the right.

One of the historically earliest mathematical uses of sinusoidal functions is to describe some of the properties of circular motion. From this history comes the custom of designating phase in degrees in the same way that circles are designated – 360 degrees is a complete circle, or a complete cycle of the grating. As shown in Figure 5.3B, two gratings of the same spatial frequency that are aligned in space are said to be *in phase*; i.e. shifted with respect to each other by 0 or 360 degrees. Two gratings shifted by one-quarter of their period (cycle length) are said to be shifted by 90 degrees, or in *quadrature phase*. A half-period shift is a shift of 180 degrees, and two gratings in this relationship are said to be *out of phase* or in *counterphase*. And a three-quarters of a period shift is a 270 degree shift – again, a quadrature shift, but in the opposite direction.

### 5.1.3 Shift-invariant linear systems

In considering the optics of the eye, there is one further simplification we can draw upon. Not only is the optical system linear, but it has shift invariance. That is, the effect of the optics does not vary across the image[2]. This gives the optics of the eye some very simple properties. In particular, if one inputs a sinusoidal grating, the output in the image is also a sinusoidal grating. Perhaps with its contrast attenuated, but still having a sinusoidal shape. Such invariance of shape holds only for harmonic functions such as sinusoids and does not hold for most functions such as a square wave. Imperfect optics (even if linear) blurs the sharp edges of a square wave so the output is no longer a perfect square wave.

## 5.2 Fourier analysis and linear systems theory

### 5.2.1 Fourier analysis

Why are sinusoidal gratings interesting? One hint has just been given in that they do not change shape when passing through the eye and forming an image. But there is more. In the late 1700s, the French mathematician Jean Baptiste Fourier described an important mathematical theorem. Fourier's theorem states that *any signal that varies in space or time can be described mathematically as the sum of a set of sinusoids that vary in frequency, amplitude, and phase.* By starting with sinusoids of the requisite frequencies, manipulating amplitudes and phases, and summing the sinusoids appropriately, any more complex function can be generated. Breaking down a complex pattern into its frequency components is called *Fourier analysis*, and recombining them to make the original pattern is called *Fourier synthesis.*

An example of Fourier analysis – the decomposition of a square wave grating into its sinusoidal components – is shown in Figure 5.4. Fourier's theorem asserts that Fourier analysis can similarly be carried out on any spatial scene (although the results are much more complex than the square wave example).

The above description is intended to convey the conceptual basis for using Fourier analysis at an intuitive level. However, we need to point out two complications. First, a stimulus cannot be specified unambiguously by its spatial frequency content alone – the phases of the components must also be specified. And second, grating patterns are a particularly simple class of stimuli. They are *one-dimensional*, in the sense that luminance varies along only one spatial dimension, and is constant along the other. In contrast, a scene is *two-dimensional* in the sense that its luminance varies along both vertical and horizontal dimensions. As it turns out, Fourier analysis of two-dimensional spatial patterns reveals spatial frequency components in an infinite number of different *orientations*: vertical, horizontal, left diagonal, right diagonal, and every orientation in between. Thus, the true amplitude spectrum of a scene is more complex than that shown in Figure 5.5, and includes spatial frequency components and their amplitudes at multiple orientations.

### 5.2.2 Why?

Why do we care about sinusoidal gratings and Fourier analysis? Most basically, vision scientists are in the business of trying to understand how the visual system codes and recodes visual stimuli. Natural visual stimuli vary in an infinite number of ways, and before the arrival of Fourier analysis vision scientists had no system in which to describe any possible image. Fourier's theorem gives us such a descriptive system – we can in principle specify any stimulus in terms of its Fourier components. For an in depth treatment of these topics, see Bracewell (1978).

[2] Shift invariance holds close to the line of sight but fails in the far corners of the eye.

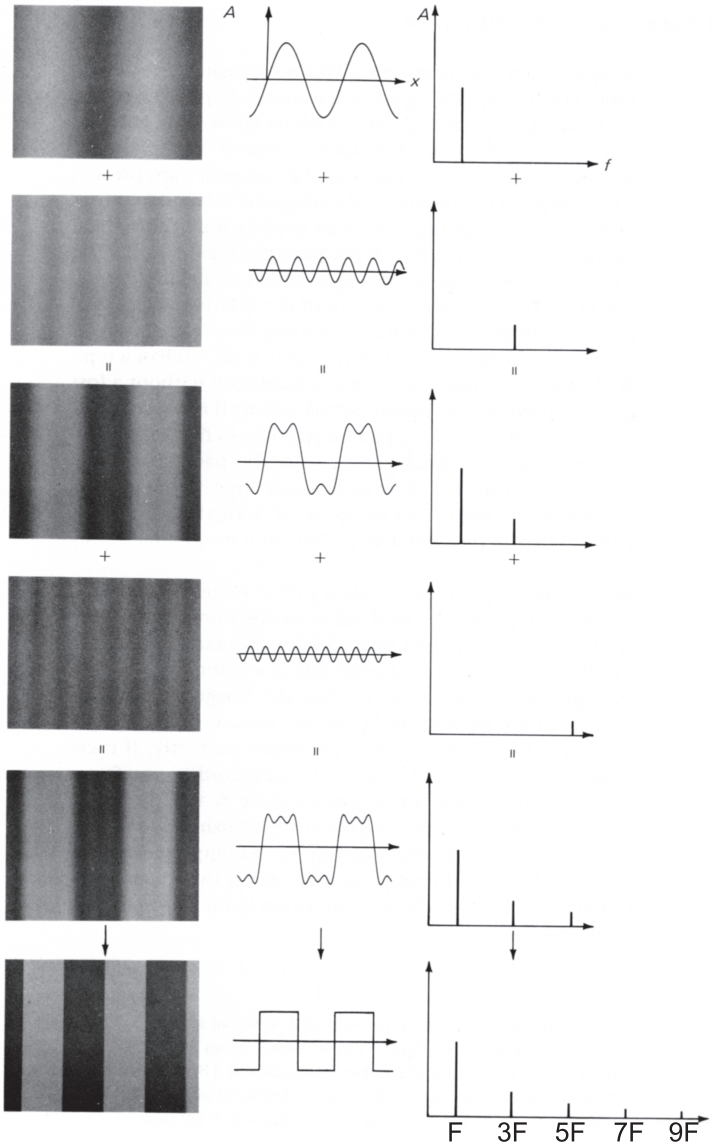


Figure 5.4: Fourier components of a square wave grating. The top row shows a sinusoidal grating of frequency f. The second row shows a sinusoid of three times the frequency, 3f. The third row shows f and 3f superimposed, yielding a somewhat squared-off waveform. The fourth row shows a sinusoid of spatial frequency 5f. The fifth row shows f, 3f, and 5f superimposed, resulting in more squaring off. Finally, the sixth row shows all of the odd harmonics of f (f, 3f, 5f, 7f, ...). When superimposed in the appropriate phases and amplitudes they produce a square wave grating. Each column shows a different way of characterizing the grating: with a picture; with a graph of luminance as a function of position; and, with a graph of the amplitude of the sine wave at each spatial frequency. [Levine and Shefner (1991, Fig. 10.6, p. 217).]

In addition, the existence of Fourier analysis led some vision scientists to the fascinating idea that, at some level of coding, the visual system might actually use a Fourier-like description to represent visual stimuli but that is getting way ahead of our story[3] (see Chapter 16).

### 5.2.3 Linear systems theory

*Linear systems theory* is a set of concepts that originated in electrical and optical engineering, and were imported into vision science in the 1950s. Suppose that we are interested in a system such as an audio amplifier, or a lens, or the human visual system as a whole. Also suppose that our goal is to be able to predict the output of the system for any arbitrary input. Rather than measuring the response to each possible input, perhaps it is possible to measure the response of the system to just a few carefully selected inputs, and use these responses to derive a general description of the system. Perhaps this description together with a standard algorithm could then be used to predict the system's response to any arbitrary input.

Now, Fourier's theorem tells us that any pattern can be represented as the sum of a set of sinusoids. Perhaps if we knew the response of the system to a sinusoid of each spatial frequency, we could calculate its response to any arbitrary visual pattern! In pursuit of this goal, then, we would need to begin by *establishing a function that specifies the response of the system to sinusoidal inputs of each different spatial frequency.*

Once this function is established, we could in principle Fourier analyze the pattern we are interested in, specifying it in terms of its frequency components; multiply each frequency component by the gain of the system at that frequency; and use Fourier synthesis to calculate the system's response to the pattern. This process is schematized in Figure 5.5. Of course, the results will be much more satisfying if the system is linear, but the conceptual framework can be interesting even if it is not.

## 5.3 Modulation transfer functions (MTFs)

To embark on this path, we can begin by measuring the response of a system – such as the optics of the eye – to sinusoidal gratings of different spatial frequencies. The response is specified in terms of the *contrast ratio*, or *gain* – the output contrast for a given input contrast – at each spatial frequency. If the system can be assumed to be linear, as optical systems are, the resulting function is called a *modulation transfer function* (*MTF*). The term MTF is a particularly meaningful one, as the MTF specifies the fraction of *modulation* (contrast) that the *system* transfers from the input to the output. For example, if at a particular spatial frequency the input contrast is 100% and the output contrast is 25%, the contrast ratio, or gain, is 0.25 at that spatial frequency.

On the other hand, if one wants to measure the response of the whole human visual system one cannot measure a MTF because there isn't an output image to measure. Instead, one can measure behavior. In particular, if you measure a common behavior (the just detectable contrast threshold)

[3] When this approach to vision was first introduced in the 1960s, Teller and her irreverent young colleagues were reminded of an old quip about Freudian theory: "You shouldn't criticize psychoanalysis until you've been psychoanalyzed." We provided an update: "You shouldn't criticize Fourier analysis until you've been Fourier analyzed!" We proceeded to amuse ourselves at meetings by Fourier analyzing the various senior scientists in our minds – the slim ones perhaps having high-pass amplitude spectra, and the round ones being particularly well endowed with low spatial frequency components.

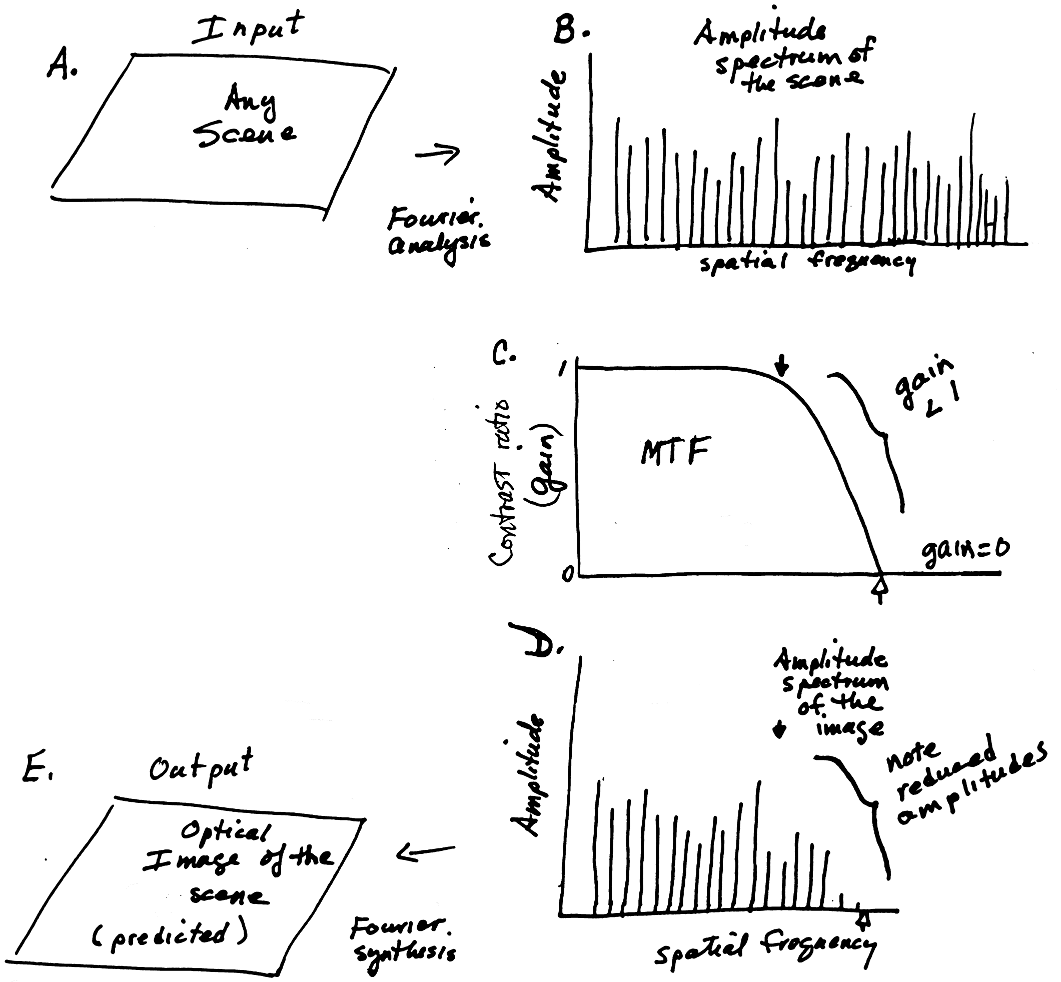


Figure 5.5: Linear systems theory. A schematic illustration of how Fourier analysis and synthesis, in combination with a modulation transfer function (MTF) for the system, allows prediction of the output of a linear system to any arbitrary input. Panel A represents any scene. Panel B represents analysis of the scene into its Fourier components (the amplitude spectrum of the input). Panel C represents the MTF of the system. The solid arrow shows the spatial frequency at which the contrast ratio (gain) falls below 1; the system reduces the amplitudes of the spatial frequencies above that value. The open arrow shows the spatial frequency at which the contrast ratio falls to zero; the system eliminates all spatial frequencies above that value. Panel D shows the amplitude spectrum of the image formed by the system. Each of the spatial frequencies represented in Panel B is multiplied by the contrast ratio of the MTF at that spatial frequency as represented in Panel C. Finally, Panel E represents the recombination of components across spatial frequency, to produce the predicted output (optical image).

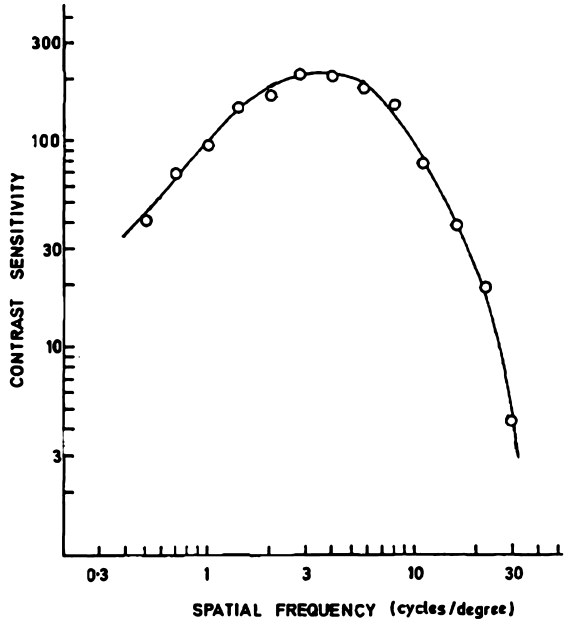


Figure 5.6: A psychophysical contrast sensitivity function, or CSF. Contrast sensitivity – the reciprocal of the contrast threshold – is plotted as a function of spatial frequency. Sensitivity is maximal in the middle of the function, at 3 to 5 cy/deg, and falls off at both lower and higher frequencies. [Modified from Robson (1966, Fig. 1, p. 1141).]

for all spatial frequencies, one can infer the stimuli that yield an equivalent effect on behavior. This function is called a *contrast sensitivity function* (*CSF*). CSFs are widely studied psychophysically, as system properties of human vision.

Obviously, the next thing we would like to do is determine both MTFs and CSFs for the human visual system. In particular, the MTF of the optics of the eye would move us toward the goal of the chapter by providing us with our much-desired method for defining human optical quality. But as it turns out, the simplest and cleanest measurements of the optical MTF are derived from measurements of the psychophysically measured CSF! So let us move back to the psychophysics laboratory.

## 5.4 A new system property: Contrast sensitivity functions (CSFs)

To measure a CSF, we ask a subject to sit in front of a video monitor, and present him with sinusoidal gratings of different spatial frequencies in turn. For each spatial frequency, we measure the subject's *contrast threshold.* That is, we measure the contrast on the video screen required for the subject to just barely detect the grating. We then take the reciprocals of the contrast thresholds to generate sensitivity values, and plot the subject's sensitivity as a function of spatial frequency.

Figure 5.6 shows a relatively early measurement of the CSF of a single human subject. Psychophysical CSFs like this one have three interesting features. First, the CSF is band pass: A band-pass function has a maximum sensitivity in the middle, and falls off to either side. For this specific function, sensitivity is maximal in the vicinity of 3 cy/deg, and falls at both lower and higher spatial frequencies. Second, the maximum contrast sensitivity is about 200 which corresponds to a minimum detectable contrast of 0.5% (1/200). This is a remarkably high level of sensitivity – in the spatial frequency range from 2 to 10 cy/deg, the subject can detect a sinusoidal luminance wiggle of less than 1% across the video monitor. Third, sensitivity falls off sharply at high spatial frequencies. By definition, the highest available stimulus contrast is 100%. Correspondingly, the highest measurable threshold is a threshold that requires 100% contrast (a sensitivity of 1); and the *high frequency cut-off* is defined as the spatial frequency at which contrast sensitivity falls to 1. Between 5 and 60 cy/deg, the visual system shows a sensitivity loss of more than two orders of magnitude. If we extrapolate the smooth curve fitted to the data, it will fall to a sensitivity of 1 somewhat below 60 cy/deg (other measurements are closer yet). Thus, the high frequency cut-off of the CSF is similar to our original estimate of grating acuity: 60 cy/deg.

Is this a coincidence? No. To measure grating acuity, we set the contrast of a square-wave grating to 100%, and vary the spatial frequency to find the highest visible spatial frequency at 100% contrast. To measure the high frequency cut-off, we vary the contrasts of high frequency gratings to find the spatial frequency with a contrast threshold of 100%. Other than the choice of which stimulus variable is manipulated, the two are the same. Contrast sensitivity functions, thus, include an estimate of the acuity threshold along with information about both the overall contrast sensitivity and how the relative sensitivity varies with spatial frequency.

Contrast sensitivity functions are of interest for several reasons. First, CSFs are recognized as the fundamental threshold-level descriptors of spatial vision. They generalize the concept of grating acuity, and describe the sensitivity of our eyes to different spatial patterns in the visual scene. Second, the measurement and modeling of CSFs is fundamental to the multiple spatial frequency channels approach to vision, which we describe in detail in Chapter 16.

And third and most immediately relevant, CSFs can be used to derive a new, more functional estimate of optical quality. Now, instead of asking about line spread functions, we can ask: what is the optical MTF of the human eye? We next describe three different techniques for measuring optical MTFs, and compare the results.

## 5.5 MTFs for the human eye

### 5.5.1 The double-pass technique

In 1963, Gerald Westheimer used the double-pass technique to measure human MTFs, using sinusoidal gratings rather than lines as the physical stimuli as we have described for Campbell and Gubisch (1966). For a grating of each spatial frequency, he compared the contrast in the physical stimulus and the contrast in what we have called the second image, and estimated the contrast in the retinal image. The results are expressed in terms of the *modulation transfer*, or *contrast ratio*: the contrast in the retinal image as a fraction of the contrast in the physical stimulus.

Figure 5.7 shows the estimated MTFs for a 3 mm pupil on log/log axes. MTFs, like CSFs, have several interesting properties. First, unlike CSFs, optical MTFs show flat contrast ratios across the low-spatial frequency range, below 1 to 5 cy/deg. Whereas the CSF is band pass, the MTF is low pass. Second, like CSFs, optical MTFs show high values at low spatial frequencies. Below a certain spatial frequency, the contrast in the retinal image is equal to the contrast in the physical stimulus, which is a contrast ratio of 1. In other words, no modulation is lost in the optics of the eye at low spatial frequencies. In short, the optics are low-pass rather than band pass.

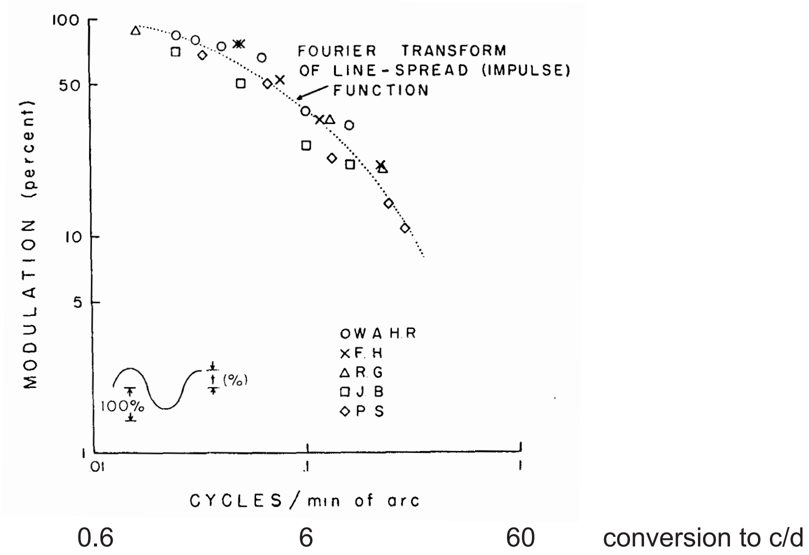


Figure 5.7: Optical modulation transfer functions, or MTFs, measured with the double-pass method. The abscissa shows the spatial frequency of the physical grating. But watch out, it is in cycles per min rather than cycles per degree (the conversion is also shown). The ordinate shows the contrast ratio (or modulation transfer) of the optical system – the output contrast (the contrast in the retinal image), as a fraction of the input contrast (the contrast in the physical stimulus). A curve of this shape is called a *low-pass* function. [From Westheimer (1963, Fig. 15, p. 93).]

Third, for higher spatial frequencies the contrast in the retinal image decreases, first slowly and then more and more rapidly, with increasing spatial frequency. The spatial frequency at which the MTF crosses the ordinate value of 0.01 (a contrast ratio of 0.01, or 1%) is called (somewhat arbitrarily) the *high frequency cut-off* of the MTF[4]. Finally, one can calculate using linear systems theory the relationship between a line-spread function and the optical MTF. The MTF predicted by a separately measured line-spread function is shown by the dotted curve.

In summary, unlike the CSF, the MTF is low pass. The high-frequency cutoff of the low-pass MTF corresponds nicely to the high-frequency cutoff of the band-pass CSF. But the MTF does nothing to account for the attenuation of sensitivity at low frequencies of the band-pass CSF.

### 5.5.2 Interferometry

A second approach to measuring the optical MTF is through the technique of *interferometry*[5]. It begins from the remarkable fact that we can produce sinusoidal gratings on the retina in two very

[4] Notice that there is a practically motivated difference in the definition of the high spatial frequency cut-off of the CSF versus the MTF. CSFs are psychophysical measurements, and the limit is taken to be the spatial frequency at which a contrast of 100% is required for threshold (no higher contrasts are available). MTFs are physical measurements, and the limit is taken to be the spatial frequency required for a contrast ratio of 1 (measurements at lower contrast values would be difficult because of noise). We will ignore this difference.

[5] For us, the story of interferometry illustrates the rich interplay among disciplines that is the charm of vision science.

different ways. The first way is, of course, by viewing a physical grating directly on a video monitor. Since the optics of the eye are linear, the retinal image will also be a sinusoidal grating, with its contrast degraded in accord with the optical MTF.

The second way of producing sinusoidal gratings on the retina makes use of the physical phenomenon of diffraction. Using a specialized optical system called an *interferometer*, two beams of light from a single (laser) source are focused at two different points in the plane of the subject's pupil (cf. Figure 4.4). The two beams then diverge to make overlapping fields of light, forming interference fringes on the subject's retina.

The interference patterns have two properties dear to the hearts of vision scientists. First, as it turns out, the variation of luminance across the interference fringes closely resembles the variation across a sinusoidal grating. And second, the formation of the grating pattern on the retina does not involve the focusing properties of the eye! The light passes through the cornea, lens, etc, but it is not focused because there is no image to focus. Defects of the eye can reduce the amount of light that creates the interference fringe, but will not reduce the contrast of the fringe. Thus, remarkably, nature allows sinusoidal gratings to be produced on the retina in two different ways, one (with *direct*, or *ordinary viewing*) that makes use of the optical focusing properties of the eye, and the other (with interferometry) that does not.

The next step is to make psychophysical measurements of CSFs, using each of the two kinds of sinusoidal gratings in turn. For direct viewing, the subject views sinusoidal gratings on the usual video monitor, and adjusts the contrasts of the gratings to threshold. These measurements provide us with a *direct*, or *ordinary CSF*, that describes the transfer of contrast through the whole sequence of stages of the visual system, including both the optics and the neural visual system.

For interferometric measurements, the subject is positioned in an optical device called an *interferometer*. He sees the interference fringes as sinusoidal gratings, and again adjusts the contrasts of the gratings to threshold. These measurements provide us with an *interferometric CSF* that, remarkably, describes the transfer of contrast through the neural visual system – retina and cortex – in isolation, but omitting any focusing by the optics of the eye. Since it is not influenced by the optics of the eye, it must depend only on neural factors; and for this reason, the interferometric CSF has also been called a *neural CSF*. We will return to neural CSFs below.

The results of a classic study by Campbell and Green (1965) are shown in Figure 5.8. The main panel of the figure shows both ordinary (open circles) and interferometric (solid contour) CSFs. The same subject was used for both sets of measurements. Since the ordinary CSF has been degraded by the optics of the eye and the interferometric CSF has not, it makes sense that the interferometric CSF falls above the ordinary CSF throughout the measured spatial frequency range.

Campbell and Green then determined the ratio of contrast sensitivities between the ordinary CSF and the interferometric CSF, as shown in the insert at the top of the figure. These contrast ratios provide an estimate of the reduction in contrast of an ordinary grating caused by passing through the optics of the eye. As we found before, the optics are low pass rather than band pass like the CSF. These results provide us with our second estimate of the optical MTF. (The two MTFs will be compared below.)

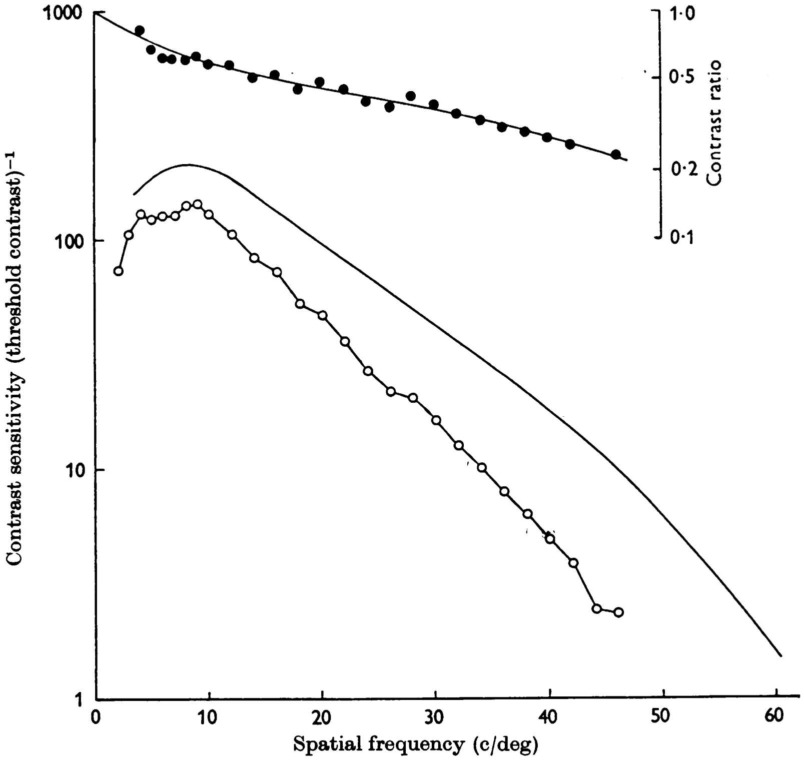


Figure 5.8: An optical MTF estimated from interferometry. In these graphs, the ordinate is logarithmic and the abscissa is linear. In the main graph, the connected open circles show an ordinary CSF measured by direct viewing of a video monitor, and therefore influenced by both optical and neural factors. The line shows an interferometric CSF based on interference fringes, and therefore influenced by only neural factors. These two functions are combined in the insert to estimate the contrast ratio between the ordinary and interferometric CSFs which is influenced by only optical factors. This function provides a second estimate of the optical MTF. [From Campbell and Green (1965, Fig. 9, p. 586).]

### 5.5.3 Adaptive optics

More recently, as part of their development of adaptive optics (see Chapter 4), Liang and Williams (1997) have added a third technique for estimating optical MTFs. The wave front sensor of the adaptive optics apparatus allows the description of optical aberrations, and from them theoretical estimates of optical MTFs can be calculated. The details of the technique, however, are beyond the scope of this book.

### 5.5.4 Comparison of three estimates of optical MTFs

When quantitative comparisons of the data of Figures 5.7 and 5.8 are carried out, they reveal that estimates of the optical MTF based on interferometry show higher contrast ratios, indicating higher optical quality, than do estimates based on the double-pass method. However, it is impossible to tell whether these differences are real; or whether they are based on individual differences among

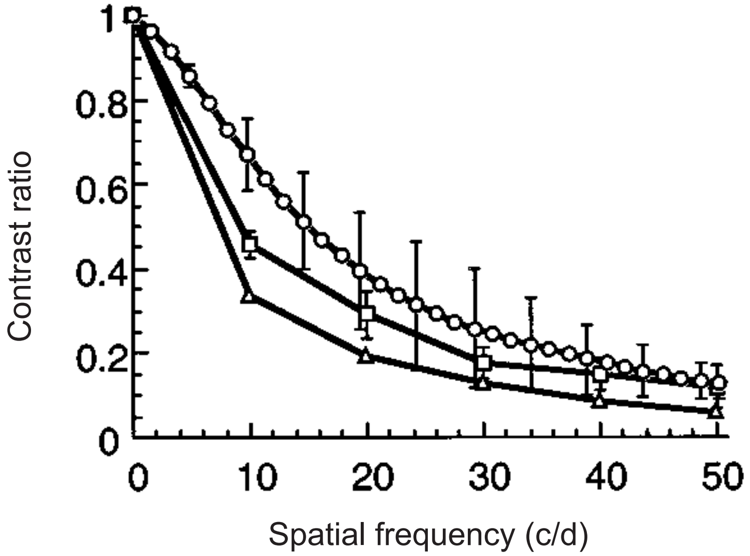


Figure 5.9: Optical MTFs measured with three techniques. Here linear-linear axes are used for the MTF. The same three subjects were tested in each case, and the data are averaged across the three. The triangles, squares and circles show the results of the double-pass, interferometric, and adaptive optics techniques respectively. [From Liang and Williams (1997, Fig. 6, p. 2877).]

the eyes of the subjects tested in the two different experiments, or on the different kinds of artifacts that potentially impact the different techniques.

To attack this question, Williams, Brainard, McMahon, and Navarro (1994) repeated measurements of optical MTFs with both the double-pass technique and the interferometric technique, under tightly parallel conditions, on the same subjects. Moreover, Liang and Williams (1997) used adaptive optics to derive a third estimate of MTFs on the same three subjects, keeping other factors as similar as possible.

The results from all three techniques, averaged across three subjects, are shown in Figure 5.9. The measured optical MTFs are poorest with the double-pass technique, medium with interferometry, and best with adaptive optics. However, individual differences are large, especially for the wave front sensor technique; and the differences among techniques are only about the size of the individual differences among subjects.

Which of the three techniques best characterizes the real MTF of the eye's optics? The double-pass technique is suspect because it depends upon light scattered within the eye and reflected from the retina. Several different retinal structures in different depth planes could in principle contribute to the scattered and reflected light. These variations could artifactually reduce the contrast in the second image, and thereby reduce the estimate of contrast in the retinal image. Consequently, the estimates of the MTF provided by the double-pass technique are probably too low.

The interferometric technique, on the other hand, is based on the actual visual performance of a human subject. That is, it provides a measure of the optical image in whatever plane within the eye is actually used for the quantal absorptions that initiate the visual signal. And, the new adaptive optics measurements are much quicker than those from the other two methods. Liang and Williams suggest that the speed of the measurements may sharpen the MTF by allowing the subject to hold a more constant accommodative state over the measurement interval.

From a broader perspective, given the wide variation of measurement techniques, the most striking thing about these three sets of measurements is the agreement among them. According to Liang and Williams, the difference in MTFs between the interferometric and adaptive optics measures would be brought about by a factor as small as a change of accommodation of only about 0.15 diopters. And all three techniques agree that the percentage of stimulus modulation transferred through the optics to the retina is about 50% at 10 cy/deg, and about 15% at 50 cy/deg.

In sum, one of the fundamental questions we posed at the beginning of this chapter was, how good are the optics of the eye? Thanks to the work described above, the question of optical quality is now largely a solved problem in vision science. The answer is that the eye's optics are excellent for low spatial frequencies, transferring virtually 100% of the contrast from the physical stimulus to the retinal image. But for spatial frequencies above about 5 cy/deg, the optics degrade the contrast in the retinal image. By 60 cy/deg, most of the contrast in the physical stimulus is lost in the optics, and does not make it to the retinal image. In short, the optics are a low-pass filter with a high frequency cut-off similar to the spatial resolution measured by grating acuity.

## 5.6 Photoreceptor spacing

We now turn to the second possible cause of the 60 cy/deg limit on grating resolution mentioned in Chapter 1: the anatomical layout of the photoreceptors. Beyond the optics, the next processing elements of the visual system are the photoreceptors – the entities that absorb light and start the neural signals in the visual system. As was shown schematically in Figure 1.6, the retina is paved with a mosaic of photoreceptors. Each photoreceptor absorbs only the quanta of light that arrive at that photoreceptor's location. Moreover, a photoreceptor sums the signals arising from all of the quanta it absorbs, without regard for the spatial location of each quantum within it. Thus, although the optical image is continuous in space, the photoreceptors sample the optical image *discretely*: that is, they sum the signal over small, separate local regions.

The photoreceptor layer is often referred to as the *receptor mosaic*. This terminology makes an analogy between the discrete sampling implemented by photoreceptors and the representation of a visual scene in a mosaic. In a mosaic, each local region of a scene is represented by a single tile of a homogeneous color. Just as the mosaic distorts the scene by representing each local region by the average color of that region, so too the photoreceptor mosaic distorts the incoming visual signal by summing the effects of quantal catches within local regions. Thus, this stage of discrete spatial sampling changes the visual signal.

### 5.6.1 The Nyquist limit

The effects of discrete sampling by a regularly spaced array of photoreceptors are shown schematically in Figure 5.10. This figure consists of five panels. In the top row of each panel is shown the luminance profile in the retinal image. In the second row is a set of schematic photoreceptors with a fixed inter-receptor spacing *d*. And in the bottom row are the signals produced by each of the corresponding photoreceptors across the matrix.

As shown in Figure 5.10A, a homogeneous field of light produces an equal quantum catch in all of the photoreceptors (give or take a little noise). In Figure 5.10B, a coarse grating falls on the photoreceptor array. Several photoreceptors fall under each dark stripe, and several under each

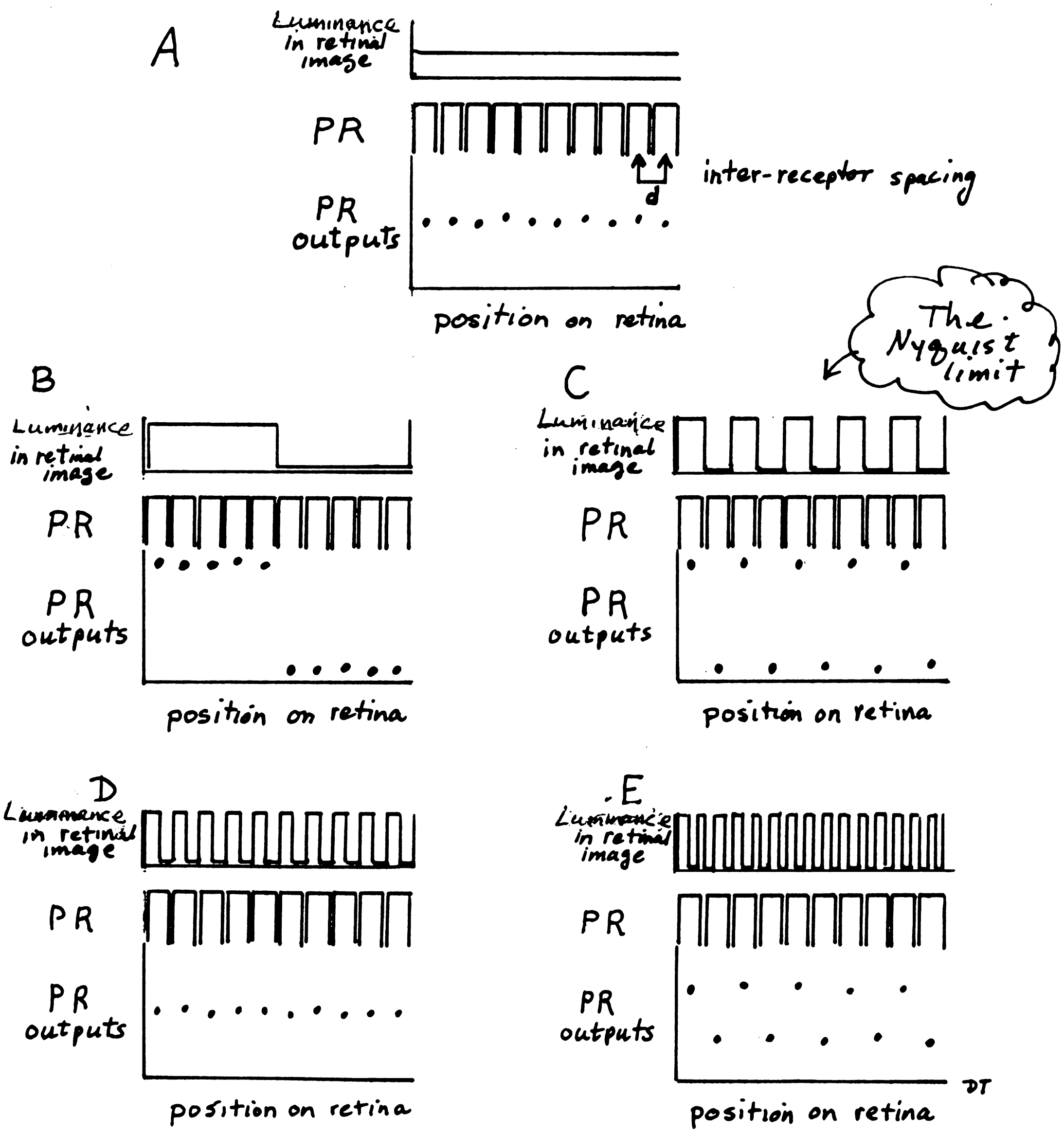


Figure 5.10: Discrete sampling and the Nyquist limit. A. A homogeneous field of light produces an approximately constant number of quantal catches in each photoreceptor (PR) across the matrix. B. A coarse grating creates a coarsely varying pattern of quantal catches, and is readily distinguished from the homogeneous field. C. A grating at the Nyquist limit creates alternating high and low quantal catches in neighboring photoreceptors. D. A grating at twice the Nyquist limit creates nearly equal quantal catches in each photoreceptor, and aliases to the homogeneous field (A). E. A grating at three times the Nyquist limit aliases to the pattern at the Nyquist limit (C).

bright stripe. That is, there will be several neighboring photoreceptors with high quantum catches, and then several with low quantum catches. Think of yourself looking out at the world through this array of photoreceptors. Analysis of the regions of high and low quantum catches would allow you to deduce that there is a spatial pattern in the world, and to have veridical (accurate) information about its spatial frequency.

In Figure 5.10C the grating is matched to the spatial separations of the photoreceptors, so that the grating produces a dark stripe on one photoreceptor and a bright stripe on its neighbor. The pattern of photoreceptor outputs would be finer in this case – one photoreceptor per stripe of the grating – but again the pattern carries veridical information about both the presence and the spatial frequency of the grating.

But there's a limit to the fineness of the grating that can be represented unambiguously by such a set of sampling units. Consider the grating in Figure 5.10D. In this case the spatial frequency of the grating is high enough so that one period of the grating – one dark and one bright stripe – falls on each photoreceptor. The pattern of quantum catches across the set of photoreceptors will wash out, and be quite similar to the pattern made by the homogeneous field in Figure 5.10A. Based on this argument, you would no longer be able to tell the grating from the homogeneous field, and information about the spatial pattern would be lost.

Intuitively, to preserve the spatial variations in the grating, one needs to sample both the bright and the dark stripe of each period of the grating, and have not more than one stripe (1/2 cycle of the grating) per photoreceptor. That is, a single cycle of the grating must occupy at least twice the inter-receptor spacing, or $2d$. The value $2d$ is defined as the *Nyquist limit* of the sampling array, stated in terms of the period of the grating. Since the spatial frequency of the grating is the reciprocal of the period, in terms of spatial frequency, $F$, the Nyquist limit is $F = 1/2d$. The grating in panel C is exactly at the Nyquist limit[6] of the receptor array.

### 5.6.2 Alias patterns in the primate fovea

But there is one more level of complication, because the Nyquist limit is not an absolute limit. Under the right conditions, some information about the presence of spatial frequencies above the Nyquist limit does get through a discrete sampling mosaic. The phenomenon is called *aliasing* because the pattern of photoreceptor responses across the sampling matrix, made by each supra-Nyquist frequency, closely resembles the pattern made by a sub-Nyquist frequency. Just as you know William Bonney by the alias "Billy the Kid", each frequency above the Nyquist limit potentially creates the same spatial pattern as a frequency below the Nyquist limit, and slips through the photoreceptor mosaic under an assumed identity – an alias.

Consider Figure 5.10 again. Figure 5.10C depicts a grating with a frequency at the Nyquist limit (call it grating $N$), and Figure 5.10E shows a frequency three times the Nyquist limit (call it $3N$). For grating $3N$, two bright stripes and one dark stripe fall on the first photoreceptor, two dark stripes and one bright stripe fall on the second photoreceptor, and so on. Even though this spatial frequency is above the Nyquist limit, the pattern of quantum catches will vary across the row of photoreceptors, and the pattern of activity across the row of photoreceptors differs from that created by the homogeneous field.

In fact, for our assumed perfectly regular matrix, the grating $3N$ would give you the *same* pattern of quantum catches as does the grating N, although at a lower contrast. If we lower the

[6] When we wrote Nyquist limit, the spell checker suggested that we might mean the nicest limit.

contrast in grating $N$, then gratings $N$ and $3N$ will yield the same spatial pattern; that is, they will alias to each other[7]. The same argument will hold true for many pairs of spatial frequencies. For each sub-Nyquist frequency, there will be a supra-Nyquist frequency that will alias to it. [Work out some examples of such *alias pairs* using diagrams like those in Figure 5.10. What pattern do you find?]

Now, how do these considerations apply to human vision? In the human fovea the interreceptor spacing, $d$, is about 0.5 minutes of arc, so (ignoring some complications) the Nyquist limit for the fovea should be about 60 cy/deg. If we could bypass the optics of the human eye, and create gratings with frequencies above about 60 cy/deg directly on our retinas, we should see alias patterns!

Simulations of the alias patterns predicted for the fovea of a primate retina are shown in Figure 5.11. Figure 5.11A shows a map of the locations of individual photoreceptors in the foveal region of a macaque monkey retina. The map shows irregular patches of regular arrays of photoreceptors. Figure 5.11B-D show patterns produced by physically laying horizontal square wave gratings of particular spatial frequencies on the matrix. The simulated spatial frequencies are 40, 80, and 110 cy/deg in panels B, C, and D respectively. In B, 40 cy/deg is below the Nyquist limit of the monkey's fovea, and the grating is represented veridically as a set of horizontal rows of dots. In C and D, 80 and 110 cy/deg are both above the Nyquist limit, and alias patterns appear. They are irregular because of the patchiness of the photoreceptor matrix.

### 5.6.3 Can we see our own alias patterns?

Conclusive experiments demonstrating the detection of alias patterns by human subjects were carried out by David Williams in 1985. Williams used interferometry to create supra-Nyquist horizontal gratings of up to 200 cy/deg directly on the retina. He then measured contrast sensitivity functions with a two-interval forced-choice experiment. Each trial of the experiment consisted of two time intervals, one of which contained a grating and the subject's task was to judge which time interval contained the grating. (Notice that this is an externally referred, forced-choice, detection experiment). The spatial frequency and contrast of the gratings varied from trial to trial.

The resulting interferometric CSFs are shown in Figure 5.12A. As expected, the subjects detected gratings below the Nyquist limit, as shown by the left hand lobe of the CSF below 60 cy/deg. But they also detected gratings far above the Nyquist limit, as shown by the right hand lobe and the tail above 60 cy/deg. With the forced-choice technique, gratings were detected all the way out to 200 cy/deg, the maximum spatial frequency that could be produced by the interferometer.

Of course a forced-choice experiment such as this one only tells us that the gratings were detected, but not what they looked like (remember, for identity experiments, don't ask; don't tell). To address this question, the subjects also described and drew the patterns they saw in the interferometer (an appearance experiment). For horizontal fringes below about 60 cy/deg, subjects reported seeing regular horizontal gratings, with spatial frequencies corresponding to the actual spatial frequencies of the interference fringes. But above about 60 cy/deg, they saw coarse, wiggly

[7] Alias patterns often occur on video systems. They are produced by aliasing between the spatial patterns in the scene and the spatial sampling characteristics of the video. Think through Figure 5.10 but with a video sensor in the place of the photoreceptors. If the input pattern is finer than the sensor spacing, the result will be aliasing. The most exotic example of aliasing Teller has ever seen occurs in the Apache dance in the movie Can-Can, viewed on a video system. The female dancer is wearing finely striped tights, and her alias patterns flash spectacularly every time she moves or changes her distance from the camera.

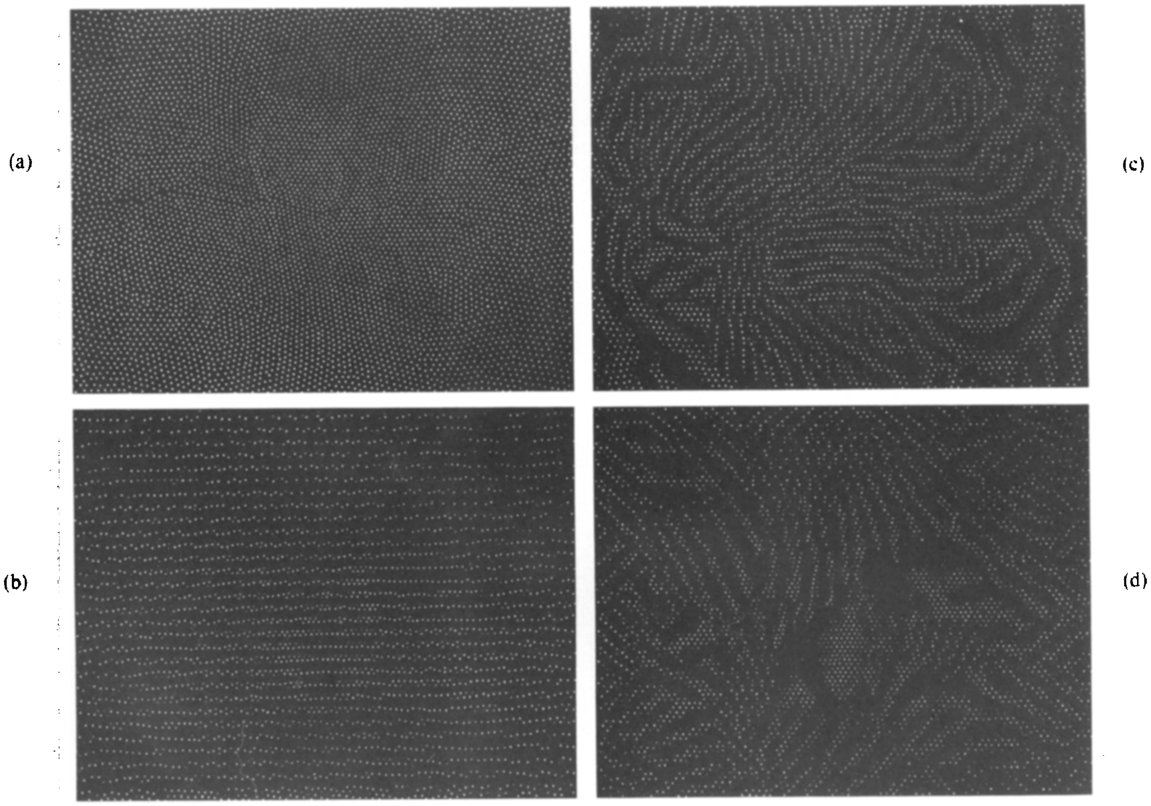


Figure 5.11: Simulated alias patterns for a monkey fovea. A. The monkey's retinal mosaic at the fovea. Each white dot represents the location of a photoreceptor. Notice the hexagonal packing in local regions, and the irregularities among local regions. B. Simulation of the pattern arising from a horizontal 40 cy/deg square wave grating. The dots are the same as in Panel (a) but whither their intensity modulated by the grating. 40 cy/deg is below the Nyquist frequency of the matrix, and the grating is represented veridically, as horizontal rows of dots. C. An alias pattern arising from simulation of an 80 cy/deg grating. D. An alias pattern arising from simulation of a 110 cy/deg grating. [Williams (1985, Fig. 6, p. 202).]

"zebra stripes" or "worms", resembling the drawings shown in Figure 5.12B-C (note the similarity to the alias patterns shown in Figure 5.11C-D).

### 5.6.4 Neural CSFs

Because of the differences in the perceived qualities of the test stimuli – regular gratings versus alias patterns – the different parts of the CSF shown in Figure 5.12A are attributed to two different origins. The region below 60 cy/deg is attributed to the processes that underlie the detection of ordinary gratings, whereas the region above 60 cy/deg are attributed to the processes that underlie the detection of alias patterns. As noted previously, the region below 60 cy/deg is called the *neural CSF*; remarkably, as discussed above, it reveals the CSF for the neural visual system in isolation, unaffected by optical factors.

Sheng He and Donald MacLeod (1996) have extended the analysis of the neural CSF. As shown

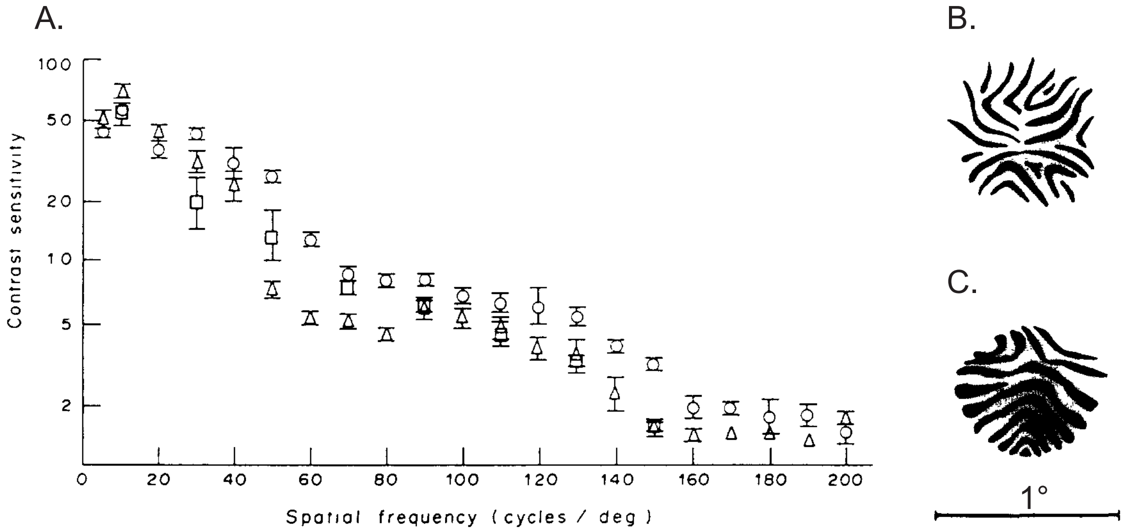


Figure 5.12: Detection of alias patterns centered on the fovea. A. An interferometric CSF based on forced-choice detection thresholds. The different symbols represent data from three observers. Gratings are detectable out to a spatial frequency of 200 cy/deg. Subjects report that gratings of spatial frequencies below 60 cy/deg are perceived veridically, whereas spatial frequencies above 60 cy/deg are visible as alias patterns. B-C. Drawings of alias patterns. The scale bar in C shows 1° of visual angle. [From Williams (1985, A from Fig. 3, p. 199; B, C from Fig. 4, p. 200).]

in Figures 5.11C-D, and 5.12B-C, the stripes in the alias patterns formed by our retinal mosaics are wiggly and variable in orientation, and in general do not conform to the orientation of the fringes that produce them. He and MacLeod proposed that subjects be tested with a forced-choice *orientation* discrimination. They reasoned that subjects should be able to make orientation discriminations for spatial frequencies that are detected veridically, but not for gratings detected only by their alias patterns.

He and MacLeod's results are shown in Figure 5.13A. Unexpectedly, their subjects could often do the orientation task above the Nyquist limit for large differences in orientation – vertical versus horizontal, or ± 45 degrees – presumably because the two gratings produced two discriminably different alias patterns. But for smaller orientation differences – gratings oriented at ± 5 or ± 10 degrees from horizontal – the subjects failed above about 60 cy/deg, and no orientation thresholds were measurable, even at 100% contrast. Thus, the lower lobes of the interferometric CSFs for orientation discrimination in Figure 5.13 provide additional estimates of neural CSFs for the human visual system.

Although this chapter concerns optical rather than neural factors, notice that we seem to have received a bonus for the future – the neural CSF, which is a description of the CSF for the neural visual system, unaffected by the optics of the eye. Notice that in both Figure 5.12 and Figure 5.13, the neural CSF declines relatively slowly below about 40 cy/deg, but falls precipitously at high spatial frequencies – more than an order of magnitude in the spatial frequency range between 40 and 60 cy/deg. We return to the implications of this finding below, when we return to the limits on grating acuity.

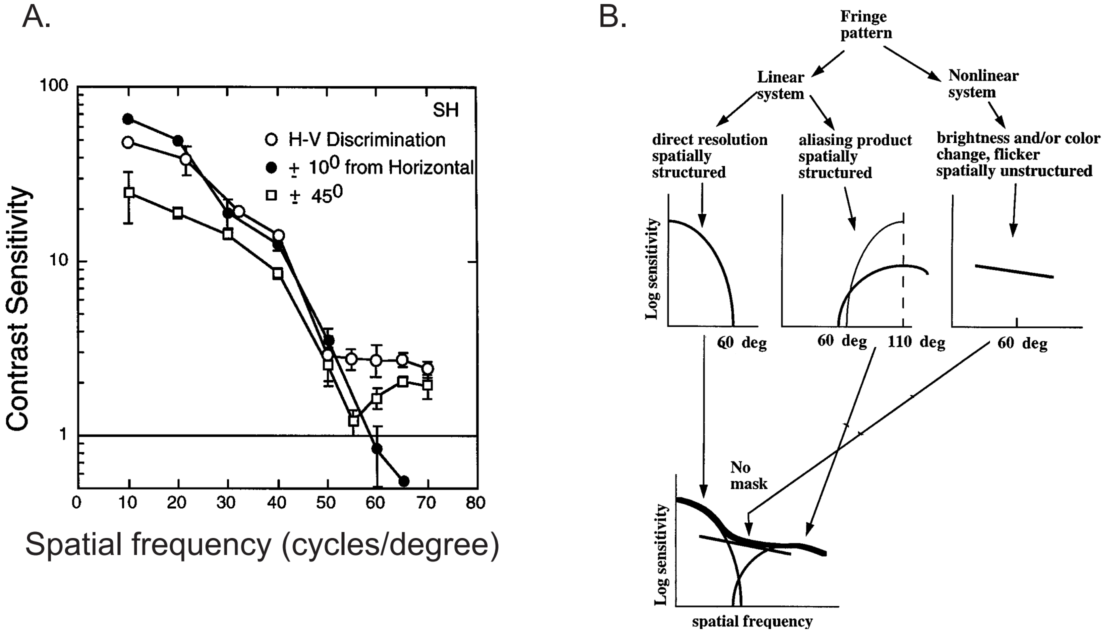


Figure 5.13: The neural CSF and a three-factor model. A. Using interferometry, subjects were asked to discriminate between gratings of two different orientations. For small orientation differences (closed symbols), discrimination became impossible above about 60 cy/deg, supporting the argument that detection above this value is mediated nonveridically by the detection of alias patterns. The lower lobe of the data, which arises from veridical grating perception, provides a new estimate of the neural CSF. B. The theory predicts functions with three lobes shown at the bottom of the figure. The lower lobe, below 60 cy/deg, is attributed to ordinary (veridical) grating detection; the upper lobe, above 60 cy/deg, to (non-veridical) detection of alias patterns; and the filling in of the gap between the two to a retinal non-linearity (see Chapter 6). [From He and MacLeod (1996, Fig. 6, p. 1143 and Fig. 8, p. 1145)]

In the meantime, He and MacLeod (1996) complicate the interpretation of the interferometric CSF by proposing a three-factor model, shown in Figure 5.13B. They argue that ordinary grating detection accounts for the left-hand lobe of the interferometric CSF – the neural CSF – below about 60 cy/deg shown on the left side of Panel B. The alias patterns account for the second lobe that rises to the right of the Nyquist limit, above about 60 cy/deg shown in the center of Panel B. However, if these were the only two factors, there should be a sharp minimum in the CSF at about 60 cy/deg, where veridical perception has cut off and aliasing is not yet available. In fact, no such minimum appears in the data, particularly in Figure 5.12 and for one subject in Figure 5.13A.

What is going on? He and MacLeod argue that detection in the spatial frequency range near 60 cy/deg is caused by yet a third process – arising from a non-linearity beyond the optics of the eye and within the neural retina. A compressive or saturating nonlinearity can affect the mean appearance of a grating to give a subtle cue of its presence. Think about how a nonlinearity affects the the minimum, maximum and average luminances. Because of the nonlinearity, the signal strength for the maximum is closer to the average signal strength than is the signal strength for the minimum. As a result, the space average signal from the grating will be smaller than a signal from the uniform field. This would not occur if the system was linear. In sum, He and MacLeod argue that a compressive nonlinearity in the photoreceptors can produce a spatially uniform change in the magnitude of the neural signal arising from the stimulus field. This change could underlie the detection of non-resolvable gratings in the range of spatial frequencies around 60 cy/deg. This argument provides us with a theoretical account of the heavy solid line that bridges the Nyquist frequency in Figure 5.13B.

## 5.7 Hyperacuities and spatial localization

In this chapter we have considered the physiological basis of spatial resolution in some detail. But, before moving on, it is worth mentioning that not all spatial judgments are limited by spatial resolution in the same way as measured by grating acuity. There are tasks that reveal finer sensitives that are sometimes referred to as hyperacuities (Westheimer, 1979).

For example, Figure 5.14A shows a target for the task of *vernier acuity*: Two vertical line with a small horizontal offset. Human subjects can see the offset and discriminate its direction when the offset is only a few seconds of arc – much smaller than the diameter of a photoreceptor. Similarly, suppose that a subject is looking at a spot of light, and at a certain point in time the spot jumps either rightward or leftward. How far does it have to jump, in order for the subject to be able to tell the direction of the jump? The *displacement threshold*, like vernier acuity, is only a few seconds of arc.

How is it possible for the visual system to have such refined sensitivity to spatial location? Figure 5.14B provides an insight. In this figure, two vertical lines are physically separated by 12 seconds of arc. Because of the line spread function of the optics, the retinal image of each line is a relatively broad distribution of light, and information about the relative locations of the lines is contained in the spatial pattern of quantal catches across each retinal image. By catching enough quanta and doing a statistical analysis of the two patterns, the visual system can estimate the relative locations of the two lines with accuracy much better than the spacing of the receptors. This highlights an important point. Limits from optics and sampling are all about spatial resolution (e.g. grating acuity). Now we are talking about spatial localization. Localization is limited in different ways than resolution.

Aside from their counterintuitive nature, hyperacuities illustrate the importance of examining closely the visual task being studied. Spatial resolution as measured by grating acuity is relevant to a great variety of tasks in which one has to resolve a fine pattern. But other tasks depend on the spatial location rather than the spatial pattern (e.g. vernier acuity). For such tasks, spatial localization is often much finer than spatial resolution of patterns. Explaining that phenomena is a story for another day.

## 5.8 Reprise: What limits grating acuity?

So, finally, how shall we answer our initial apparently simple locus question: what limits grating acuity? We now have quantitative descriptions of three possible limiting stages – optics, photoreceptor spacing, and (collapsing retinal and central stages) neural processing. The limits imposed by all three loci – the high frequency cut-off of the optical MTF, the Nyquist limit of the pho-

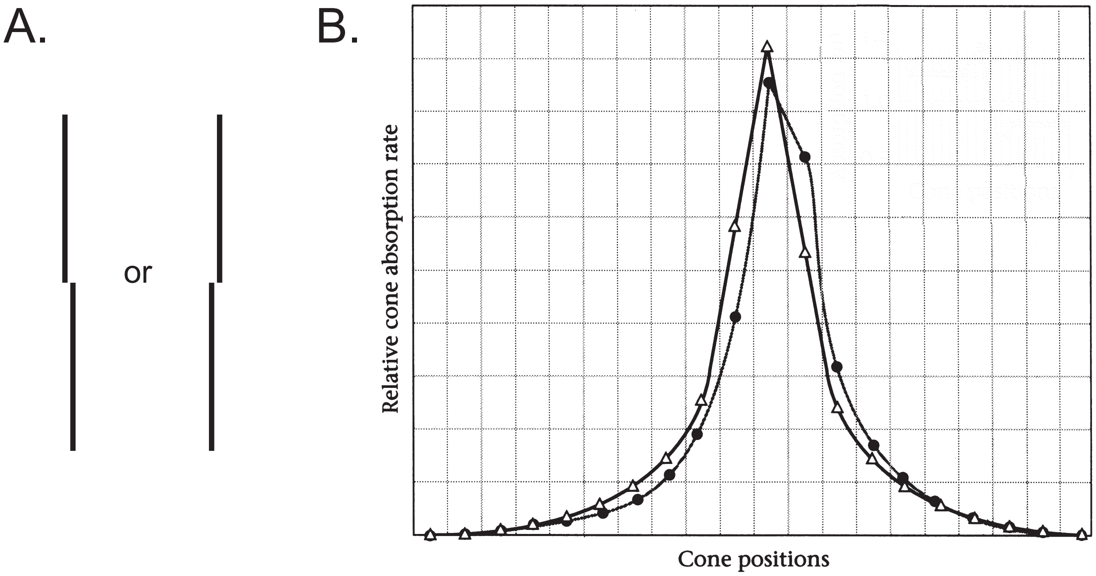


Figure 5.14: Hyperacuity. Panel A illustrates the stimulus which are two lines at horizontal locations separated by 12 seconds of arc – about 1/3 the width of a foveal photoreceptor. Panel B shows, the line spread function made by the two lines with the centers of the two functions separated by 12 seconds of arc. The abscissa denotes a row of photoreceptors (one at each grid line). The relative locations of the two lines can be determined by statistical analysis of the distribution of quantal catches. [From Wandell (1995, Fig. 7.27, p. 242).]

toreceptor matrix, and the high frequency cut-off of the neural CSF – all converge at the original grating acuity limit of about 60 cy/deg. What conclusion[8] shall we draw?

The simplest option, of course, is to stay with our original answer – the optics of the eye. Since the optics are the first stage of processing, the incoming light encounters the optics before it encounters the photoreceptors or later levels of neural processing. The optics remove high spatial frequencies from the retinal image, and thereby impose the initial limit on spatial resolution. They get the blame, even though if they had a higher cut-off frequency, the receptor mosaic and the neural CSF would still impose approximately the same resolution limit.

But, taking a slightly more sophisticated perspective, why are these three limits so closely matched? Of course, the matching of components in a serial processing system can be considered a sensible design feature, implemented by natural selection. There is no point in having the sampling limit or the neural limit better than the optical limit, or vice versa.

But there's still a deeper argument. Remember that if the optics weren't cutting out the spatial frequencies above the Nyquist limit, aliasing would occur. Aliasing can yield false perceptions, which could in principle be disadvantageous – we don't want to be seeing zebras when we look at picket fences or venetian blinds! Assuming there is no alternative to discrete sampling, how might one build a system that would still avoid the false perceptions that would arise from alias patterns? One solution would be to make the sampling matrix irregular, and indeed there are irregularities in the

[8] Other than: Beware of simple questions!

photoreceptor mosaic in Figure 5.7. However, we know that these irregularities are not sufficient to eliminate alias patterns, because we see alias patterns with interference fringes.

For the human eye, the solution to the alias problem appears to be a limit the spatial resolution of the optics to a point below the Nyquist limit of the receptors. An optical system that greatly reduces the contrast of gratings above 60 cy/deg would obviate the problem of alias patterns. On this argument, the optics of the eye may have evolved to be as poor as they are, in order to function as an *anti-aliasing device* for the photoreceptor mosaic. The important design principle here is, processing stages are not just selected to have similar limits, but also to cover for each others' deficiencies.

We can further speculate that the ultimate limit of fineness of the photoreceptor mosaic is very likely set by the minimum possible size of photoreceptors. Perhaps the organelles inside the photoreceptors can only be made so small, and no smaller. Perhaps human foveal photoreceptors are as small as they can be without a radically new design. Perhaps some limit on photoreceptor design sets the Nyquist limit at 60 cy/deg, and the Nyquist limit necessitates the optical limit at 60 cy/deg in order to protect the system from alias patterns.

If this argument were true, then what limits grating acuity to 60 cy/deg? At the functional level we might choose to blame the optics. But at the system level we might choose to blame the size and spacing of photoreceptors, which require the optics to be poor if aliasing is to be avoided[9].

The neural CSF also poses a puzzle. In evolutionary terms, it makes no sense for the neural CSF to process spatial frequencies beyond the cutoff frequency of the optics, so it is not surprising that it too cuts off near 60 cy/deg. However, why does the neural CSF fall off so rapidly at spatial frequencies between 40 and 60 cy/deg? In principle, in order to preserve all of the information transferred by the optics, the neural CSF should be low pass, with close to 100% modulation transferred right out to 60 cy/deg.

Why doesn't the neural CSF take this shape? One would have to argue that preserving the available contrast information above 40 cy/deg is too expensive to be worthwhile in terms of neural processing. But why discard information needed for spatial pattern, but preserve information about spatial location? Perhaps refined hyperacuities are more important than high contrast sensitivity. A quantitative design argument is needed if these issues are to be taken further.

Stepping back from the specifics, this discussion provides our first detailed example of linking theories. We started out with several possible physiological explanations for the limits on grating acuity. The results show that all three linking theories have some merit and suggested more elaborate linking theories that relate the various aspects of the physiology. As we proceed, we will see many examples of alternative linking theories that connect different aspects of perception and physiology.

## 5.9 Summary: Recoding the visual signal

In Chapters 4 and 5 we have talked about the first stage of information processing within the visual system – the optics of the eye. In this book we will call the mapping of the physical world to the retinal image, via the optics of the eye, the *optical transformation.*

[9] This conclusion is controversial. He and MacLeod argue that the optics are even poorer than would be needed to avoid aliasing. If so, there could be some other factor that requires a limited optical quality. Perhaps this factor is the broad line spread function needed to support the hyperacuities for spatial location.

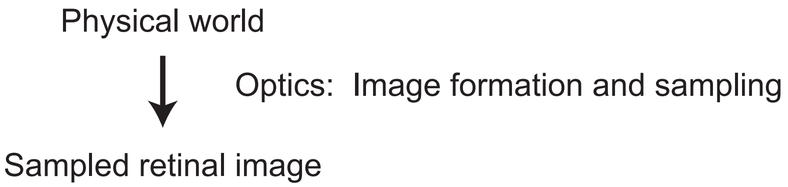


Figure 5.15: The optical transformation of the visual signal.

The optics of image formation imposes several information losses on the incoming visual signal. First, the physical world and the objects within it are three dimensional, but the retinal image is only two dimensional. The optics of the eye collapse all points in the same line of sight to the same location in the retinal image. Second, the size of the retinal image of an object is not tied to the size of the object, but varies with its distance from the observer. At this stage it is hard to see how we will ever be able to perceive the sizes and distances of objects veridically. Yet, since we can, the information must be available somehow. We return to this question in Chapter 24. Third, the physical world contains all spatial frequencies. But the optical system low-pass filters the incoming visual signal, reducing the contrasts of spatial frequencies above about 5 cy/deg, and imposing a cut-off at about 60 cy/deg. We see the world low-pass filtered, through the "window" of our optical MTF.

As we move from one stage of visual processing to the next in this book, we will try to capture the essential form of the code at each level with a slogan, "in 25 words or less". In Figure 5.15, this is further reduced to a single sound bite. These slogans are intended to serve as mnemonics for remembering the effects of each stage of processing on the visual code. We are now ready to summarize the effects of optics. The optics transform the incoming signal by making it *two-dimensional*, and *low-pass filtered*. This signal is further transformed by the process of discrete sampling. Here sampling was introduced in the abstract; in later chapters the details of the photoreceptors will make the process more concrete.

The artistic style called *pointillism*, shown in Figure 5.16, provides a nice analogy for the form in which spatial information is coded by the photoreceptors. In pointillism, the artist represents a visual scene by using discrete dots of paint on the canvas. At the level of the photoreceptors, the visual scene is represented by sets of spatially discrete quantum catches. There is no signal that binds together the parts of a given object in the scene. (Of course there was no such signal in the retinal image either.) The work of assembling the image into meaningful parts remains to be done by higher levels of the visual system. In sum, discrete sampling makes the incoming signal *pointillistic*. We pursue this in Chapter 6 with a more detailed analysis of the photoreceptors. How do they work, and what limits do they place on our vision?

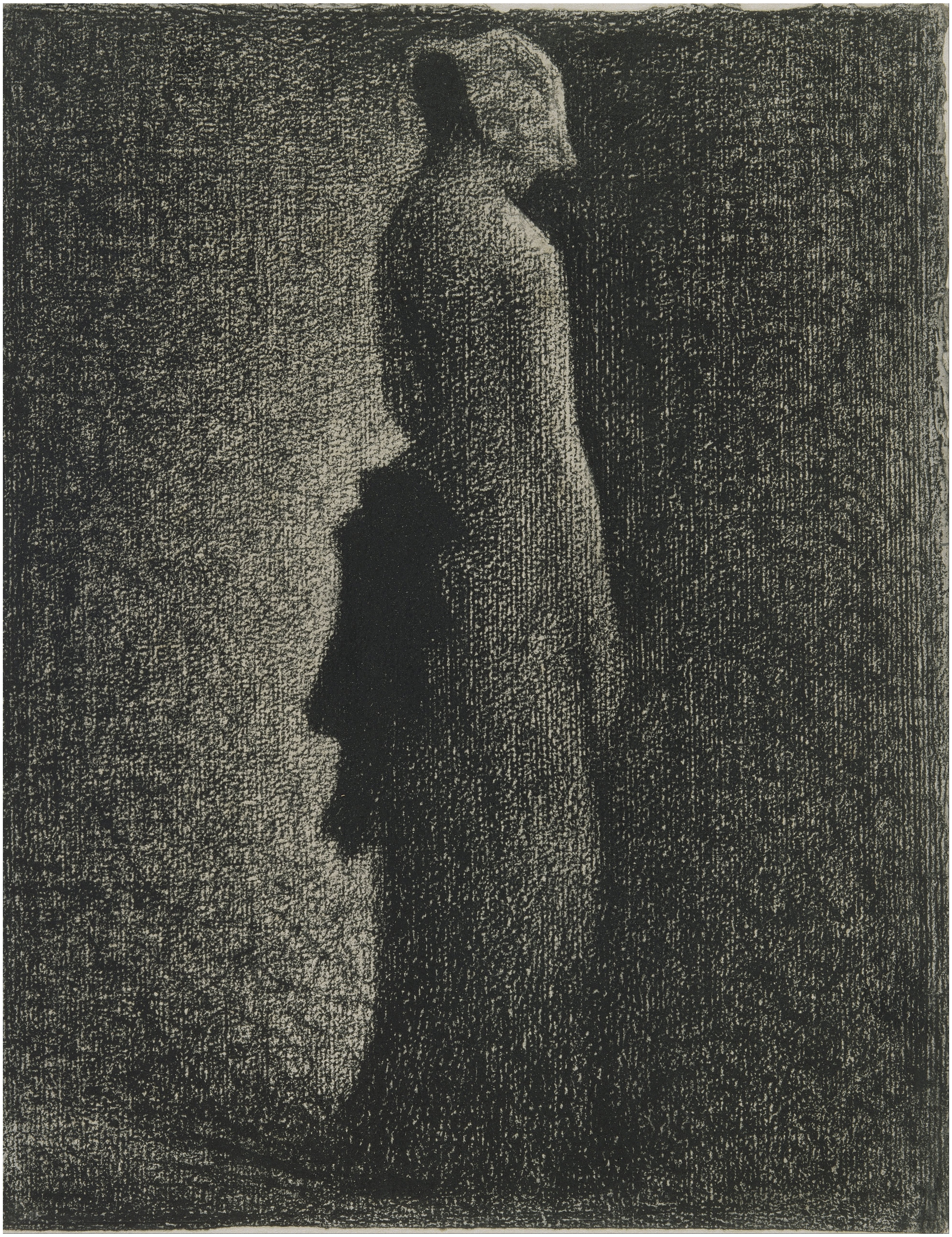


Figure 5.16: A pointillistic painting: Seurat, The Black Bow. [See copyright information on the painting; via Ratliff (1965, Fig. 7.27, p. 242).]

# Chapter 6

# Photoreceptors and Vision

## Contents

In this chapter we leave behind the purely physical aspects of vision – light and optics – and begin a section on retinal physiology. In Chapters 6 and 7 we explore the anatomy, physiology, and function of photoreceptors. In Chapters 8-12 we explore the properties of other retinal neurons. In each case, we first describe the anatomy and physiology of the neurons themselves, and then introduce some of the linking theories of how the characteristics of these neurons leave their marks on the system properties of vision.

Logically speaking, photoreceptors have two major tasks to perform: *phototransduction* and *signal transmission*. The first goal of this chapter is to create at least a qualitative (if not quantitative) appreciation of these two processes.

In the *phototransduction* process, each individual photoreceptor – rod or cone – absorbs quanta of light. A quantum of light – a physical entity – ends its existence, and creates a neural signal. For many vision scientists, phototransduction holds a special fascination, because it forms the immediate interface between the physical and physiological worlds. A part of the universe becomes a part of the individual.

The second task of photoreceptors is *signal transmission*. The absorption of quanta occurs in the outer segment of the photoreceptor. But in order to influence vision, the photoreceptor must transmit a neural signal all the way to its synaptic terminal, at which the photoreceptor communicates with the other retinal neurons. The processes involved are complex, and the technical details require a background in biochemistry and cell biology. However, if you don't have this background, we hope to provide an intuitive appreciation.

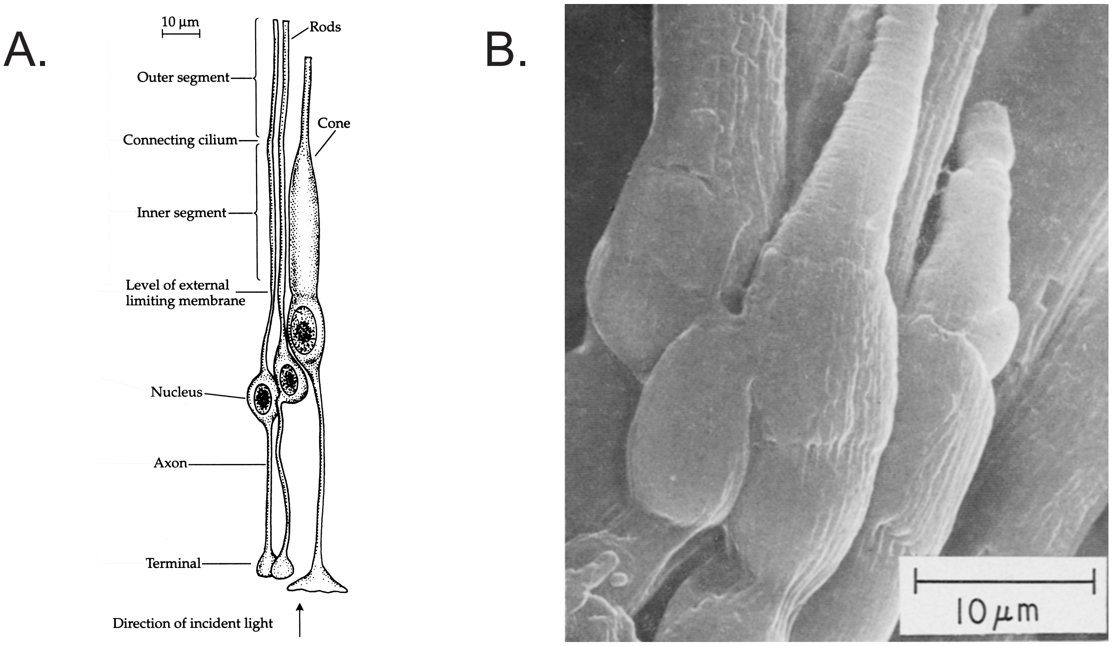


Figure 6.1: Rods and cones. A. Drawings of a rod and a cone as seen under a light microscope. It shows the three basic parts of the photoreceptor: outer segment, inner segment, and synaptic terminal. The outer segments lie against the pigment epithelium, at the back of the eyeball, and the synaptic terminals lie closest to the center of the eyeball. Light entering the eye passes through the synaptic terminals and inner segments of the photoreceptors before being absorbed in the outer segment. B. A scanning electron micrograph of two cones and several rods. [A from Oyster (1999, Fig. 13.5, p. 550), after Polyak (1941). B from Lewis, Zeevi, and Werblin (1969, Fig. 1, p. 561).]

Vision scientists can now carry out physiological recordings from single living, functioning photoreceptors. At low light levels, these recordings show that, amazingly, the absorption of a single quantum in a rod outer segment creates a physiological signal that is sufficient to affect the output of the rod. At higher light levels, they show that the responses of rods increase with increasing light levels, but eventually saturate; that is, they provide evidence for a saturating non-linearity very early in visual processing.

This chapter also continues our analysis of linking theories. How do the properties of photoreceptors leave their marks on the system properties of vision? We consider two examples. The first deals with transduction in the rods and the spectral characteristics and wavelength information losses of scotopic vision. The second deals with the the exquisite sensitivity of rods and its consequences for the absolute threshold of light detection. Transduction and signal transmission processes in the cones also have profound consequences for color vision and for photopic spectral sensitivity. However, the color story is too long to tell within the present chapter, and is postponed to Chapter 7.

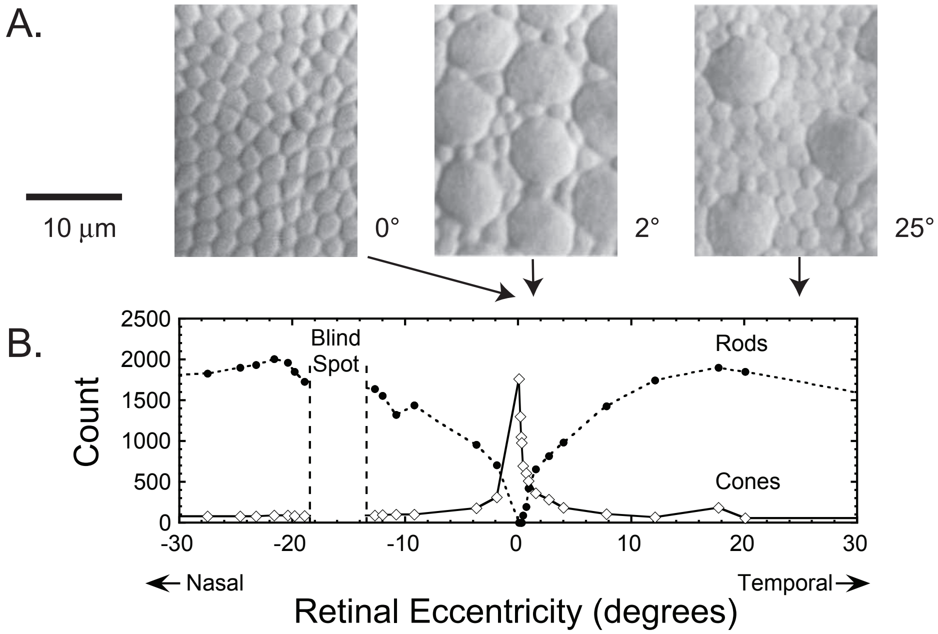


Figure 6.2: Distributions of rods and cones across the retina. A. Photomicrographs at the level of the inner segment that show the relative sizes and concentrations of rods and cones at three retinal eccentricities. At zero degrees, in the fovea, all of the photoreceptors are cones. Outside the fovea, there is a mix of cones and rods with the cones becoming larger in diameter. B. The counts of rods and cones in a 0.0069 mm$^2$ patch as a function of retinal eccentricity (think 83 by 83 $\mu$m) along the horizontal meridian. The concentration of cones is highest at the fovea, drops off rapidly out to about 10° eccentricity, and remains roughly constant across the rest of the peripheral retina. The rods are absent from the fovea, increase to a maximum concentration at about 20° eccentricity, and then taper off toward the far periphery (not shown). The blind spot – the region at which the optic nerve exits the eye – contains no photoreceptors of either type. [A. From Curcio et al. (1990, Fig. 2 and 3, p. 501-502); B. Replotted from Osterberg (1935, Table 3. p. 64-74)]

## 6.1 The anatomy of photoreceptors

As shown in the schematic overview of Figure 1.4A, the photoreceptors lie in the *outer* portion of the retina, against the back wall of the eyeball. Quanta of light coming in through the lens traverse several other types of neurons before they arrive at the photoreceptors and are finally absorbed.

### 6.1.1 Rods and cones

There are two kinds of photoreceptors in the eye – *rods* and *cones*. Figure 6.1 shows some family portraits of rods and cones. Figure 6.1A shows a drawing of a primate cone and two primate rods, as seen through a light microscope. Figure 6.1B shows a scanning electron micrograph of two cones and several rods. Anatomists divide each photoreceptor into four basic parts: the *outer segment*, the *inner segment*, the *nucleus*, and the *synaptic terminal*.

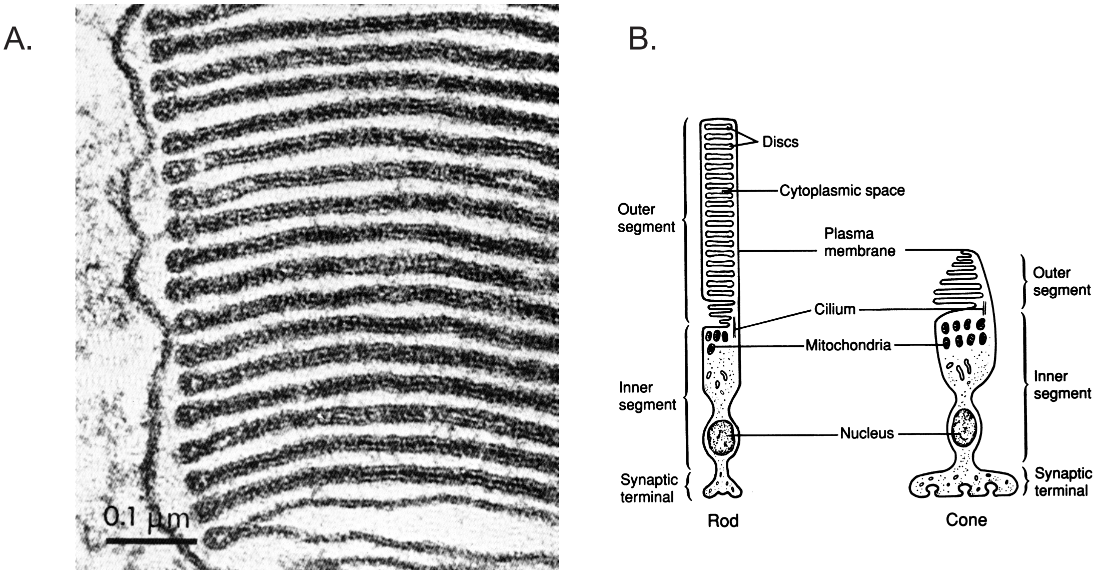


Figure 6.3: Membrane specializations in the outer segments of rods and cones. A. This panel shows a high magnification electron micrograph of part of the outer segment of a monkey rod. The tightly packed horizontal striations are the disks. B. This panel shows the different arrangements of the specialized membranes in rod and cone outer segments. In the rod, the disks are separate from the cell membrane. In the cone, the cell membrane folds back and forth, making a continuous comb-like structure. [A. from Dowling (1987, Fig. 7.4, p. 194) and B. from Kandel et al. (1991, Fig. 28-2B, p. 403).]

### 6.1.2 Retinal distributions of rods and cones

The numbers of rods and cones vary across the retina in different ways, as shown in Figure 6.2. The photomicrographs in Figure 6.2A show the varying sizes and densities of the two types of photoreceptors. In the central fovea (eccentricity 0°), all of the photoreceptors are cones, and the outer segment diameters of foveal cones subtend only about 30 seconds of arc. At the other eccentricities both rods and cones are present, with the cones being increasingly larger in size, and the rods being increasingly more numerous. Figure 6.2B shows the relative concentration of rods and cones as a function of retinal eccentricity.

Figure 6.3A shows a high power electron micrograph of a rod outer segment. It is a highly specialized structure of tightly packed membranes, called *disks*. There are about 1000 disks per rod outer segment. Each disk is composed of two membranes joined at the ends with a space between, like a stack of pita bread. The whole stack of disks is contained within the outer membrane of the rod outer segment. In cones these structures are slightly different (see Figure 6.3B), in that the “disk” membranes are actually continuous with the outer membrane of the cell. The cross section of a cone is like a comb.

These highly organized structures hold the machinery for catching quanta and starting neural signals in the photoreceptors. Because of structural differences and other factors, the details of transduction differ slightly between rods and cones. Our descriptions will apply most strictly to rods.

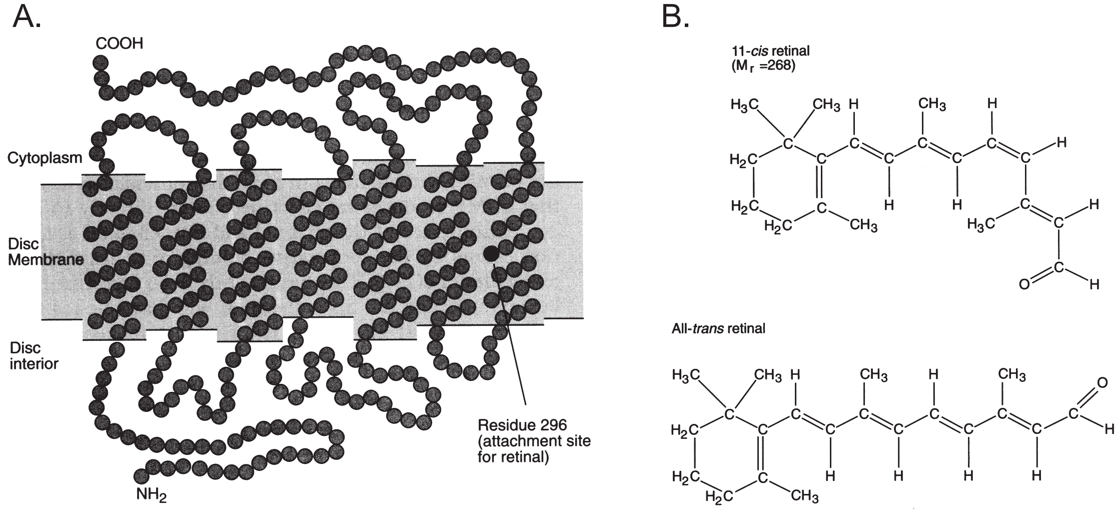


Figure 6.4: The rhodopsin molecule. A. The rod opsin twists into a helical structure, and loops back and forth through the disk membrane seven times, with one end outside the disk and the other inside it. B. The chromophore in its 11-cis and all-trans configurations. [From Kandel et al. (1991, Fig. 28-4, p. 405).]

## 6.2 Phototransduction

### 6.2.1 The rod photopigment: Rhodopsin

The substance that absorbs quanta of light in the photoreceptor outer segment is called a *photopigment.* In rods, the photopigment is called *rhodopsin*[1]. Rhodopsin molecules sit tightly packed in the disks of the outer segments of the photoreceptors, as well as in the surrounding membrane that contains them. In mammalian retina, each rod contains about 1000 disks, and each disk contains about $10^5$ rhodopsin molecules, for a total of about $10^8$ rhodopsin molecules per rod outer segment.

The structure of the rhodopsin molecule is shown in Figure 6.4. It has two parts – the *opsin molecule* and the *chromophore.* The opsin is by far the larger part – for those with a background in chemistry, it is a protein composed of 348 amino acids. The molecular weight of the opsin molecule is about 39,000.

As shown in Figure 6.4A, the opsin molecule twists into a helical structure, and loops back and forth through the disk membrane a total of seven times, crossing between the outside and the inside of the disk.

The chromophore – also called *retinal* – is by far the smaller part of the molecule with a molecular weight of only 285 compared to 39,000 for the opsin. Retinal is the form of vitamin A commonly found in carrots and other vegetables. As shown in Figure 6.4B, the chromophore commonly exists in either of two forms, called *11-cis* and *all-trans retinal.* These are terms used to

[1] In an eye that is fully adjusted to the dark, the retina has a rosy appearance – hence the name *rhodopsin* (= red vision substance) for the rod photopigment. When the eye is exposed to high levels of light, the retina changes color – it loses its color, or "bleaches". In vision jargon, light is said to *bleach* photopigments.

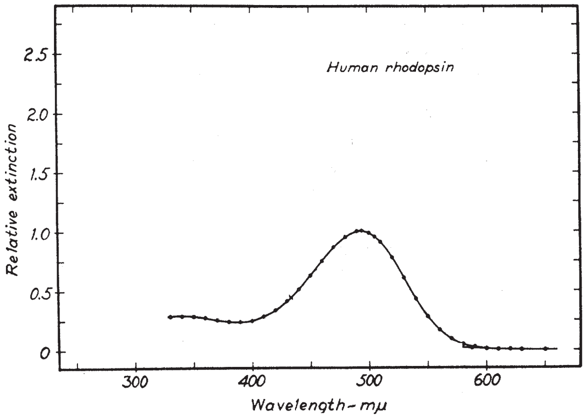


Figure 6.5: The absorption spectra of human rhodopsin. The graph shows the relative absorption (relative extinction) as a function of wavelength. The spectra for rhodopsin peaks in the visible spectra near 500 nm and falls off sharply at long wavelengths. For short wavelengths, it falls off to a broad shoulder for wavelengths from 300 to 400 nm. [Modified from Wald and Brown (1958, Fig. 1, p. 223).]

describe the atomic connections, and hence the three-dimensional shape, of the chromophore. The two states of the chromophore are called *isomers* – the chemical composition of the chromophore remains unchanged, but its three-dimensional shape changes. In the 11-cis configuration, the carbon backbone of the molecule is bent at the 11th carbon atom; in the all-trans configuration, this bend is straightened. The chromophore is attached to the opsin within the membrane, in the middle of the seventh loop (as labeled in the rightmost loop in Figure 6.4A). It lies in wait in the 11-cis configuration in the middle of the barrel, primed for action when a quantum of light arrives.

Now, recall that a quantum is an indivisible packet of energy. It cannot be divided up among different photoreceptors, but can only enter a single photoreceptor and be absorbed by a single rhodopsin molecule. Moreover, absorption is an all-or-nothing event: either a quantum is absorbed or it is not. But it is also probabilistic: when the quantum arrives at the outer segment, there is a certain probability that it will be absorbed by a molecule of rhodopsin. That probability varies with the wavelength of the light. For rhodopsin that probability is maximal at about 500 nm, and it falls off at both shorter and longer wavelengths. If the quantum is not absorbed, it is lost to visual processing.

### 6.2.2 The absorption spectra of rhodopsin

In prior chapters we have discussed the encoding of wavelength information using the idea of spectral sensitivity. Recall the effects of wavelength on absolute threshold as summarized by Figure 2.6. It showed several ways to plot the sensitivity of human scotopic vision as a function of wavelength. In studying the photopigments, we can ask the same question of the pigment. What is the sensitivity of the pigment to light?

This measurement was first made by Koenig in 1894 (see Hecht, 1937 for a review of the early literature) and has been refined over the years. A detailed and convincing set of measurements were made by Wald and Brown (1958). They measured the amount of light absorbed by a thin film of rhodopsin that had been extracted from a human eye. Care must be taken with such measurements because as soon as rhodopsin is exposed to light it bleaches into the opsin and the all-trans chromophore. Wald and Brown measured this bleaching process and the estimated underlying absorption spectra of rhodopsin is shown in Figure 6.5. It plots a measure of absorption (extinction) as a function of wavelength. Human rhodopsin has a peak just below 500 nm and sensitivity falls off sharply for longer wavelengths and falls off more modestly for shorter wavelengths. This spectra has the same general shape as the scotopic spectral sensitivity except that it shows more sensitivity at short wavelengths. Why might that be?[2]

### 6.2.3 The mechanism of transduction: Cis-trans isomerization

So far, so good. But how does light act on the rhodopsin molecule? When a quantum is absorbed into the structure of a rhodopsin molecule, the energy of the quantum is used to excite an electron, and the decay of the excited electron leads to a change in the conformation of the chromophore. In other words, the only thing light does in the entire visual process is to trigger a change in the shape of the chromophore. Moreover, the change of shape is always the same, regardless of the wavelength of light that has been absorbed.

After the chromophore absorbs the quantum, the rhodopsin molecule goes through a series of very rapid changes in three-dimensional shape, finally including separation of the chromophore from the opsin. With the help of enzymes located in the pigment epithelium, the chromophore eventually gets changed back into 11-cis retinal, and rejoins the opsin in the ultimate recycling process. On average, it takes several minutes for a rhodopsin molecule to reform.

## 6.3 Signal transmission

As we said earlier, the transduction process is the first of two tasks that each photoreceptor needs to accomplish. The second task is signal transmission: the photoreceptor must create and transmit a signal, passing the information that the quantum has been absorbed, all the way from the rhodopsin molecule to the synaptic terminal.

Photoreceptors are neurons, but it turns out that they are very atypical neurons. In particular, the more typical neurons, of which students have often heard, have axons and fire action potentials (spikes). To work through the properties of photoreceptors, it will be useful to be able to compare them to the properties of typical neurons. In this section, we first review the properties of typical neurons, and then proceed to the unusual properties of photoreceptors.

### 6.3.1 A typical neuron in a nutshell

Figure 6.6 shows the anatomy of a typical neuron. The figure illustrates its *dendrites*, *cell body*, and *axon*. The direction of signal transmission is in through the dendrites, across the cell body and out the axon. The axon ends in a set of structures called the *axon terminals*, and it contacts the dendrites of the next set of neurons across intercellular spaces called *synapses*.

[2] Hint: This measurement was taken with the pigment removed from the eye so the light did not have to pass through the eye's optics.

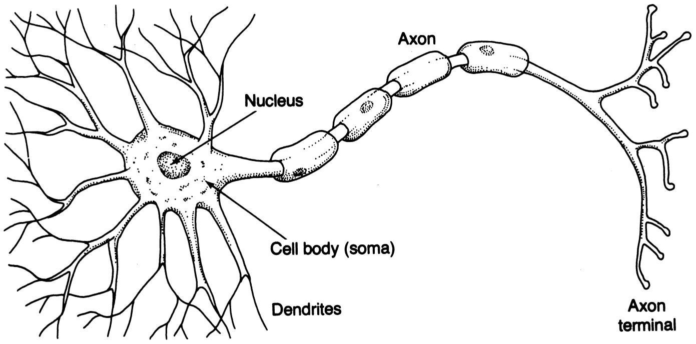


Figure 6.6: A typical neuron. A. Anatomy of a neuron: dendrites, cell body with nucleus, axon, and synaptic terminals. [From Levine and Shefner (1991, Fig. 3-11, p. 36).]

How does a typical neuron work? As shown in Figure 6.7, in its resting state a typical neuron sets up a small electrical voltage across its outer membrane. It does this by maintaining different concentrations of ions with different electrical charges inside versus outside the cell membrane. Part of the reason for the charge difference is that nerve cell membranes are *semipermeable*. That is, like the filter in a coffee maker, the openings (channels) in the membrane pass particles of a certain size and electrical charge, and exclude others. So for each kind of channel in the membrane, some chemical ions can enter and leave whereas others cannot. Moreover, the ions always diffuse from a high to a lower concentration. So if one can create an imbalance, there will be a drift of ions through semipermeable membrane to restore the balance.

To create an imbalance between the ions on either side of the cell's membrane, all neurons have a *sodium-potassium pump* – an active transport mechanism that pumps sodium ions ($Na^+$) out of the cell and pumps in potassium ions ($K^+$). Because more sodium ions are pumped out than potassium ions are pumped in, the net charge is negative on the inside with respect to the outside. For most neurons, the resting membrane potential – the magnitude of the charge – is about -70 mV, with the minus sign indicating that the *inside* of the cell is negative with respect to the *outside*.

How do typical neurons process incoming signals? When a neuron receives a signal across a synapse from a neuron earlier in the causal chain, the incoming signal perturbs the voltage across the membrane in the vicinity of the input synapse. These perturbations, called *graded potentials*, can be either *depolarizing* (excitatory) or *hyperpolarizing* (inhibitory). They spread passively along the dendrites and the cell body, with decreasing effect the greater the distance from the input site. Graded potentials from many input sites combine their positive and negative effects across the dendrites and cell body of the neuron.

At the base of the axon there is a specialized location called the *axon hillock*. When the neuron is *depolarized*, so that the voltage across the membrane of the cell is sufficiently *decreased* at the axon hillock, the properties of the axon membrane suddenly change. The mechanisms for this change are well understood, but beyond the scope of this chapter. Suffice it to say that the end product is an *action potential* or *spike* – a brief, all-or-none wave of *depolarization* that travels rapidly along the axon, propagating itself all the way to the synaptic terminal.

A reasonable analogy to a spike traveling down an axon is a flame travelling down a long match. The act of striking the match starts a flame at one end. Each segment of the match that burns

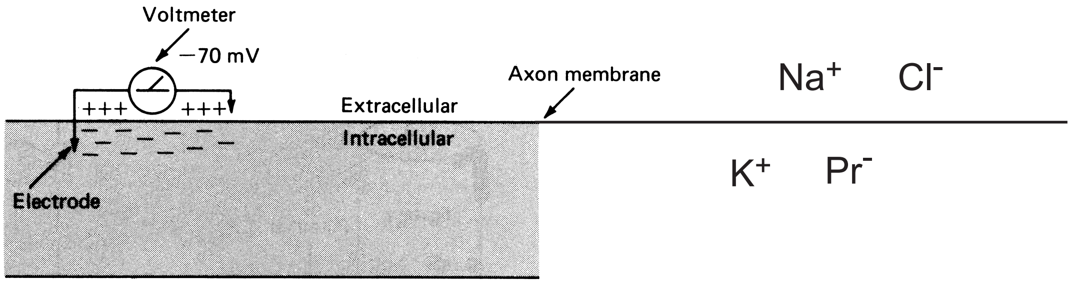


Figure 6.7: Ionic differences across a neuron's membrane. A typical neuron maintains different concentrations of different ions (charged particles) on the inside of its cell membrane compared to the outside. The inside contains relatively high concentrations of potassium (K+) and proteins (Pr–); the outside contains relatively high concentrations of sodium (Na+) and Chloride (Cl-). These charge differences create a resting potential – an electrical voltage – of -70 mV between the inside and the outside of the cell. [Modified from Levine and Shefner (1991, Figs. 3-3 and 3-4, p. 38).]

provides the energy to ignite the next segment, so that the flame travels all the way down the match to its end. The analogy would be even better if the segments of the match regenerated themselves after the flame had passed, to be ready for the next traveling flame.

At the synaptic terminal, the depolarization brought about by the arrival of the spike produces an increase in the release of a *neurotransmitter* – a chemical substance specialized to transmit signals across the synapse. The neurotransmitter in turn creates graded potentials in the dendrites and cell bodies of postsynaptic neurons. Any given postsynaptic neuron can receive and combine inputs from thousands of presynaptic cells. The process is repeated countless times throughout the complex neural network of the brain.

In sum, whenever a neural signal must travel relatively long distances, the signal is carried by the patterns of spikes in the axons of typical neurons. We will return to typical neurons in Chapter 8, when we explore the properties of ganglion cells – the neurons whose axons carry information from the eye to the brain.

### 6.3.2 Photoreceptors are not typical neurons

In the meantime, we return to photoreceptors. Vertebrate photoreceptors are extremely atypical neurons – in fact, as we will see, many of their properties differ from those of a typical neuron. Most importantly, photoreceptors do not produce spikes; they transmit messages only with graded potentials.

### 6.3.3 Photocurrents

The unique physiological properties of photoreceptors are shown in Figure 6.8. The first notable difference between a photoreceptor and a typical neuron is that the resting potential of a photoreceptor is only about -40 mV rather than the -70mV seen in the typical neuron. Why is this so?

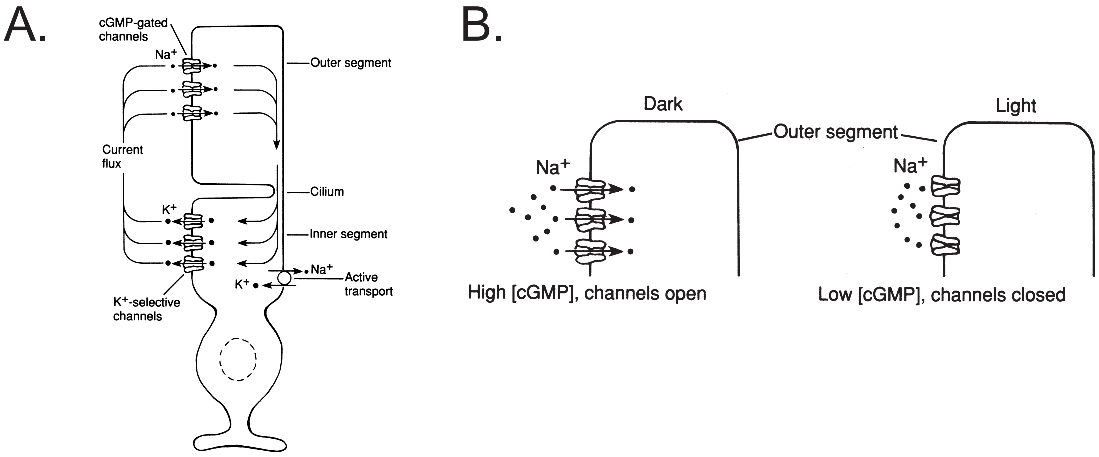


Figure 6.8: The dark current. A. The arrows show the current that flows in darkness, when the sodium permeable (cGMP gated) channels in the outer segment are open. B. In the light the channels close, and sodium ions are excluded from the outer segment. [From Kandel et al. (1995, Fig. 22-5, p. 415).]

As with typical neurons, there are sodium permeable channels in the outer membrane of the outer segment of the photoreceptor (shown in Figure 6.8B). In the dark, these channels are open, and they allow sodium to leak in from the higher concentration outside the cell. Meanwhile, the inner segment is permeable to potassium ions which leak out from the higher concentration inside the cell. This state of affairs produces an electrical current – a flow of ions – from the outer to the inner segment inside the cell, and back again on the outside of the cell (Figure 6.8A). This continuous depolarizing current, called the *dark current*, results in the membrane potential of about -40 mV compared to the -70 mV seen in a typical neuron.

In the dark, the constant depolarization produces a constant release of transmitter from the synaptic terminal. This property of photoreceptors is consistent with the corresponding property of typical neurons, in which depolarization (at the axon hillock) produces an increase in the release of transmitter (at the synaptic terminal, at the far end of the axon). In summary, in the dark, the resting membrane potential is about -40 mV, the photocurrent flows continuously, and the photoreceptor continuously releases transmitter from its synaptic terminal.

What happens when light is absorbed? It turns out that (for reasons discussed immediately below) sodium channels in the outer segment close, excluding sodium ions, and thus reducing the dark current. In consequence, the cell *hyperpolarizes* toward -70 mV; and the release of transmitter that occurred continuously in the dark is *reduced* by the action of light.

We should pause for a minute to consider this mechanism of action. Intuitively, most of us would probably have guessed that increased light on the photoreceptor would yield an increase in transmitter release from the photoreceptor. But in fact the opposite is the case – increased light absorption in the photoreceptor yields a *decreased* signal from the photoreceptor. Is this a logical problem? Not really. The fact that the signal for an increase in light level is a decrease of transmitter (and vice versa) is logically perfectly OK. It is the *change* of transmitter release with the change of light level that matters, not the absolute direction of change.

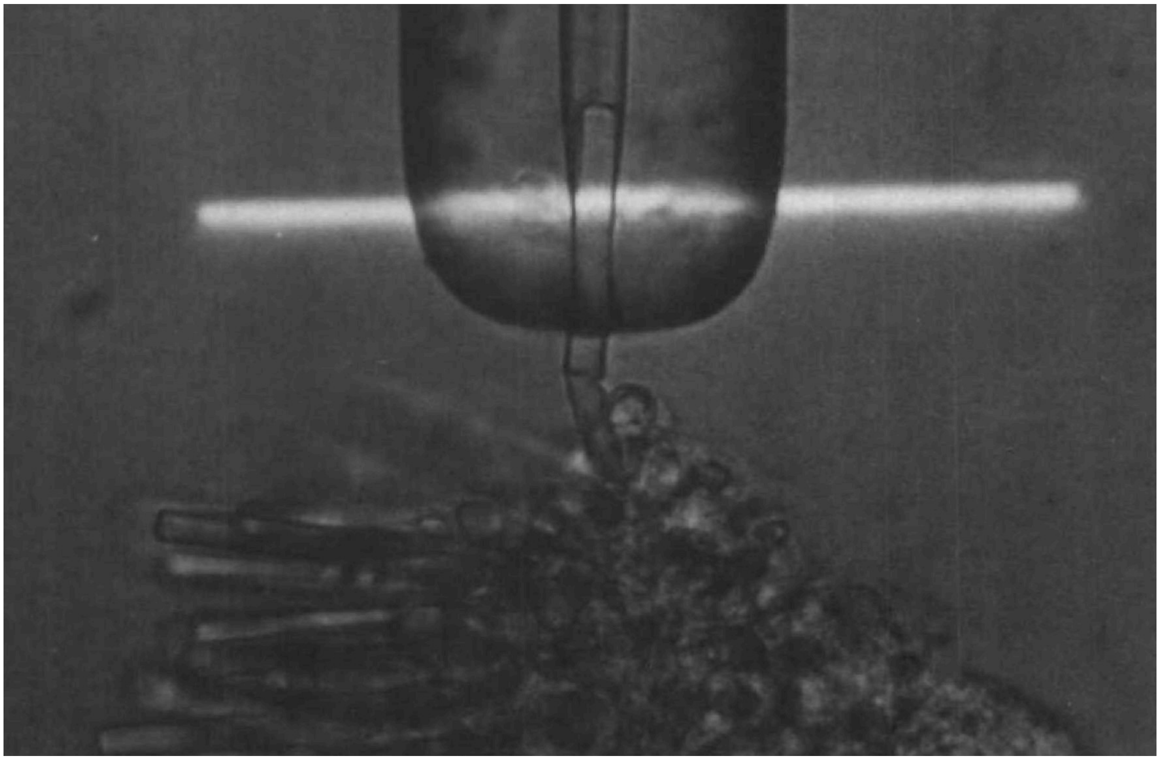


Figure 6.9: The suction electrode technique. The photomicrograph shows a piece of retina from a toad (*Bufo marinus*). The outer segments of the rods are at the left. One of the outer segments has been sucked into the electrode so that its membrane current can be recorded. The horizontal bar across the electrode is the stimulus – a bar of light used to deliver quanta to the rod. [From Baylor (1987, Fig. 4, p. 36).]

### 6.3.4 The chemical cascade

Returning now to the outer segment: what happens after a photon is absorbed? In particular, how does the message that light is absorbed get from the rhodopsin molecule to the outer membrane of the cell, and bring about a closing of the sodium channels? To make a long and technical story short, the absorption of a quantum triggers a complex series of biochemical changes, called the *chemical cascade*, that results in the closing of sodium permeable channels in the outer membrane of the outer segment of the photoreceptor.

From our perspective, the bottom line is that the chemical cascade produces an enormous *amplification* of the signal. Absorption of a single quantum in the outer segment of a rod results in the closing of several hundred sodium permeable channels, and each closed channel blocks the entry of as many as 10,000 sodium ions per second into the rod's outer segment. As we will see, the change in sodium flux is so large that it causes a detectable change in the charge across the photoreceptor membrane, as well as in the output of the photoreceptor.

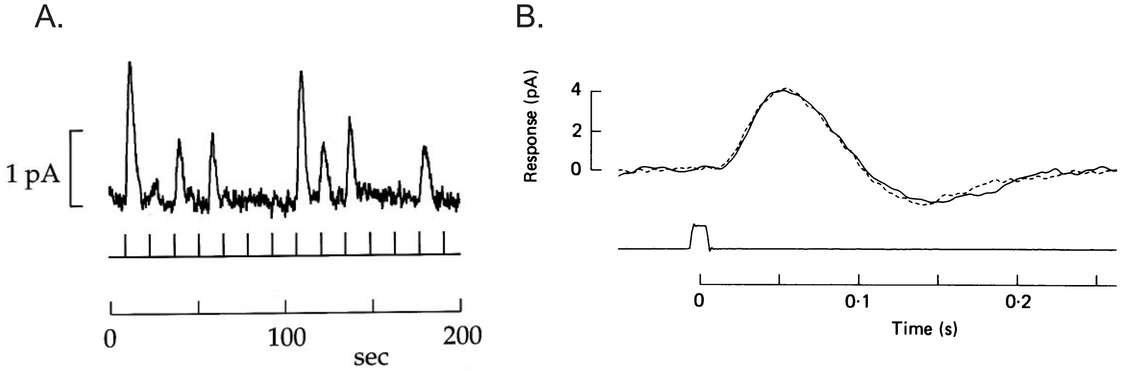


Figure 6.10: The responses of rods to very dim flashes of light. A. Responses to a series of very dim flashes spread over a 200 second period. Individual flashes occur at each tick mark. The ordinate shows the change in the photoreceptor current in picoamps (pA; 1 picoamp = $10_{-12}$ amperes). The responses of the photoreceptor come in three sizes, corresponding to the absorption of zero, one, or two quanta from each particular flash. The response to one isomerization is about one pA. B. The response to a single quanta on an expanded time scale. The responses to quanta of different wavelengths, such as the 559 and 659 nm cases shown here, are identical. [A from Rieke and Baylor (1998, Fig. 4, p. 1030); B from Baylor, Nunn, and Schnapf (1987, Fig. 2, p. 150).]

## 6.4 Physiological responses recorded from rods

In the 1970s, a remarkable new technique was developed: the *suction electrode*. A suction electrode preparation is shown in Figure 6.9. Using an excised retina, it is possible to draw the outer segment of a rod or a cone into a closely fitting hollow glass electrode. The membrane current that would ordinarily flow along the outside of the cell then flows inside the electrode, and changes in the current can be measured. By shining lights of various intensities and wavelengths on the outer segment of the photoreceptor within the electrode, one can record the changes in current flow in response to light, all the way down to the responses to absorption of an individual quantum. This work was extended to primate photoreceptors, including human photoreceptors, in the early 1980's.

### 6.4.1 Low light levels: Responses to single quanta

Suppose that a rod outer segment has been drawn into a suction electrode, and the current flowing through the photoreceptor is being recorded. Now suppose that the experimenter produces a series of flashes of light so dim that on average only a single quantum of light will be absorbed by the photoreceptor. Because of the quantal nature of light, the actual number of quanta absorbed will vary from one flash to the next – sometimes zero, sometimes one, sometimes two, and occasionally more than two. If the response of the photoreceptor to a quantal absorption is consistent and repeatable, then the change in current on each flash should take one of only a few stereotyped forms, corresponding to the absorption of zero, or one, or two, or (occasionally) higher numbers of quanta.

The results of such an experiment are shown in Figure 6.10A. The tracing shows the current recorded as a function of time. The tick marks under the tracing show the times at which the

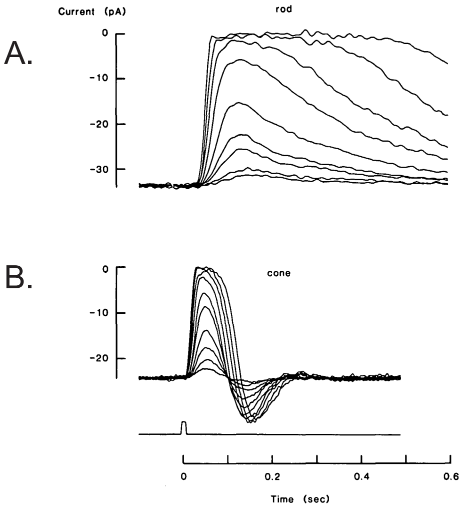


Figure 6.11: Saturation in rods and cones. A. Superimposed responses to flashes of light of increasing intensity, recorded from a monkey rod with a suction electrode. The average number of quanta absorbed per photoreceptor per flash increases by a factor of two from each trace to the next. In the top panel, for traces 1-7, the higher the intensity the higher the peak membrane current. For traces 7-9, the peak response shows little if any additional increase; these traces reveal physiological saturation in the rod. B. A similar trend can be seen in the responses of cones. [From Baylor (1987, Fig. 11, p. 42).]

flashes were nominally delivered – a little more than one flash every 10 seconds. The response of the cell is variable from one flash to the next, as predicted. Changes in the membrane current occur on some but not all trials; and when they occur, they are usually of a stereotyped form, either small or large in size. In sum, and remarkably, rod photoreceptors can indeed initiate a measurable signal from the absorption of a single quantum of light[3].

Similar experiments show that the photocurrent produced in a rod by a single quantal absorption is the same regardless of the wavelength of the quantum. Figure 6.10B shows responses to flashes of 550 and 659 nm lights that each produced one quantal absorption. Rather than different responses to a single quantal absorption, differences in sensitivity for different wavelengths are due to differences in probability of absorption.

### 6.4.2 Higher light levels: A saturating non-linearity

What about the photoreceptors' responses to multiple photons at higher light levels? Figure 6.11 shows photocurrents measured from dark adapted primate rods (A) and cones (B), measured with suction electrodes. The traces show current flow in response to brief flashes of lights with varying radiance. For both rods and cones, as light levels increase, the amplitude decreases (goes from

[3] In cones as in rods, single quanta are caught by single photopigment molecules, and single quantal absorptions must initiate functional physiological signals. But a rod produces a much larger signal than a cone does in response to the absorption of a single quantum. Some estimates suggest that the difference in the magnitude of current produced is as much as 100 to 1. Thus, the magnitudes of the cone responses to individual quanta are too small to measure, even with suction electrodes.

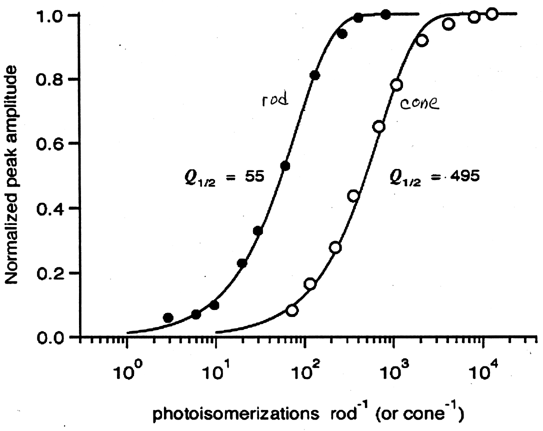


Figure 6.12: Dynamic ranges of rods and cones. Peak responses of a rod (left) and a cone (right) to flashes of light of increasing intensity. The abscissa shows the number of photoisomerizations per rod or per cone. $Q_{1/2}$ is the light level required to produce a half-maximal response; this is about 55 quanta for rods and 495 quanta for cones. Both rods and cones show saturating nonlinearities. The range over which the cell's response changes with changes in the light level is called the dynamic range of the cell. [Schnapf, personal communication.]

negative to zero). The time course of the response also changes with the light level. For rods, as the quantal catch increases, latency decreases. The cone response is biphasic but shows the same trend. Eventually, at high enough light levels all of the sodium channels are closed, the current is reduced to zero, and no further changes in current amplitude can occur (especially for the rods).

Figure 6.12 shows the peak change in the membrane current of rods and cones as a function of the number of quantal absorptions. The size of the response increases with the number of quanta absorbed, but the response eventually saturates. That is, photoreceptors show a *saturating non-linearity.*

The *dynamic range* of the photoreceptor is defined as the range of inputs over which the photoreceptor's output changes. Dynamic ranges for both rods and cones can be inferred from Figure 6.12. Defined in terms of the peak change in current flow, the dynamic ranges of both the rod and the cone cover about a factor of 100, or two log units: from about 2 to about 200 quanta absorbed for rods, and from about 20 to 2,000 for cones. We will see in Chapter 10 that these descriptions apply only to the dark-adapted rod and the dark-adapted cone, and things become more complex when light adaptation processes are included.

## 6.5 Linking theory for wavelength encoding

In Chapter 2, we introduced two properties of scotopic vision. First, scotopic spectral sensitivity varies with the wavelength of light, with a maximum at about 500 nm and a sharp falloff of sensitivity to either side. And second, the ability to preserve wavelength information is lost in scotopic vision: lights of all wavelengths look whitish, and they can be matched to one another – made indiscriminable – simply by adjusting relative light levels. Some people see these two

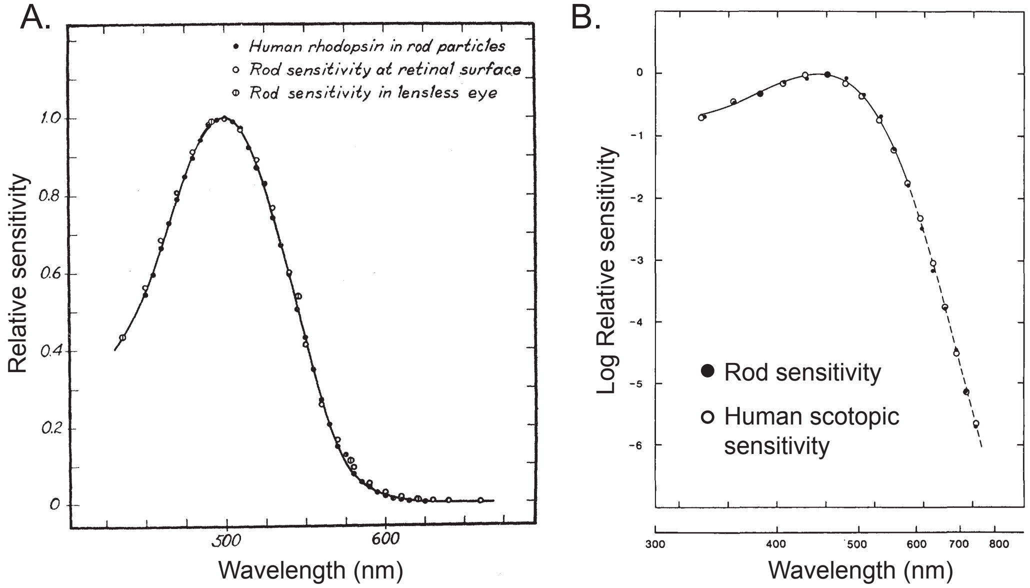


Figure 6.13: Comparing spectra. A. Comparing the absorption spectrum of rhodopsin and scotopic spectral sensitivity. All conditions show sensitivity as a function of wavelength relative to peak sensitivity. The rhodopsin measurement was from a suspension of rod particles from a human eye. The psychophysical data are human scotopic sensitivity measurements from two sources: one an observer without a lens and the other an average of several normal observers with a correction for the lens. The correction for lens, or the lensless observer are necessary to match the results on rhodopsin alone. B. Comparing the spectral sensitivity of responses by single monkey rods to human scotopic sensitivity. This comparison has a larger range of wavelength and a logarithmically scaled sensitivity axis to show the correspondence of small values at the tails of the function. As with the first comparison, the scotopic sensitivity has been corrected for the wavelength dependent effect of the lens. There is a close correspondence between the physiological and the behavioral spectra.[A. From Wald and Brown (1958, Fig. 6, p. 225). B. Modified from Baylor (1987, Fig. 14, p. 44) based on the data of Baylor et al. (1984).]

properties as intuitively contradictory. How can sensitivity vary with wavelength if wavelength information is lost? In fact, the transduction process precisely accounts for both.

Figure 6.13A shows a comparison of the scotopic sensitivity curve of human subjects to the absorption spectrum of human rhodopsin. The physiological rhodopsin data are shown by the solid symbols. The figure includes two kinds of psychophysical data. Scotopic sensitivity from an observer without a lens is shown by the open-and-dashed circles and the average of several normal observers with correction for the wavelength dependent effect of the lens are shown by open circles. The correction for the differential absorption of light of different wavelengths by the optics of the eye (Figure 4.9) is critical for the data to match at short wavelengths. For the normal observers, the correction for the lens yields the same results as an observer without a lens that needs no correction. All three data sets have their maxima at about 500 nm, and fall off virtually identically, both at shorter and at longer wavelengths. The fit between the data sets is remarkable, particularly given it compares psychophysical and physiological data.

This is such an beautiful comparison it deserves an encore. Figure 6.13B shows a comparison of the scotopic sensitivity curve of human subjects to the spectral sensitivity of individual rods measured with the suction electrode method. The rod sensitivity data are shown by filled symbols and the human scotopic sensitivity data are shown by the open symbols. Both are fit by a theoretical curve that is used to describe a variety of photopigments[4]. As with the other comparison, the scotopic sensitivity curve has been corrected for the effect of the lens. Once again, the physiologically defined rod sensitivity data closely corresponds with the psychophysical defined scotopic sensitivity curve.

There is a subtle but important difference between the spectra in Figure 6.13 and the spectra for scotopic spectral sensitivity in Figure 2.6. The spectra in Figure 6.13 were measured by retinal physiologists that were primarily interested in rods and photopigments. The measurements were made outside the eye and consequently did not involve the eye's optics. For the behavioral data, they removed the effect of the lens in plotting the scotopic spectral sensitivities. The result is a spectra with a broader shoulder for short wavelengths (e.g. Figures 6.5 and 6.13). In contrast, the measurements of spectral sensitivity in Figure 2.6 were made by vision scientist that was interested in the whole system including the optics. The result is the sharp fall off for short wavelengths that is primarily due to the lens and not the photopigment. In sum, keep a sharp eye on what is being measured. Is it a particular part of the physiology (a pigment) or is the the action of the whole system (pigment and optics).

When a rod absorbs a quantum, an isomerization occurs, but the effect is exactly the same regardless of the wavelength of the quantum. It follows that the rod has *equivalence classes*: sets of stimuli which, even though they are physically different from each other, are rendered identical by the transduction process. Stimuli that lead to a rod catching equal numbers of quanta are in an equivalence class for that rod. It is these quantal equivalences that lead to the suprathreshold discrimination failures seen in scotopic vision. So these two properties of transduction – the variation of the probability of quantal absorption with wavelength, and the loss of wavelength information at the instant of quantal absorption – are exactly sufficient to model the two system properties of scotopic vision – the shape of the spectral sensitivity curve, and the existence of equivalence classes.

By what criteria do we evaluate the quality of a linking theory? The more fully established the facts in both psychophysics and physiology, the better the match of details between the two sets of facts, the fewer the free parameters, and the fewer the reasonable alternative explanations, the more compelling the linking theory. In this case, both the rhodopsin spectrum and the scotopic spectral sensitivity are known from direct measurements, and equivalence classes exactly like the ones originally discovered psychophysically can be demonstrated physiologically by recording rod signals to light of different wavelengths (Figure 6.10B). The story fits together perfectly, with no questionable assumptions and no free parameters. In short, this is one of the most compelling linking theories in vision science.

[4] Figure 6.13B differs from most spectra shown in this book in using wavelength scaled by wavenumber (1/wavelength). This was used because the theoretical curve in the figure has the a constant shape for many different pigments when using a wavenumber axis.

## 6.6 Linking theory for absolute thresholds

In 1942, Hecht, Schlaer, and Pirenne carried out a psychophysical experiment on absolute detection thresholds. The stimulus was a test spot that subtended 10 minutes of visual angle. It was placed 20 degrees eccentric to the fovea, near the region of maximum density of rod photoreceptors (Figure 6.2). After the subject adjusted fully to the dark, detection thresholds were measured using the Yes/No method of constant stimuli.

Hecht and his colleagues then made careful calibrations of the light source, and combined these with estimates of the fraction of quanta that are lost within the eye rather than absorbed by the rod photoreceptors. Based on these estimates, they concluded that the subjects could detect the test spot when a total of only 5-10 quanta were absorbed by the whole set of photoreceptors covered by the test spot. This system property implies that the absorption of 5-10 quanta by a set of neighboring photoreceptors is sufficient to initiate a signal that traverses every stage of processing in the visual system. An elegant elaboration of this argument is presented in the opening chapters of Cornsweet (1970).

Moreover, comparisons to retinal anatomy showed that the 10 minute-of-arc test spot covered several hundred rod photoreceptors. Hecht and colleagues calculated that with only 5-10 quanta required for detection, the probability was very low that any one rod would have absorbed two or more quanta. Thus, they also concluded that the absorption of a single quantum must be sufficient to make a detectable signal in an individual rod[5].

About 40 years later, this prediction was confirmed by direct physiological measurements using the suction electrode technique described earlier in this chapter. The key data are shown in Figure 6.14. It shows a summary of the responses of individual rods to weak flashes of light that were previously illustrated in Figure 6.10. Data from different rods are presented in each of the four panels. Within each panel there are two graphs. The main graph is a histogram of the responses to a weak flash of light. In addition, there is a second inserted graph that is a histogram of the responses on trials in which no light was presented. Consider Panel A: The insert shows that all of the responses on trials with no flash had amplitudes that were between -0.5 and 0.5 pA. With a flash, the responses ranged from about -0.5 to over 2.0 pA. The other three cells showed a similar pattern. Clearly the flash of light results in quite a few trials with much higher amplitudes than occur without the flash of light.

How can we use these distributions to understand just how sensitive is a single rod? This can be done using the theory of signal detection introduced in Chapter 2 for analyzing yes-no behavioral data. Recall Figure 2.4 that illustrated how the behavioral responses can be predicted from two theoretical distributions: one for trials with the critical stimulus (a flash) and one for trials without the critical stimulus (no flash). To generate a response, one chooses a criterion that at least partially discriminates the two kinds of events. Taking the data from the rod in Panel D, one could pick a criterion of 0.5 pA. All values below that would be judged "no" and all values

[5] The claim of such exquisite sensitivity was not immediately accepted by all vision scientists. Teller and a friend both went to graduate school in the early 1960's, Teller in psychology and the friend in biochemistry. It turned out that we were both assigned to read the Hecht, Shlaer, and Pirenne paper (for the friend it was part of an examination question on cell biology). The friend argued that the energy in a quantum was not sufficient to make a signal that could traverse the whole rod photoreceptor, and therefore that Hecht et al's conclusion must be wrong. Teller maintained, on the basis of a bumblebees-can-fly argument, that a single quantum must be sufficient; and that new mechanisms of photoreceptor function, consistent with the psychophysical data, must remain to be discovered. The photocurrent and the amplification provided by the chemical cascade eventually resolved the argument.

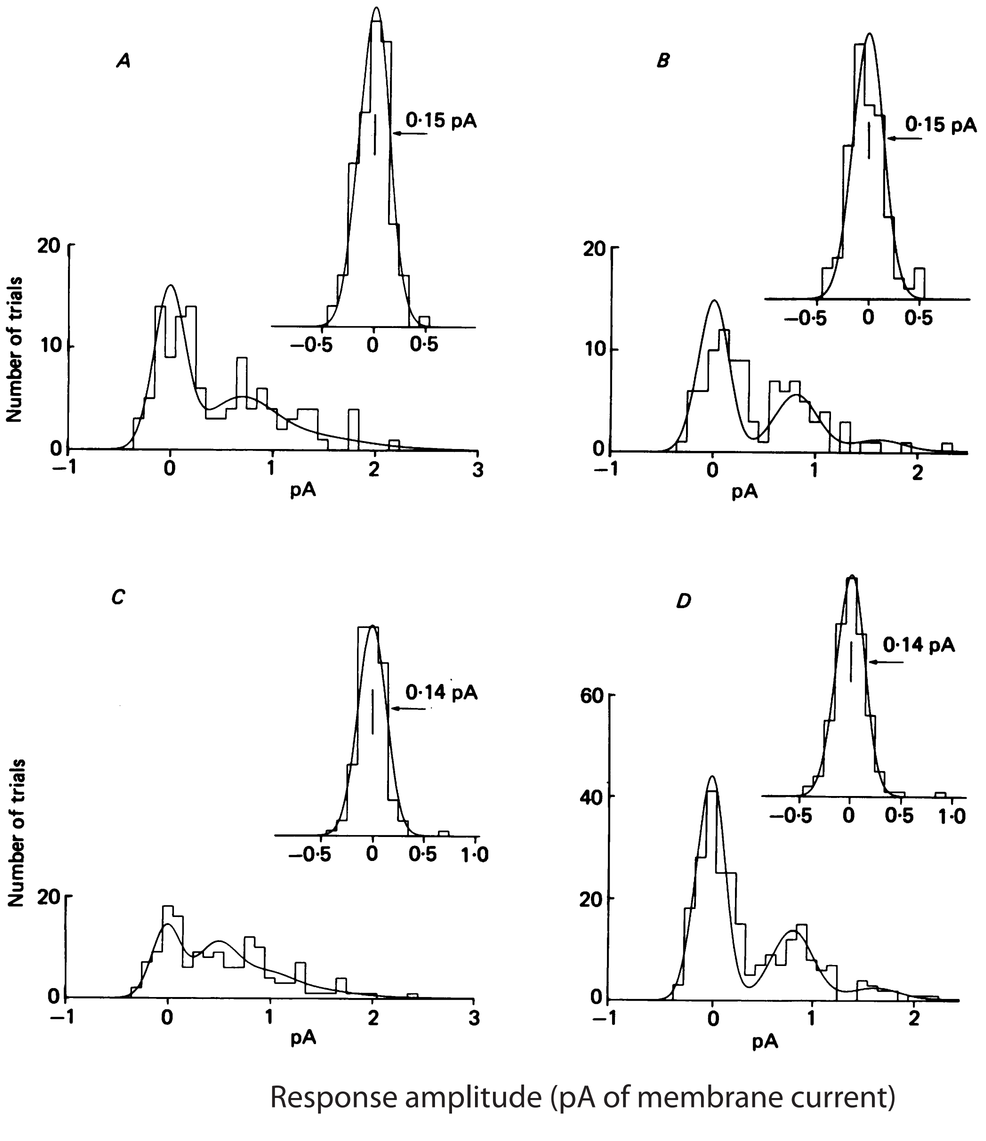


Figure 6.14: Rod responses to weak flashes of light. Each panel shows a histogram of the response amplitude from a single rod to a weak flash of light. Specifically, the histograms plot the number of trials on which a given level of membrane current was observed during that trial. The main graph in each panel is the response to a weak flash of light. The inserted graph in each panel is the response during a trial with no flash of light. Responses during these "blank" trials were very consistent with values ranging from -0.5 to 0.5 pA. In contrast, the responses to flashes of light included a much larger range of amplitudes from -0.5 to 2.0 pA. The smooth curve drawn through the histogram is the fit of a theory in which the cells response is a mixture of responses to zero, one, or very infrequently two quanta. One can see the separate peaks for zero and one quanta in these smooth curves. These results are consistent with the information from a single rod being sufficient to detect a single quanta on a substantial fraction of the trials. [From Baylor, Nunn, and Schnapf (1984, Fig. 14, p. 44).]

above would be judged "yes". For the trials without a flash, this criterion would result in nearly perfect performance: all correct rejections except for one false alarm. For trials with a flash, this criterion would result in intermediate performance with about 50% hits and 50% misses. Put these two kinds of trials together, and the information from this cell allows for judgments of the presence or absence of a flash that are correct on about 75% of the trials. Not bad for a single rod and light flashes that on average result in the absorption of only a single quanta. In future chapters, we replace this informal calculation with systematic evaluations of the sensitivity of individual neurons using the methods of signal detection theory.

The linking theory, then, is that the detection of lights that yield a quantum catch of only 5-10 quanta over a $10'$ field is mediated by the exquisite sensitivity of individual rod photoreceptors that can detect single photons with considerable accuracy[6].

## 6.7 Relevant linking propositions

There are no new linking propositions involved in the two linking theories just described. Both depend only on either threshold or matching judgments, and thus both linking theories rely on identity propositions. Consider them in turn.

The measurement of spectral sensitivity depends on detection experiments to define a set of stimuli that have the same effect on behavior. This is a the set of wavelengths and intensities that are just visible. The measurement of absorption spectra similarly depend on identifying a set of stimuli that have an equivalent amount of light absorbed by a photopigment. Thus, each data set in Figure 6.13 is an equivalence class of lights defined by identity.

The measurement of responses to a few photons of light also involve simple detection experiments. The twist is that these judgments are subject to random variation from trial to trial. The quantal nature of light makes this true even of the stimulus. To interpret such inherently noisy data, we will depend on ideas from signal detection theory.

### 6.7.1 The Nothing Mucks it Up proviso

We now introduce another kind of linking proposition that philosophers call a *ceteris paribus condition* – other unspecified things being equal, the argument holds. Teller (1980) calls it the *Nothing Mucks It Up proviso.* It is the implicit assumption that nothing else in the visual system interferes with the identified physiological processes determining the relevant psychophysical phenomenon.

For example, in the case of equivalence classes for scotopic spectral sensitivity, the Nothing Mucks It Up proviso includes the assumption that the rods are the only photoreceptors that mediate vision across the whole range of conditions tested. That is, there are no neural elements that are more sensitive than the rods at any wavelength. If rods do in fact provide the only source of information about these lights, then no code transformations can occur that changes the equivalence classes of either the psychophysical and physiological data. In fact, we know that the human retina contains cones as well as rods. And in fact, they do "muck up" this linking theory at higher light levels, as we will see in Chapter 7. So this theory of rods and scotopic vision applies to only the low light levels in which rods are in fact the only active photoreceptor.

[6] To psychophysics chauvinist Teller, part of the fun of the story is that the psychophysics came first.

## 6.8 Summary: Photoreceptors and system properties of vision

In this chapter, we introduced the anatomical structure of rods and cones. We learned that rods contain the photopigment rhodopsin, which has an absorption spectrum with a maximum at about 500 nm. We described the molecular structure of rhodopsin, and the cis-trans isomerization of the rhodopsin molecule. It is this small change in the shape of the molecule that accomplishes the transduction from light to physiological signals.

We then reviewed the properties of typical neurons, in order to stand them in contrast to the properties of photoreceptors. We introduced the photocurrents that flow around photoreceptors in the dark, the way they are changed by the action of light, and the chemical cascade that amplifies the neural signal at the outer membrane.

We then had a look at physiological recordings from individual photoreceptors. In particular, we saw that the responses to individual quantal absorptions can be recorded from photoreceptors. We also showed how the photoreceptor response changes with light levels and the saturating nonlinearities at high light levels.

Finally, we examined two examples of linking theories about how rod photoreceptors leave their marks on our visual perception. The first asserts that the transduction process in the rods accounts for the two system properties of scotopic vision developed in Chapter 2: spectral sensitivity and equivalence classes. The second asserts that a rod can signal the absorption of a single quantum and that signal determines the value of the absolute threshold in human vision. These two linking theories seem highly credible. We are on a roll, and the question becomes, how much longer can we go on before our linking theories begin to leave more room for doubt?

In Chapter 7, we turn to the ways in which the cone photoreceptors leave their marks. In particular, we examine the consequences that the presence of three cone types has for the processing of wavelength information and for color vision.

# Chapter 7

# The Trichromacy of Color Vision

## Contents

In Chapters 2 and 3 we introduced some psychophysical facts concerning how our visual perceptions change (or don't change) with changes in the wavelength and intensity of light. At low light levels (scotopic vision), all wavelengths of light look the same whitish color. Some patches of light will look brighter than others, but given variations in only physical intensity, lights of all wavelengths can be made to look identical. That is, in scotopic vision metamer sets include all wavelengths of light, and all wavelength information is lost.

In Chapter 5 we developed a model to explain this fact. The model assumes that only a single photoreceptor type, the rods, is functional at scotopic light levels. A single photoreceptor type cannot preserve wavelength information because wavelength information is discarded in the transduction process. Each rod sums quantal catches linearly, and any two lights that lead to the same total quantal catch in the rods will be metameric – they will look identical to a human subject.

When stimuli are at photopic light levels, however, subjects experience color variations, as discussed in Chapter 3. Thus, they readily discriminate among lights of different wavelengths because lights of different wavelengths differ in perceived color, and the colors are sustained (at least approximately) across variations of intensity. We therefore know immediately that photopic vision cannot be based on a single univariant photoreceptor class like the rods. The physiological model adopted for scotopic vision must be rejected for photopic vision because it fails to account for the system property – the preservation of wavelength information.

You may be surprised to learn, however, that even in photopic vision there are sets of lights of different wavelength compositions and intensities that are indiscriminable from each other. As described in earlier chapters, such sets are called *metamers*. The nature of these psychophysically defined sets of lights is described by a psychophysical law, *the law of trichromacy*, which will be presented in detail below. And the question is: why do these metamer sets occur?

To address this question we depart from our practice of avoiding mathematical formulations, and present a mathematical model of trichromacy. We do this because the model is a particularly simple use of algebra (three simultaneous linear equations in three unknowns), yet it is elegant and sufficient to the modeling task. Moreover, the historical interplay between the psychophysical law and the mathematical model, leading on to the discovery of the physiological and genetic entities that instantiate the model, provides one of the loveliest examples of progressive explanation in vision science (see Mollon, 2003, for a historical account).

In addition, trichromatic matches provide a new and interesting example of the use of identity matching, and thereby of the Identity family of linking propositions. The Identity family was first introduced in Chapter 2 in our account of thresholds and scotopic matching. Watch for the use of Identity propositions as we go along.

## 7.1 The system properties of trichromacy

Let us begin by introducing three psychophysical facts about wavelength discrimination. First, as already discussed, we can discriminate among lights of different wavelengths because different wavelengths look different colors. Second, some mixtures of physical wavelengths can be discriminated from other mixtures and from any single wavelength selected from the spectrum. Whites, purples, and desaturated colors (pink, baby blue, light green, etc.) are examples of colors that arise only from wavelength mixtures. They are called *non-spectral* (or *extra-spectral*) colors, meaning that we cannot match them to any individual wavelength.

### 7.1.1 Metamer sets in photopic vision

Third, metamers occur in photopic as well as in scotopic vision. That is, even at photopic light levels there are sets of lights of very different wavelength compositions and intensities that look identical. The membership in these metamer sets is initially counterintuitive and odd.

For example, Figure 7.1 shows a plot of the *complementary wavelengths* of light. In vision science, the term complementary wavelengths is used to describe pairs of wavelengths that *look white* when mixed together in proper proportions. This diagram tells us that many different mixtures of wavelengths all look white[1]. Moreover, by varying the intensities of the different mixtures you could match them all in perceived brightness, with the result that they would all look *identical*, even though they are *very* different physically. And there is an infinite number of other combinations of three or more wavelengths that all look white. Similarly, we can make a set of stimuli of many

[1] Unfortunately the term "white" is used in both physics and psychophysics, and this causes confusion. In physics the term *white light* is often taken to mean an equal energy mixture of all wavelengths, like the last mixture in Figure 7.1B. This definition is troublesome in vision science, because many different physical stimuli (wavelength mixtures) actually appear white, as the rest of Figure 7.1 shows. In fact, a whitish appearance tells us remarkably little about the wavelength composition of a light. To our knowledge there is no word in either physics or psychophysics for "physical stimuli that are members of the perceptual white metamer set". The closest phrase is "metameric to an equal energy light".

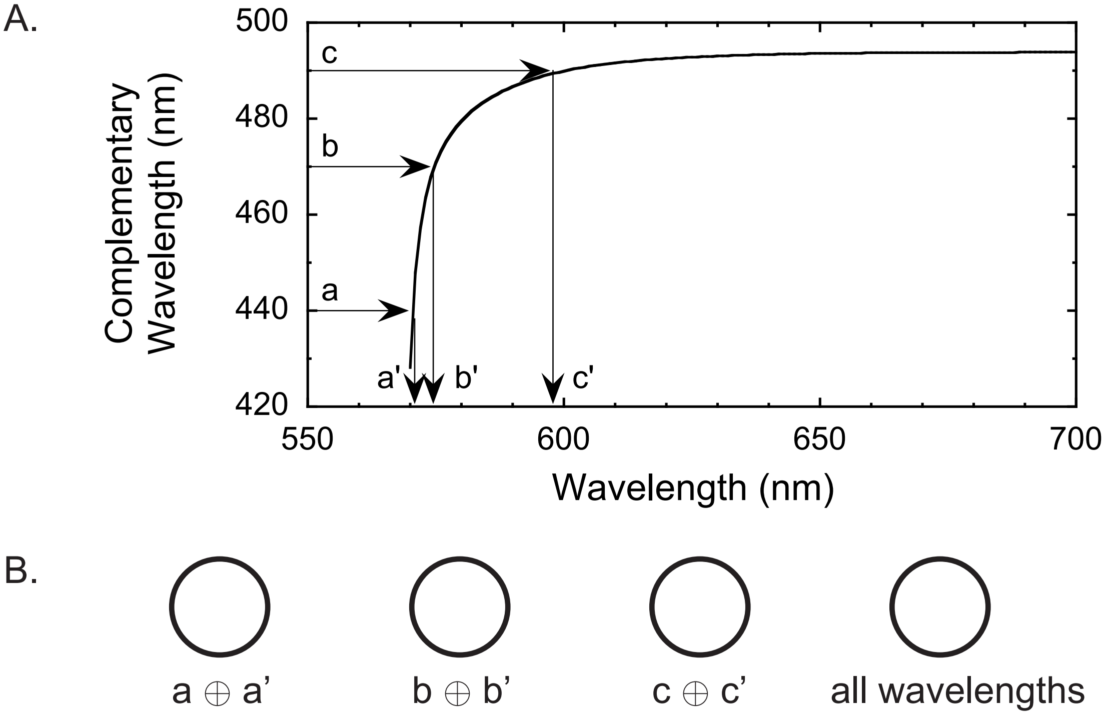


Figure 7.1: Complementary wavelengths: mixtures that look white. A. For each wavelength from about 570 nm to 700 nm, it is possible to achieve a white-appearing spot by combining that wavelength in the proper proportion with the properly chosen complementary wavelength, which will fall somewhere between 420 and 500 nm. B. For properly chosen intensities, the row of lights shown here would all look white, and would be perceptually indiscriminable. The last spot on the right is a mixture of equal energies of all wavelengths. In addition to those shown, many other mixtures would also look white – for example, any of many mixtures of three wavelengths, four wavelengths, and so on. (The plus sign in the circle ⊕ is the symbol for superposition: we are superimposing one light on another.)

different wavelength compositions that all look identical and a particular shade of yellow; another set that all look identical and a particular shade of light blue; and so forth. Each of these sets of lights is a metamer set.

In ordinary experience we usually don’t notice the existence of metamer sets, because metamers are such perfect perceptual facsimiles of each other that the fact that they are physically different passes unnoticed. But here’s an example from our experience. Imagine a light fixture that consists of two light bulbs inside a translucent globe. Around Christmas time, someone takes out the two ordinary light bulbs and replaces them with a “red” bulb (a bulb that emits a band of long wavelengths, say, above 620 nm) and a “greenish-yellow” bulb (a bulb that emits a band of middle wavelengths, say, between 530 and 560 nm). After replacing the globe, one half of the it looks red and the other half greenish-yellow. But between the two, there appears a band of very distinct and saturated yellow.

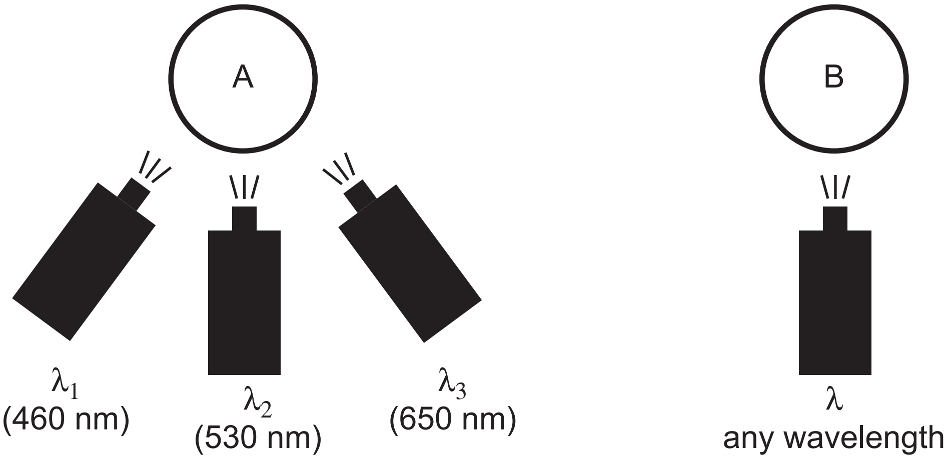


Figure 7.2: Set-up for a demonstration of trichromacy. Light from the three projectors on the left is superimposed to make patch A. The projectors are fitted with narrow-band filters to provide wavelengths of (say) 460, 530, and 650 nm. Each projector also has a knob that controls its intensity. Patch B comes from a fourth projector that can be set to provide any wavelength or combination of wavelengths at any intensity.

Why was the yellow band there? Not because the band was illuminated by a light of an isolated wavelength that looks yellow (say, 575 nm), but because it contained just the right mixture of light from the "red" and "green" bulbs. The mixture of wavelengths coming from the band belonged to the yellow-appearing metamer set. But most people who walked by would not have even wondered why the yellow band was there. Of those who wondered, most would probably have assumed that the globe must have contained a source of 575 nm light. Only a few would have guessed that nothing but broadband "red" and "green" bulbs were hidden inside the globe.

To emphasize again the oddity of color mixture, the perception of combining lights of different wavelengths is very different from the perception of combining sounds of different frequencies. With sound vibrations of different temporal frequencies, we hear tones of different pitch. Playing these tones together, we hear *chords* that still perceptually contain the original tones; we don't hear an intermediate tone, much less a completely novel tone or no tone at all. Why is the mixing of lights different?

### 7.1.2 The psychophysical law of trichromacy

As it turns out, metamer sets are not as arbitrary as they originally seem. They follow a particular rule, called the *law of trichromacy* (tri = three).

Figure 7.2 shows a laboratory set-up for demonstrating trichromacy. It involves a set of three slide projectors, fitted out with devices for allowing continuous variation of their intensities. In front of each projector is a narrow-band color filter, and overlap the three beams on a projection screen to make a patch of light, A. In other words, patch A is a mixture of three lights of different wavelengths, $\lambda_1$, $\lambda_2$, and $\lambda_3$. A useful set of choices (cf. Figure 7.4) is to let $\lambda_1 = 460$ nm (which looks predominantly violet), $\lambda_2 = 530$ nm (which looks predominantly green), and $\lambda_3 = 650$ nm

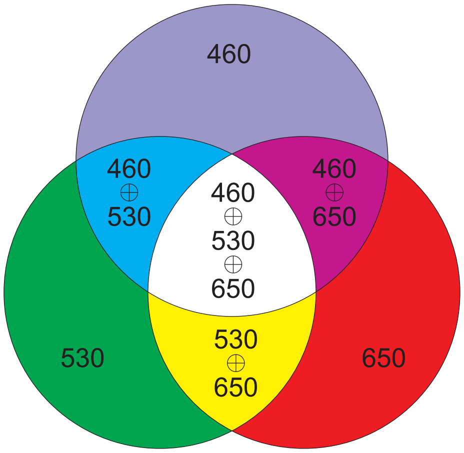


Figure 7.3: A simulation of color mixture. In this figure, the three superimposed beams of Figure 7.2 have been partially separated in space. The outer crescents simulate the colors of each of the three original wavelengths. The outer "triangles" simulate the colors of combinations of two wavelengths, and the central "triangle" simulates the color resulting from the combination of all three wavelengths. (This figure is only a simulation and not a true demonstration, because the colors of the various segments are simulated with the broadband pigments used in printing, rather than being made from narrow wavelength bands.)

(which looks predominantly red)[2]. The set-up also has a fourth projector to make a second patch of light, B. Then by using a variety of filters in turn, we can make patch B appear with any color and brightness.

The law of trichromacy, informally stated, is that by varying only the intensities of the three wavelengths in patch A, we can make a perfect perceptual match to (almost) any other light in patch B; that is, we can make A $\equiv$ B (A is metameric to B). Different colors and brightness in patch B will require different intensities of the three wavelengths in patch A, but an exact perceptual match will almost always be possible.

In fact, we can exactly match any light in patch B if we are allowed to move one of the three wavelengths from patch A to patch B. This additional provision leads to the formal statement of the law of trichromacy: Given any four lights, we can arrange them with three in patch A and one in patch B, or two in patch A and two in patch B; vary the intensity of any three of them, and end up with a perfect perceptual match between the two patches. Figure 7.3 shows a simulation of the appearance of all combinations of the three wavelengths.

In 1928, W.D. Wright carried out a classic study of trichromatic color matching. Wright tested 10 subjects with mixtures of 460, 530, and 650 nm in patch A. He set up patch B with each of a series of individual wavelengths of light (e.g. 405 nm, 410 nm, 415 nm, etc). For each wavelength in patch B, the subjects adjusted the intensities of the three lights in patch A to make metameric matches between the two patches.

[2] The three lights we use as the mixture set are sometimes called *primaries*. However, this term is used in different ways in other contexts. For example, the "primary" colors you learned about in kindergarten make use of a different meaning of the term.

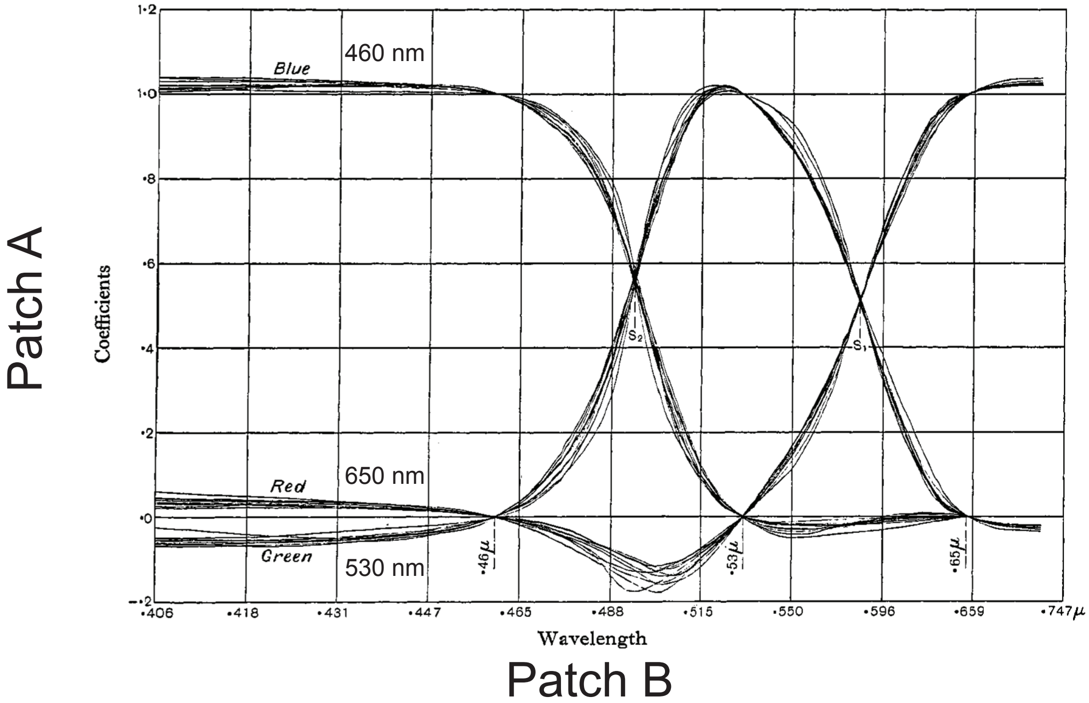


Figure 7.4: Data from a trichromatic color mixture experiment. The abscissa shows the wavelength of light in patch B. The ordinate shows the proportions of 460, 530, and 650 nm lights in Patch A required to make the two patches metameric. The lines show fits to the data from 10 subjects. The negative coefficients of the primaries in for some wavelength regions imply that the primary was moved from Patch A to Patch B. [From Wright (1928, Fig. 2, p. 148).]

Wright's data are shown in Figure 7.4. In the short wavelength range, below about 460 nm, color changes very little with wavelength. These lights all look predominantly violet, and each can be nearly matched with the 460 nm (labeled "blue" in the figure) primary alone. However, a small amount of the 650 nm ("red") primary must be added, and a small amount of the 530 nm ("green") primary must be subtracted – moved to the other side and mixed with the light in patch B – in order to make the matches. At 460 nm, of course, the 460 nm primary provides an exact match. Between 460 and 650 nm, perceived hue changes more rapidly with wavelength, as do the proportions of the different primaries required for matches. Notably, the 650 nm ("red") primary must be subtracted for all wavelengths between 460 and 530 nm, as must the 460 ("blue") primary for all wavelengths between 530 and 650 nm.

Trichromacy is a remarkable and puzzling system property of photopic vision. Why do wavelength mixtures behave the way they do? Why are the metamer sets as they are? Why are three lights enough? The answer lies in our visual systems.

### 7.1.3 Reprise on the Converse Identity proposition

Let us do the exercise of ferreting out a linking proposition. The Identity family of linking propositions was introduced in Chapter 2 (Table 2.3), in the context of matching and thresholds. Trichromatic metamers are sets of stimuli that are very different physically but appear identical perceptually. That is, they are another case in which subjects are carrying out a matching task. The data are perceptual, so to explain them physiologically we will be trying to reason from perception to physiology. Thus the two relevant Identity propositions are the Contrapositive and the Converse.

Moreover, the basic perceptual observation is that patches A and B match, so the relevant linking proposition is the Converse: perceptual identity implies physiological identity. Thus, if the theorist assumes the truth of the Converse Identity proposition, metameric matches imply that the signals that arise from a set of metameric stimuli are rendered physiologically identical somewhere within the visual system.

The linking theory for trichromacy then becomes a locus and coding question. Where within the visual system do the signals originating from the physically different stimuli become identical, and by means of what physiological processes and computations?

## 7.2 A mathematical model of trichromacy

### 7.2.1 Assume three Fundamentals

The mathematical model of trichromacy starts with the set of mathematical or physiological assumptions shown schematically in Figure 7.5. The model assumes that photopic vision is served by three Fundamentals – three mathematical/physiological entities with different spectral sensitivity curves. The peak sensitivities of the three Fundamentals are assumed to differ, but the ranges are assumed to overlap substantially (for simplicity, they overlap entirely in Figure 7.5A). The model also assumes that each Fundamental forms a linear summation of the signals resulting from different wavelengths of light.

For concreteness, in the following pages we identify the Fundamentals with three types of cones. But it's interesting to notice that the mathematical model of trichromacy preceded any direct evidence of the numbers of cone types or their spectral sensitivity curves. We will return to this point below.

As was the case for rods in Chapter 2, we can represent each cone type with a funnel that counts the quanta it catches, without keeping track of their wavelengths (Figure 2.7, in which the marbles can now be identified as quanta). In the case of photopic vision there are three funnels, each with its own counter. The three-funnel analogy is shown in Figure 7.5B.

Now we need to develop some symbols. To identify each Fundamental (cone type) with the wavelength range of its maximum sensitivity, the three Fundamentals will be called L, M, and S[3]. The letters *L, M,* and *S* in italics will denote the quantum catch rates generated in the L, M, and S cones respectively. Because we are dealing with two patches of light, *A* and *B*, there will be two sets of cones (two sets of funnels in the analogy), one for patch A and one for patch *B*. Let the quantum catches resulting from patch *A* be $L_A$, $M_A$, and $S_A$, and from patch *B* be $L_B$, $M_B$, and $S_B$ respectively.

[3] The different cone types have historically been called "red cones", "green cones" and "blue cones". Vision scientists avoid this terminology, in order to maintain the clear separation of perceptual and physiological terms. If the terminology is sloppy, it is easy to think that color vision is simple – we see red because we have "red cones"! We use color names to refer to perceived colors, and a different set of names – *S*, or *short-wavelength-sensitive, M,* or *mid-wavelength-sensitive*, and *L*, or *long-wavelength sensitive* – to refer to cones.

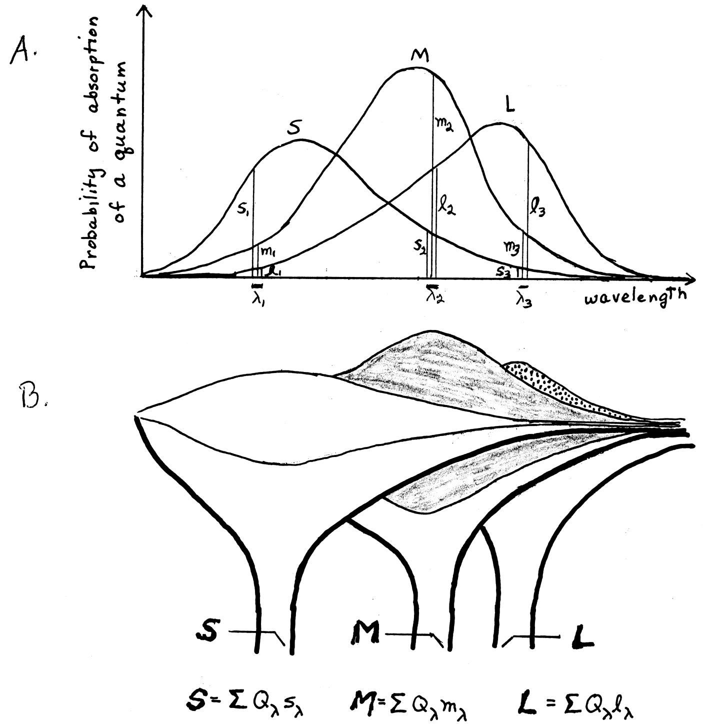


Figure 7.5: An illustration of the mathematical model of trichromacy. A. Completely overlapping spectral sensitivity curves assumed for the three Fundamentals for purposes of illustration. B. Extension of the funnel analogy to the case of trichromacy. Each funnel captures a broad range of wavelengths with probabilities determined by its width at each wavelength, and counts its total quantum catch without regard to wavelength. The result is a set of three variables, *S, M,* and *L.*

### 7.2.2 The condition for metamerism

If lights of different wavelength composition yield identical quantum catches in *each* of the three hypothetical photoreceptor types, then these lights will be metamers. Using the symbols just defined this statement can be formalized as:

$$\text{if } L_A = L_B \text{ and } M_A = M_B \text{ and } S_A = S_B, \text{ then } A \equiv B.$$

This statement can be called the *condition for metamerism.* But under what circumstances is the condition for metamerism satisfied? Is it a fool's dream, or a realistic basis for a model?

### 7.2.3 The color equations

What does light of a given wavelength, $\lambda$, do when it encounters a three pigment system like that shown in Figure 7.5? It makes a triplet (a set of three) quantum catch rates, one in each of the three cone types. For any given wavelength such as $\lambda_1$, the heights of the three curves at that wavelength tell us the probabilities of absorption of a quantum by each cone type. Let,

- $\mathrm{l}_1$ = the height of the L curve at $\lambda_1$,

- $\mathrm{m}_1$ = the height of the M curve at $\lambda_1$, and
- $\mathrm{s}_1$ = the height of the S curve at $\lambda_1$.

In addition, let $l_2$, $m_2$, and $s_2$, and $l_3$, $m_3$, and $s_3$ be similarly defined. Since we assumed the shapes of the three pigment curves in Figure 7.5, all of these values of curve heights are constants in the equations below.

Now we need symbols for the *intensity* of the lights from each of the three projectors; that is, for the rate of arrival of quanta of each wavelength $\lambda_1$, $\lambda_2$, and $\lambda_3$. These intensity values are just the intensities of the three projectors in Figure 7.2. The intensities of projectors 1, 2, and 3 will be called $Q_1$, $Q_2$, and $Q_3$ respectively. Since the subject varies these intensities to make the metameric matches, the $Q$'s will turn out to be the variables in the equations below.

We are now in a position to write expressions for the rate of quantal absorptions from each of the three wavelengths in each of the three cone types. The rate of quantal absorptions of the wavelength $\lambda_1$:

- by the L cones is: $Q_1 l_1$,
- by the M cones is: $Q_1 m_1$, and
- by the S cones is: $Q_1 s_1$.

and similarly for $\lambda_2$ and $\lambda_3$. Moreover, to calculate the total quantum catches in each cone type in response to any wavelength mixture, we just add up the quantum catches from all of the available wavelengths for each photoreceptor:

- L cone quantum catch is $L = \sum Q_\lambda l_\lambda$,
- M cone quantum catch is $M = \sum Q_\lambda m_\lambda$, and
- S cone quantum catch is $S = \sum Q_\lambda s_\lambda$.

Now let's return to our two patches, $A$ and $B$. Patch $A$ is composed of three wavelengths, $\lambda_1$, $\lambda_2$, and $\lambda_3$, with variable intensities $\mathrm{Q}_1$, $\mathrm{Q}_2$ and $\mathrm{Q}_3$. We can now write three equations to describe the three cone quantum catch rates produced by patch $A$:

$$L_A = Q_1 l_1 + Q_2 l_2 + Q_3 l_3,$$

$$M_A = Q_1 m_1 + Q_2 m_2 + Q_3 m_3,$$

$$S_A = Q_1 s_1 + Q_2 s_2 + Q_3 s_3.$$

What about patch $B$? Patch $B$ is a light of any chosen wavelength composition and intensity. For any specific choice of wavelength composition and intensity, patch $B$ generates a specific triplet of cone signals, $L_B$, $M_B$, and $S_B$. Once the wavelength composition is chosen, these entities are constants.

Now, the fundamental question is, by varying only the intensities $Q_l$, $Q_2$, and $Q_3$, can the triplet of values $L_A$, $M_A$, and $S_A$ be made identical to the triplet $L_B$, $M_B$, and $S_B$? That's the condition for metamerism.

Assume for the moment that the two lights *are* metamers. Then by the condition for metamerism we can substitute $L_B$, $M_B$, and $S_B$ for $L_A$, $M_A$, and $S_A$ to produce:

$$L_B = Q_1 l_1 + Q_2 l_2 + Q_3 l_3,$$

$$M_B = Q_1 m_1 + Q_2 m_2 + Q_3 m_3,$$

$$S_B = Q_1 s_1 + Q_2 s_2 + Q_3 s_3.$$

This set of three simultaneous equations contain three unknowns: the intensities $Q_l$, $Q_2$, and $Q_3$. Each of the little $l$, $m$, and $s$ values are known because they are specified by the heights of the spectral sensitivity curves $L$, $M$ and $S$, and $L_B$, $M_B$, and $S_B$ are known because they are specified by the choice of the light $B$. Each of the three cone types contributes an equation, and each of the three wavelengths contributes a variable.

But remember that it is an elementary property of linear algebra that three simultaneous linear equations in three unknowns are guaranteed to have a solution, so we know that for any specified values of $L_B$, $M_B$, and $S_B$ we can solve for the values of $\mathrm{Q}_l$, $\mathrm{Q}_2$, and $\mathrm{Q}_3$. It follows that for *any* specified set of values of $L_B$, $M_B$, and $S_B$ – that is, for any light in patch $B$ – these equations can be solved. Thus by varying only the radiances of the three wavelengths in patch $A$, patch $A$ can be made metameric to light of any wavelength composition in patch $B$. That is the informal statement of the law of trichromacy! So in sum, the equations provide a sufficient mathematical model of the law of trichromacy.

But there's one possible flaw in the argument. Remember that, although we are guaranteed a solution to three simultaneous equations in three unknowns, there is no guarantee that all of the values for $\mathrm{Q}_l$, $\mathrm{Q}_2$, and $\mathrm{Q}_3$ will be positive. One or more of them might be negative. But real lights cannot have negative intensities, so how do we interpret the negative values? The answer is, in algebra, one can move the negative term to the other side of the equation and make its value positive. In the matching experiment, one can move the corresponding light, $\lambda_1$, $\lambda_2$, or $\lambda_3$, from patch $A$ to patch $B$. From this convention results the formal statement of the law of trichromacy: Given any four lights, we can arrange them in two patches to make $A \equiv B$.

Return now to a more historically accurate picture. For the sake of concreteness we initially identified the three Fundamentals with three cone types, and the linear summation property with the loss of wavelength information in an individual photoreceptor caused by the properties of the transduction process. But historically, both the psychophysical fact of trichromacy and the mathematical model of trichromacy were established by about 1860, before we had any other evidence of the number of cone types, or the univariance of photoreceptors, or even any modern notion of quantum theory. It was a deep mathematical insight to see that a set of three simultaneous linear equations in three unknowns would provide a sufficient model for the perceptual fact of trichromacy, and to posit three Fundamentals with overlapping spectral sensitivity curves to provide the three variable system of simultaneous linear equations.

Finally, let's return to the properties of wavelength discrimination with which the chapter began. First, we can discriminate among lights of different wavelengths because they have distinctive colors regardless of variations in intensity. As shown in Figure 7.5, each different wavelength sets up a different set of *relative* quantum catches, l vs. m vs. s, in the L vs. M vs. S cones. Putting it another way: wavelength information is lost in each individual cone type, but it is preserved in the *ensemble* of three cone types by the *relative* quantal catches among them. This ensemble code is the form in which wavelength information passes through the photoreceptor level of processing.

Similarly, information about the *intensity* of the light is preserved in the *absolute* quantum catches in the three kinds of cones.

Second, some mixtures of wavelengths can be discriminated from other mixtures and from any single spectral wavelength. Why? Because not all of the possible ratios of cone signals are created by individual wavelengths of light, and some mixtures of wavelengths create these novel ratios. For example, by inspection of Figure 7.5, there are no individual wavelengths that create L/M/S ratios of 1:1:1, or 2:1:2, and so on; but you can find mixtures of wavelengths that will do so. When mixtures of wavelengths create these ratios, non-spectral colors appear.

The third property, metamerism, blends into the law of trichromacy, and has already been explained in detail.

### 7.2.4 Reprise on the Initial Identity proposition

Here's an exercise on linking propositions. Whereas the inference from behavioral trichromacy to three Fundamentals rests on a Converse Identity proposition, the mathematical model from fundamentals to behavioral trichromacy rests on the Initial Identity proposition. The mathematical model begins by assuming (seemingly arbitrarily) the existence of three physiological entities – the three Fundamentals. We argued mathematically that three such entities, acting together, would process physical stimuli in just such a way as to create the metamer sets observed psychophysically in human subjects, and summarized by the law of trichromacy.

The Initial Identity proposition – that identical physiological states imply identical perceptual states – enters the argument because we are trying to reason from (assumed) physiological states to perceptual states. Assuming Initial Identity allows us to use physiological identity to infer perceptual identity, and thus use the model to provide an account of the psychophysical data.

### 7.2.5 The physiological implications of trichromacy

Now let's step back a little. Clearly the system property of trichromacy, together with its mathematical model, places major constraints on physiological models of the visual system. Up until this point, for the sake of specificity and simplicity, we have identified the three mathematical Fundamentals with three cone types. This is the linking theory at he heart of of our understanding of color vision. But the true constraints are actually somewhat more general.

What the mathematical model of trichromacy actually suggests is that information available for discriminating among wavelengths and intensities of light passes through a serious bottleneck – or rather, three bottlenecks – somewhere on its way through the visual system. That is, there is a stage at which this information is limited to three variables. (A statistician would say the visual signal has only three degrees of freedom, and an engineer would say the system has only three information channels).

In the specific linking theory we have introduced, this three-channel stage is instantiated by three kinds of cones with three different photopigments. But logically, one could have a linking theory with more than three kinds of photoreceptors and have the information reduced to three variables by passing through a three-channel stage somewhere later in the visual system. And as it happens, in peripheral vision we do have rods as well as the three cone types, and the reduction to three channels does come later.

A final thing to note is that even if the photoreceptors are the bottlenecks, the argument above does not depend on assuming any particular set of shapes or wavelengths of maximum sensitivity for the spectral sensitivity curves of the three photopigments. We assumed three Fundamentals, L, M, and S, and the heights of these three curves at particular wavelengths were constants that entered into the color equations. But logically we could have assumed any of many shapes for the spectral sensitivity curves, as long as they are consistent with color mixture data like that shown in Figure 7.4.

## 7.3 Searching for the fundamentals

The psychophysical fact and the three-Fundamental model of trichromacy were both well established by about 1860 (Mollon, 2003). But the situation left vision scientists in a state of acute frustration. They knew that there are (probably) three cone types; but, they did not know their spectral sensitivity curves. Determining the spectral sensitivity curves of the three Fundamentals has thus been a fundamental challenge (pun intended) for vision scientists for 150 years, and scientists from several disciplines set out to determine the shapes of these curves. We will review three historical approaches.

First, some of the earliest relatively accurate estimates of the Fundamentals were derived from psychophysical measurements. Of these the most successful approach was based on the assumption that certain “color-blind” individuals (see below) have lost one of the Fundamentals, but that their two remaining Fundamentals are identical to the Fundamentals of normal subjects. Without the contribution of the third pigment, the available pigments in “color-blind” human subjects could be estimated psychophysically, and by hypothesis used to estimate the spectra of the normal pigments.

Excellent psychophysically-based estimates of the sensitivity maxima and the curve shapes of the cone Fundamentals emerged in the 1970s. They are shown with data from other, later techniques in Figure 7.6B, and it can be seen that the data correspond closely. However, the challenge persisted of verifying these estimates with more direct measurements that were free of the assumption that normal pigments occur in color-deficient subjects.

In the next historical iteration, the technique of *microspectrophotometry* was developed. In microspectrophotometry, one dissociates the cells of an excised retina, until individual photoreceptors can be seen floating free under the microscope. One can then shine a tiny beam of light through an individual photoreceptor and onto a photocell, and the percent of light absorbed can be measured for each wavelength in turn.

Cone spectral sensitivity curves measured with microspectrophotometry are shown in Figure 7.6B. Microspectrophotometry confirms that the spectral sensitivity curves of individual cones are smooth and U-shaped, and these particular data suggest absorption maxima at about 420, 530, and 560 for the three cone types. Microspectrophotometry, however, is beset with signal/noise problems that limit the measurements to a relatively narrow range of wavelengths around the spectral maximum (notice the limited wavelength range of the squares in Figure 7.6B).

In the mid-1980’s, the problem of cone spectral sensitivities was attacked with suction electrodes. This technique has the advantage that the physiological response of the photoreceptor itself is used for the actual measurements. Since very small amounts of light are sufficient to produce measurable changes in photocurrents, the suction electrode produced an increase in sensitivity over earlier techniques, with a corresponding increase in the wavelength range over which meaningful measurements could be made. Data from an early suction electrode study are compared with those of earlier techniques in Figure 7.6B. The suction electrode data, plotted on a linear sensitivity axis,

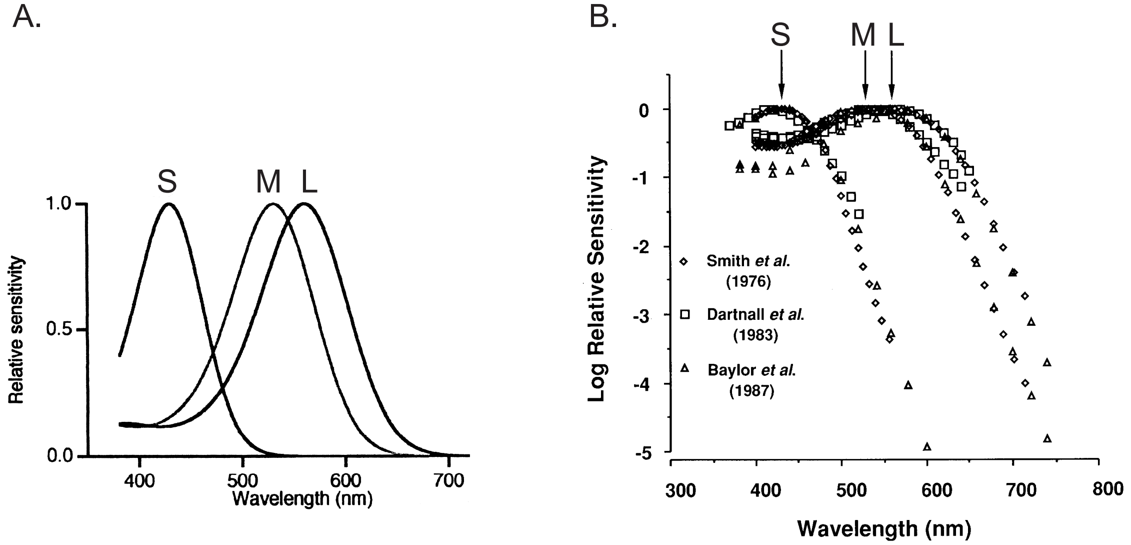


Figure 7.6: The spectral sensitivities of the three cone types in color-normal subjects. A. L, M and S cone spectra from macaque retina, recorded with suction electrodes, shown on a linear ordinate. B. The three cone spectra estimated with three different techniques, shown on a log ordinate. The results of psychophysical (diamonds), microspectrophotometric (squares), and suction electrode (triangles) techniques are shown. The three sets of measurements agree remarkably well. [A From Baylor (1987, replotted courtesy of J. Schnapf); B From Lennie and D'Zmura (1988, Fig. 10, p. 344).]

are shown in Figure 7.6A. These data replace the conceptual Fundamentals introduced in Figure 7.5A.

All of these techniques have evolved over the authors' scientific lifetime, and each decade has brought a surer answer. Nowadays there is excellent agreement from all of these very different kinds of measurements, and the cone spectral sensitivity curves are a solved problem. The most recent estimates suggest that the spectral maxima of the S, M, and L cones are very close to 430, 530, and 560 nm. Moreover, as needed to sustain the three channel mathematical model developed above, the three spectral curves are broadly overlapping, at least between 400 and 550 nm, so that in this wavelength range, a single wavelength can set up a quantum catch rate in each of the three kinds of photoreceptors, and set up the ensemble code.

Notice, however, that the S cones have negligible sensitivity above about 550 nm. As a result, color-normal subjects actually have only two functional cone types in the mid to long wavelength spectral range. And in fact, color-normal subjects can match any wavelength above about 550 nm with a mixture of only two primaries, such as a 530 nm "green" and a 650 nm "red".

## 7.4 Linking theory for cones and trichromacy

Let us consider where we've been. First, psychophysical measurements (color matching) on human subjects quantified the puzzling system properties of metamerism and trichromacy. Second, a mathematical model was developed to explain these system properties. The model – three Fundamentals with broadly overlapping spectra and linear summation of effects across wavelengths, expressed mathematically in three simultaneous linear equations with three unknowns – provides a sufficient account of the psychophysical data.

But the model also provided specific suggestions about neural elements one should find within the visual system: three kinds of cones with different spectral maxima but overlapping spectral sensitivity curves. We then went looking for independent evidence for the existence and spectral properties of three kinds of cone photoreceptors, and found it. Moreover, the linear summation across wavelengths assumed to occur within each Fundamental finds its explanation in the loss of wavelength information in the transduction process. The linking theory between the cones and trichromacy is quite complete, and tightly tied into the network of surrounding sciences.

This linking theory between the cones and trichromacy builds on the linking theory for rods and wavelength encoding already described for scotopic vision in Chapter 6. In that chapter, we found a close correspondence between metamer sets for behavior and for the rods. On can think of this theory as linking the monochromacy of scotopic vision to the univariant response of a rod to different wavelengths. In this chapter, we linked the trichromacy of photopic vision to the univariant responses of the three cone types. In both cases, the selectivity of wavelength encoding was used to link behavioral effects to neural elements.

The story of trichromacy also illustrates the difference between a mathematical and a physiological model. Mathematical modeling is challenging, and it's a major achievement to invent a model that exactly accounts for a set of psychophysical findings. But given a satisfying mathematical model, the question arises: will the model be instantiated in the physiological visual system? Vision scientists become keenly interested in finding physiological entities that embody the mathematical entities or parameters assumed by such a model. This is the search for the "link" in linking theory.

For the authors, it's thrilling to understand that three simultaneous equations in three unknowns provide an account of trichromacy. But it's even more thrilling to learn that the hypothesized discarding of wavelength information within a Fundamental corresponds to the actual discarding of wavelength information by the transduction process, and that the mathematical property of adding up the terms in the linear equations across wavelength corresponds to the physiological property of combining the effects of individual cis-trans isomerizations across wavelength.

In sum, the linking theory of trichromacy is vision science at its very best. Trichromacy and its explanation provide an important example to which to aspire as one tries to invent linking theories to explain other system properties of vision.

## 7.5 Why three and only three cone types?

A design question: Why did humans evolve to have exactly three cone types rather than two or four? It has been argued that ancestral primates had only two types – an S cone and a prototypical LM cone with a spectral maximum in the mid to long wavelength region. But since the S cones have negligible sensitivity above 550 nm, the ancestral primate would have had only a single pigment available above 550 nm, and would not have been able to discriminate among middle and long wavelengths, nor among surfaces that reflect different combinations of middle and long wavelengths. Roughly speaking, the ancestral primate would not have been able to discriminate among objects or surfaces that we perceive as yellow-greens, yellows, oranges and reds.

The splitting of the LM prototype into two separate classes – L and M – probably occurred about 30-40 million years ago in the old-world primates from which humans evolved[4]. It has been proposed that such a trichromatic system allows the discrimination of red, orange, and yellow fruit from green trees, and thus to tell ripe from unripe fruit. Consequently, the third cone type allowed ancestral primates to exploit an important new food source, and probably carried a selective advantage in evolutionary terms.

Why not keep on evolving, and have more than three cone types? After all, the larger the number of cone types, the smaller the metamer sets, and the more wavelength information is preserved. Considering the entire range of vertebrate visual systems, some have one, two, three and even four cone pigments (Bowmaker, 2008). The speculation here is that, although it is easy to create trichromatic metamers in the lab, they rarely occur in nature. In nature most surfaces reflect broad bands of wavelengths and most light sources produce broad bands of wavelengths. Thus, most of the objects that produce lights in any one metamer set probably have relatively similar spectral characteristics (Marimont and Wandell, 1992). So additional photopigments might not allow us much more useful color discriminations than we can already make with three.

## 7.6 Color vision deficiencies: Dichromacies and anomalies

People whose retinas contain the three standard pigments are said to have normal color vision, or to be *color-normal trichromats*. But not everyone is color normal. Some people are missing one of the three kinds of cones. Others have three cone types, but one or more of the pigments is shifted in spectral sensitivity. Such changes make predictable changes in color matching and discrimination. Look back at Figure 7.5 as you read the next few paragraphs.

First, what would happen if you were missing one of the three cone photopigments? Suppose you were missing the L photopigment. In that case, in Figure 7.5 you would have two rather than three funnels, and two rather than three cone output signals. You would have only two rather than three equations in your set of color equations, because the equation for quantum catches in the L cones would not be needed. Since there would be only two equations, you would need only two wavelengths in patch A to match any wavelength in patch B (two equations need only two unknowns to be guaranteed a solution). All of the metamer sets of a trichromatic individual would also be metamer sets for you; but your metamer sets would be larger than those of the trichromat – you would confuse patches of light that would be readily discriminable for your trichromatic friend, and you would probably be worse than they are at finding yellow, orange and red fruit in green trees.

This form of color vision deficiency is well known, and occurs quite frequently in human beings. Since the color vision system is reduced from three to two variables, such individuals are called *dichromats* (di = two). The two most common types are *protanopia*, in which the person is missing functional L cones (proto = first; protanopia = the first kind of color deficiency); and *deuteranopia*, in which the person is missing functional M cones (deutero = second; deuteranopia = the second kind). Each of these types of color vision deficiency is sex-linked, and occurs in about 1% of the Caucasian male population. The third kind of dichromacy, *tritanopia* (tri = three; tritanope = the third kind), in which the person is missing functional S cones, is much rarer, and occurs with equal

[4] There was also an independent development of trichromacy in the new-world howler monkeys (Hunt, Dulai, Cowing, Julliot, Mollon, Bowmaker, Li, and Hewett-Emmertt, 1998).

frequency in both males and females.

Second, what would happen if the spectral sensitivity of one of your photopigments were shifted along the wavelength axis? Suppose your L cone pigment were shifted toward your M cone pigment. How would your color equations change? Since the height of the L curve would have changed a little at each wavelength, all of the little l's in the equation for the quantum catch in L cones would change. You would still be trichromatic, because your color vision would still be described by three equations in three unknowns. But for each wavelength composition of patch B, the change in values of $l$'s would make a change in the intensities of the three lights in patch A needed to make the match to patch B. That is, your color matches would be different than those of your color-normal friend. Similar changes would occur if the M or the S cone spectral sensitivity curve were shifted.

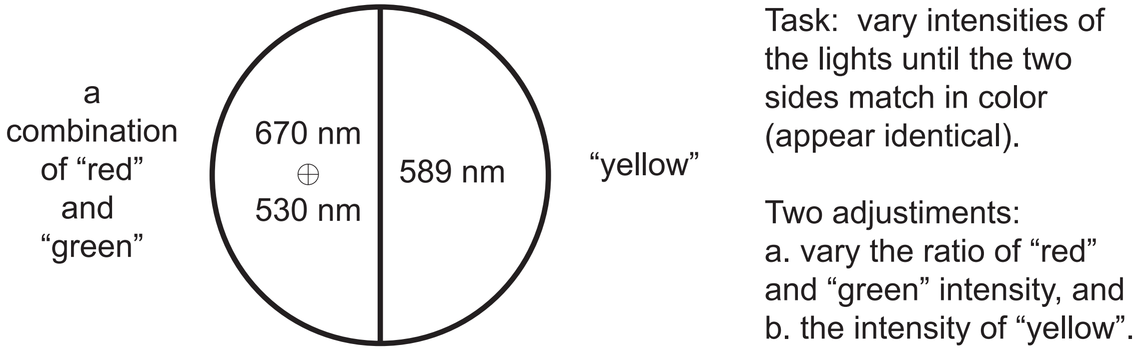
This form of color vision deficiency is also well known, and people with trichromatic vision but non-normal metamer sets are said to be *color anomalous*. The color vision of people with a shifted L cone pigment is called *protanomalous*, while that of people with a shifted M cone pigment is called *deuteranomalous*. Protanomaly and deuteranomaly occur in about 1% and 3% of the Caucasian male population respectively. *Tritanomalous* color vision, which results from a shifted S cone pigment, is much rarer and is difficult to discriminate from an intrusion of rods.

In sum, in the Caucasian population about 8% (one in 12) of the male population and less than 1% of the female population have a color deficiency caused by losses or spectral shifts of either the L or the M cone photopigment. As a group, these forms of color vision are often called the *red/green color deficiencies*. Another small percentage (less than 1%) have *tritan* deficiencies – losses or anomalies of the S cone photopigment.

Color-normal and color-deficient individuals live in different perceptual worlds. To illustrate this point, we here discuss a clinical color mixture test that diagnoses among color-normal, dichromatic and anomalous trichromatic subjects. The test is the *Rayleigh match*, carried out with a device called an *anomaloscope*. The stimulus for the test is illustrated in Figure 7.7. In an anomaloscope, the subject sees a mixture of 550 and 670 nm lights (which ordinarily look green and red respectively to a color-normal subject) in one half of a circular field, and a 589 nm light (which ordinarily looks roughly yellow to a color-normal subject) is presented in the other half of the field. The subject is asked to vary the proportion of the 550 vs. 670 nm lights in the one half-field, and the intensity of the 589 nm light in the other half field, to make a metameric match between the two halves of the field.

For a color normal subject, there will be a particular ratio of intensities of the 550 and 670 lights (say, 50:50) that is metameric to the 589 nm light. Both halves of the anomaloscope field will look roughly yellow. For a dichromat, who cannot discriminate the 550 from the 670 light in the first place, the 589 nm field can be matched by any ratio of the 550 and 670 nm fields, including 100% of either of these wavelengths. In other words, lights that look red, yellow and green to color-normal subjects all look the same – are in the same metamer set – for a dichromat. For an anomalous trichromat, the normal trichromat's metamers look different colors, but the 550 and 670 lights can be mixed in some *other* proportion (say 70:30) to match the 589 nm light. These half fields that match for the anomalous trichromat look different to the color normal subject.

Red/green color deficiencies are so common that in any class of 30 school children, one or more of the boys is likely to have a red/green color deficiency. These children are likely to find it more difficult to learn color names, use the "correct" color in drawing with crayons, and so on. An adult color-deficient individual can have trouble choosing two socks that match, and may wear color combinations that seem bizarre to his color-normal friends. If you argue with your friends over the



Task: vary intensities of the lights until the two sides match in color (appear identical).

Two adjustiments:
a. vary the ratio of “red” and “green” intensity, and
b. the intensity of “yellow”.

Figure 7.7: Rayleigh matches. An illustration of the spatial layout of the visual field in an anomaloscope. On the left side, the field is composed of a combination of 530 and a 670 nm primaries. On the right side, it is composed of a single 589 nm primary.

color of things, you may be a dichromat or an anomalous trichromat and not realize it.

## 7.7 Genetics of color vision

In 1986, a team of geneticists and psychophysicists led by Jeremy Nathans isolated and sequenced the genes that control the production of each of the three human cone photopigments. The DNA encoding was characterized for both color-normal and red-green color-deficient individuals. This research was extremely exciting to vision scientists, since it took the search for the Fundamentals of color vision all the way to the molecular level. A review can be found in Neitz and Neitz (2011).

Each of the red/green color deficiencies described above shows an X-linked pattern of inheritance (the particular deficiency is passed from grandfather to grandson with the mother being a carrier). The L and M pigment genes are located on the X chromosome. Tritanopia shows an autosomal pattern of inheritance with the S pigment gene located on chromosome 7 (Nathans, Piantanida, Eddy, Shows, and Hogness, 1986).

As expected from the X-linked inheritance patterns of red/green color vision deficiencies, the genes for two photopigments were found next to each other on the X chromosome. Unexpectedly, many individuals had several rather than just one copy of the second gene in the sequence. Protanopes turned out to be missing the first gene of the sequence, which was therefore identified as the L cone pigment gene. Deuteranopes often had only the first gene of the sequence, and the second and later genes were therefore characterized as the M cone pigment genes. More complex genetic patterns – genes made up of pieces from both the L and the M pigment genes – were also found, especially in color-anomalous individuals. Moreover, polymorphisms were found in the normal L and M pigment genes, allowing an explanation of more subtle variations of color vision among color-normal individuals[5].

This work gives a new path to study the linking theory between trichromacy and the photopigments. Simple blood tests can be used to identify the genes that specify the photopigments in an

[5] Geneticists (and the rest of us) sometimes slip into talking as though complex human behaviors can be attributed to single genes. A deuteranopic student in one of Teller’s classes wrote a term paper on the genetics of color vision deficiencies. In parody of the single-gene assumption, his opening sentence was, “I have a gene for mismatched socks”.

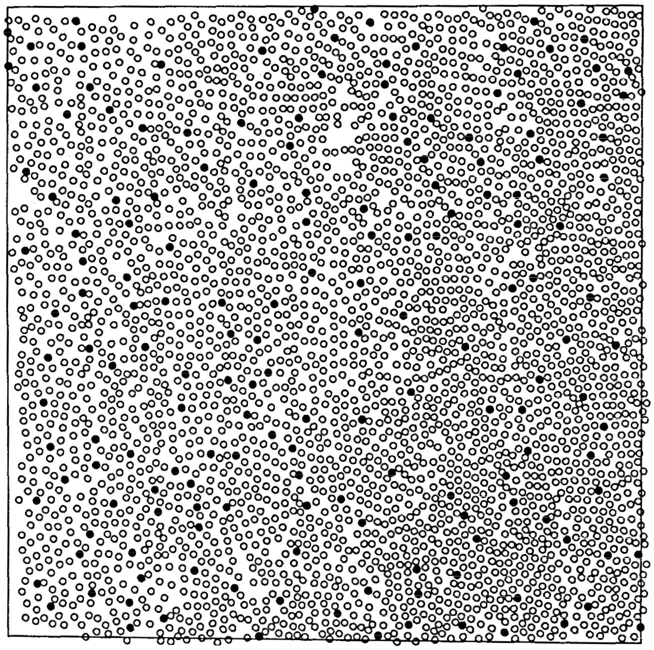


Figure 7.8: The matrix of S cones. The picture shows the distribution of S cones over a small region just off from the center of the fovea (fovea is to the lower right). The black dots are S cones; the open dots are L and M cones. [From Curcio et al. (1991, Fig. 10, p. 620).

individual. These can be compared with the spectral sensitivity estimated by color matching and other behavioral measures to test the correspondence between trichromacy and the details of the photopigments.

## 7.8 Spatial mosaics for the S, M and L cones

The numbers and distributions of rods and cones across the retina, shown in Figure 6.2, have been known for over 75 years. But what are the numbers and distributions of each of the three cone types?

These questions have been of major theoretical interest, but the answers proved elusive for many years. The S cone mosaic has been of interest because acuity is poor under conditions that isolate S cones, and it has therefore been speculated that there might be only a small number of S cones. The L and M cone mosaics have been of interest because there are individual differences in photopic spectral sensitivity curves, and it has been suspected that these might be due to individual differences in the proportions of L vs. M cones: the *L/M cone ratio.*

Within the last 20 years or so, the numbers and distribution of S cones has been determined with several different techniques. In the most definitive early study, Christine Curcio and her colleagues (Curcio, Allen, Sloan, Lerea, Hurley, Klock, and Milam, 1991) used a newly developed stain specialized to reveal the S cone opsin. Use of this stain exposes the whole S-cone matrix.

A sample of Curcio et al's results are shown in Figure 7.8. In their data, S cones provide only about 10% of the cones in the human retina. In fact, although it does not show in the figure, S

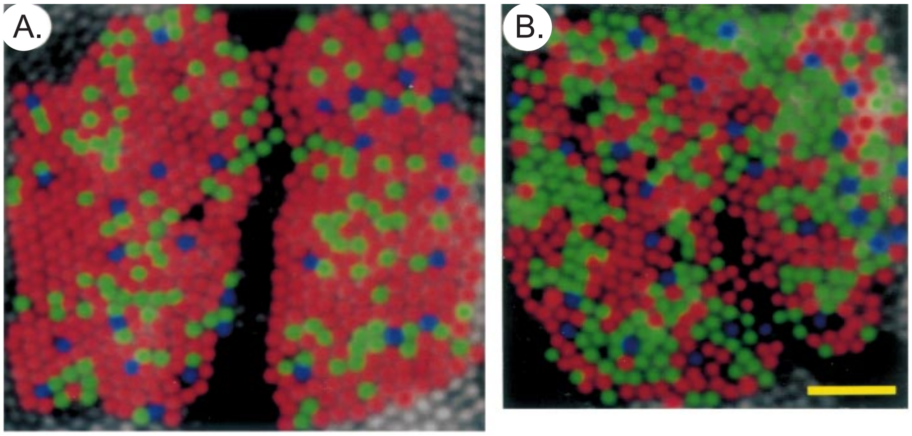


Figure 7.9: Cone mosaics in two color-normal subjects. A. The pseudocolor image of the retina for subject JW. B. The pseudocolor image of the retina for a second subject, AN, whose photopic spectral sensitivity curve was elevated in the mid wavelength region, suggesting that he might have a higher than average proportion of M cones. He does. [Roorda and Williams (1999, Fig. 3, p. 522).]

cones are usually missing entirely from the central fovea; and they are spaced relatively far apart throughout the rest of the retina.

Why? The suspected explanation for this sparse representation is that chromatic aberration defocusses the retinal image formed from short wavelength light, so that fine spatial sampling would be wasted in the S cone system. The S cones provide another example of the idea that "poor" optics – in this case chronically defocussed images, due to chromatic aberration – limit our acuity for short wavelength light, and do so through an evolutionary mechanism that matches the spacing of the photoreceptor matrix to the quality of the optical image.

The numbers and distributions of L and M cones were next attacked with the techniques of adaptive optics. In our discussion in Chapter 4 we suggested that one of the major uses of adaptive optics will be to *look in* through a person's corrected optics and be able to see the structures in his living retina. Omitting many details, one can expose the same patch of retina to lights with three different combinations of wavelength and use the contrast between the images to estimate which cones are L, M, or S. The results are shown in Figure 7.9 with pseudocolor images of the retinas of two different subjects. The proportions of S cones were about 5% in both retinas. But the L/M cone ratios differed considerably: 3.8 for JW vs. 1.2 for AN. Measurements of larger samples of subjects confirm the existence of this large range of ratios (Carroll et al., 2002). We will return to these individual differences in the following section on photopic spectral sensitivity.

More recently, adaptive optics have yielded answers to some of the classic questions concerning dichromacy. It has long been speculated that the loss of a photopigment gene could lead to the functional or actual loss of a type of cone: the L cones for protanopes, the M cones for deuteranopes, and the S cones for tritanopes. The images shown in Figure 7.10, taken with adaptive optics in the living eyes of two dichromats, show that this speculation is correct. The retina of the deuteranope shown in Figure 7.10A is missing its M cones, and the retina of the protanope shown in Figure

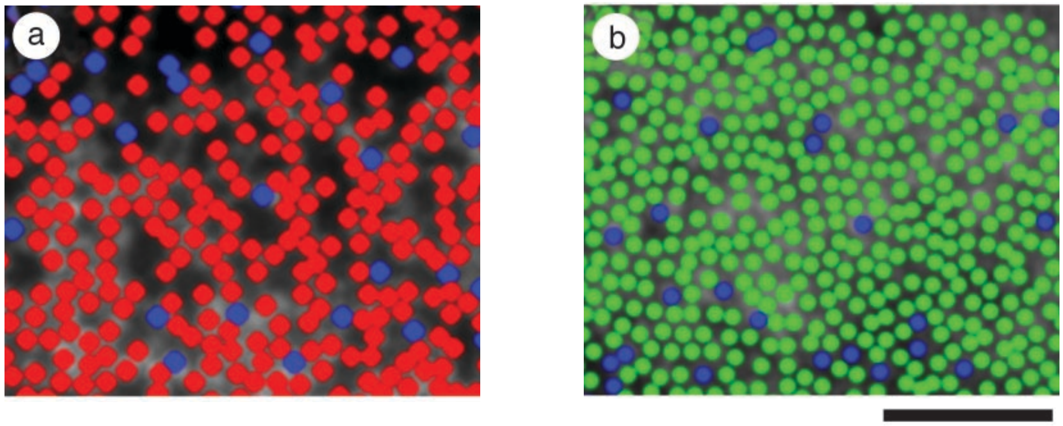


Figure 7.10: Cone mosaics in two dichromats. A. Deuteranope NC, missing M cones. B. Protanope MM, missing L cones. NC has a "patchy" retina, with some cones apparently missing, whereas MM appears to have the full complement of cones. The scale bar at the lower right is for 50 $\mu$m. [From Carroll, Neitz, Hofer, Neitz, and Williams (2004, Fig. 4, p. 8465).]

7.10B is missing its L cones.

These images also address a second question. There have been two rival theories for how the loss of a pigment plays itself out at the level of disabling the photoreceptors that would have contained that pigment. *Loss theories* suggest that the cones that would have contained the missing pigment are literally lost from the retina, leaving holes in the mosaic where the missing class of photoreceptors would have been. On the other hand, *replacement theories* suggest that the cones that would have contained the missing pigment are filled with a different pigment – for example, that a protanope would have its potential L and M cones both filled with the M-cone pigment – so that the full complement of cones is retained in the retinal mosaic.

It turns out that both theories are probably right. In studies of the molecular genetics of color vision, two different kinds of genetic changes have been found in different dichromats. A dichromat can be missing the gene for the L or M cone pigment. Alternatively, he can have a mutation that makes the L or M cone pigment misshapen and therefore nonfunctional. Perhaps the absence of a gene leads to replacement, and a misshapen pigment leads to loss.

These arguments are supported by the images in Figure 7.10. Two subjects are shown: NC, who is a deuteranope because of a mutant M pigment gene, and MM, who is a protanope because of a missing L pigment gene. The retina of the deuteranope NC is shown in Figure 7.10A. It shows L cones but no M cones, and a reduced overall number of cones, with "holes" between them, as predicted by loss theory. In contrast, the retina of the protanope MM is shown in Figure 7.10B. It shows M cones but no L cones, but a normal number of cones overall. Apparently the L cones have been filled with M cone pigment, and retained in the retinal mosaic, as predicted by replacement theory. Thus ends a controversy that lasted a century – loss of a cone type occurs in some dichromats, and replacement of the missing photopigment by the available one occurs in others.

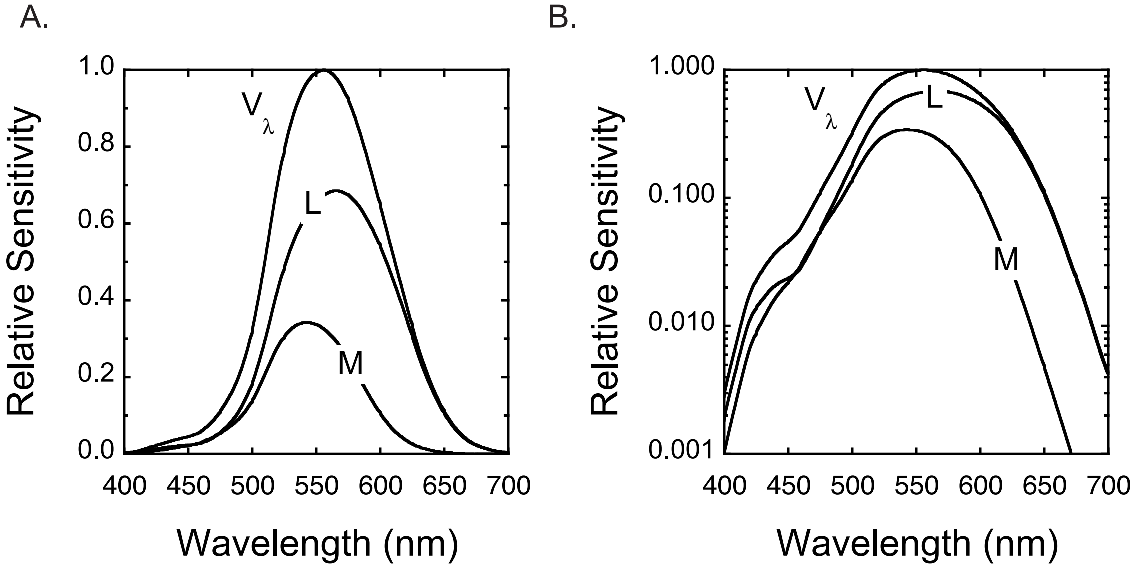


Figure 7.11: A model of $V(\lambda)$ based on a weighted sum of L and M cone inputs. Panel A shows the calculation on a linear ordinate, and Panel B shows it on a logarithmically scaled ordinate. The curves labeled L and M show the spectral sensitivities of the L and M cones; the curves labeled $V(\lambda)$ show the synthesis of $V(\lambda)$ from the sum of the L and M curves. For the average color-normal subject, the weighting needed to fit $V(\lambda)$ is about 2:1 for the L vs. M cone signals; that is, $V(\lambda)$ = 2L + M. The differential weighting is incorporated into the diagram by increasing the height of the L curve with respect to the M curve.

## 7.9 Photopic spectral sensitivity

Knowing the spectral sensitivities of the three Fundamentals provides us with several theoretical bonuses. First, in Chapter 3 we introduced the photopic spectral sensitivity curve, $V(\lambda)$. Now that the spectral sensitivity curves of the actual L and M cones are available for use, it turns out that $V(\lambda)$ can be readily modeled by a weighted linear sum of L and M cone signals. This idea is illustrated in Figure 7.11, and the physiological model fits the psychophysical data well.

Second, we also mentioned in Chapter 3 that although vision scientists have adopted a standard curve for photopic spectral sensitivity, there are actually small but consistent individual differences in the empirical photopic spectral sensitivity curves for different individual subjects. It has long been speculated that these individual differences could come about from variations in L/M cone ratios among subjects. Inspection of Figure 7.11 reveals that independent sliding of the L and M spectral sensitivity curves up and down will allow interesting changes in the overall photopic curve, and this model provides reasonable fits to the known individual differences.

Moreover, we have had a chance to examine the cone mosaics of two color-normal subjects, JW and AN, in Figure 7.9. The L/M cone ratios differed markedly between these two retinas. The photopic spectral sensitivity curves of these two subjects were also measured, and the differences are in the right direction to be modeled by the differences in L/M cone ratio.

And third, in Chapter 3 we also described the fact that $V(\lambda)$ emerges from many different kinds of psychophysical experiments – flicker photometry, motion photometry and minimally distinct

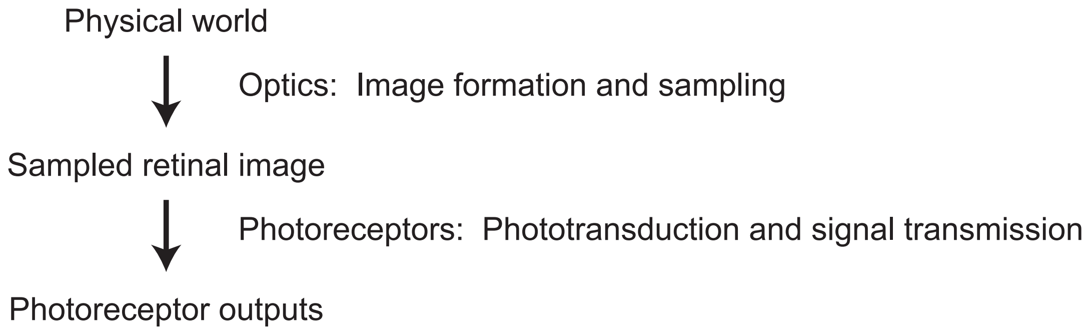

Physical world
Optics: Image formation and sampling
Sampled retinal image
Photoreceptors: Phototransduction and signal transmission
Photoreceptor outputs

Figure 7.12: Adding the photoreceptor transformation.

border judgments, among others. We argued that when a characteristic psychophysical "signature" emerges frequently from the data, it's a good guess that that characteristic has a physiological instantiation. That is, vision scientists would be drawn to speculate that individual neurons that sum inputs from L and M cones, with spectral sensitivity curves corresponding to $V(\lambda)$, will be formed within the visual system. However, neurons that embody this prediction are simply not present at the level of the photoreceptors. We are left to speculate that they will emerge at a later level of processing.

## 7.10 Summary: The photoreceptor transformation

At the end of Chapter 5 we summarized the effects of the transformation from the physical world to the retinal image. This optical transformation rendered the retinal image *two-dimensional* and *low pass filtered*.

In addition to the optics, the incoming visual signal also encounters a stage of discrete sampling by the photoreceptors that rendered it (poetically) *pointillistic*. The discrete sampling stage could be considered either part of optical processing (since the signal is still carried by light), or part of processing by the photoreceptors (since they are the spatially discrete elements). As such, it could be considered part of the effect of having photoreceptors. Following the first of these choices, we summarized sampling along with the optical transformations at the end of Chapter 5.

We are now ready to summarize the transformations due to the photoreceptors: this transformation is from the retinal image to the quantum catches in photoreceptors via the phototransduction process, and on to the photoreceptor outputs via a complex set of chemical and electrical information transmission processes. These two stages of processing are summarized in Figure 7.12. In combination, these two processes create a spatial array of quantal catches in the rods and L, M and S cones, and transform it into a spatial array of synaptic outputs from these four kinds of photoreceptors.

The transduction process and the combination of four kinds of photoreceptors provide us with a linking theory for some of the system properties of scotopic and photopic vision introduced in Chapters 2 and 3, as well as for the trichromacy of color vision discussed in the present chapter. The scotopic spectral sensitivity curve is maximal at about 500 nm, and wavelength information is lost in scotopic vision, because scotopic vision is served by rods and rods alone. Wavelength information is preserved in photopic vision, up to the limits described by trichromacy, because photopic vision is served by three and only three kinds of cones.

The information transmission process also leaves its mark. For example, at low light levels, each rod is so exquisitely sensitive that it produces a detectable signal in response to the absorption of a single quantum. This sensitivity enables human subjects to detect the absorption of only 5-10 quanta in an extended test field, and therefore of a single quantum in an individual rod. In addition, a saturating non-linearity, probably within the cone photoreceptors, provides a signal that allows the detection of interference fringes in the vicinity of 60 cy/deg. Of course, the properties of photoreceptors influence all aspects of vision, but additional examples are beyond our scope.

Information processing by the three cone types together also illustrates the concept of *ensemble codes* (or *pattern codes*). Because of the nature of transduction, each photoreceptor individually loses wavelength information. Yet, working together as an ensemble, the three types of photoreceptors preserve at least some information about the wavelength composition of each region of the retinal image. By comparing the signals from L, M and S cones, later levels of the system have access to a fair bit of wavelength information. The concept of a pattern code will recur frequently throughout the remainder of this book. Later we will see neurons that compare signals from the L, M and S cones, to create a new wavelength/color code.