ON THE USE OF MLDS IN THE STUDY OF DEPTH AND LIGHTNESS PERCEPTION

vorgelegt von
Guillermo Andres Aguilar Cornejo, M.Sc.
aus Santiago, Chile

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Klaus Obermayer
Gutachterin:	Frau Dr. Marianne Maertens
Gutachter:	Prof. Dr. Marc Alexa
Gutachter:	Prof. Dr. Felix Wichmann
Gutachter:	Prof. Dr. Kenneth Knoblauch

Tag der wissenschaftlichen Aussprache: 4. September 2017

Berlin 2017

To my grandmother,
who was always there.

1923 – 2015

ABSTRACT

An open question in vision research is how to measure the perceptual dimension evoked by the stimulus in a reliable way. Although a variety of psychophysical procedures are available, it is still a challenge to find methods that are efficient and avoid critical confounds, such as strategies triggered by difficult and unnatural tasks used by discrimination methods. In this doctoral thesis I propose the use of Maximum Likelihood Difference Scaling (MLDS, Maloney & Yang, 2003) as a reliable tool for measuring perception. MLDS is a method based on judgments of appearance of clearly visible stimulus differences in an easy and intuitive task, and it allows the estimation of perceptual scales in an efficient way.

Here I first use numerical simulations to test the accuracy and precision of the scales derived with MLDS, and I also tested the effect of violations of the model assumptions. The results of these simulations establish the validity of MLDS as a method for measuring appearance. Then, we evaluated MLDS experimentally in the domain of lightness perception. We measured perceptual lightness scales under different viewing conditions and we validate the derived scales empirically by predicting lightness matches that were derived in a classical asymmetric matching task. A large practical benefit of MLDS is that it renders the task easy for the subject and thus minimizing the potential influence of strategies. At the same time the perceptual scales provide a more direct estimate of internal variables against which theoretical models of appearance can be tested.

In a third part I study the relationship between MLDS and discrimination methods as suggested by Devinck & Knoblauch (2012). In simulations MLDS was more efficient than the traditional 2-AFC discrimination method while at the same time providing analogous sensitivity estimates. I also tested this equivalence experimentally in a slant-from-texture task, for which sensitivity has been previously studied in the literature. Here I found varying degrees of equivalence and it remains to be tested in the future whether these differences are due to

true differences in the perceptual representation, or to violations of the model assumptions.

Together with the use of realistic stimuli, MLDS offers a reliable method to measure the perceptual dimension, and in that way enabling the testing of theoretical models of perceptual inference.

ZUSAMMENFASSUNG

Eine offene Frage in der visuellen Wahrnehmungsforschung ist, wie sich eine durch einen Stimulus evozierte Wahrnehmungsdimension reliabel messen lässt. Trotz einer Auswahl existierender psychophysischer Methoden, bleibt es eine Herausforderung effiziente Methoden zu finden, die kritische Konfundierungen verhindern, wie es zum Beispiel bei Diskriminationsaufgaben mit schwierigen und unnatürlichen Aufgaben der Fall sein kann. In dieser Dissertation stelle ich *Maximum Likelihood Difference Scaling* (MLDS, Maloney & Yang, 2003) als eine reliable Methode zur Messung von Wahrnehmungseindrücken vor. MLDS basiert auf der Bewertung deutlich sichtbarer Stimulusunterschiede in einer intuitiven und einfachen Aufgabe und ermöglicht die effiziente und reliable Schätzung perzeptueller Skalen.

In einem ersten Schritt wird MLDS als eine zuverlässig Methode zur Messung von Wahrnehmunseindrücken etabliert, indem ihre Genauigkeit und Präzision bestimmt wird sowie Verletzungen von Modellannahmen numerisch simuliert werden. In einem nächsten Schritt wird MLDS im Bereich der Helligkeitswahrnehmung experimentell evaluiert, indem gezeigt wird, dass MLDS erfolgreich Wahrnehmungsskalen für verschiedene visuelle Kontextbedingungen bestimmt. Die gemessenen Skalen weisen weitgehend Helligkeitskonstanz auf. MLDS erforderte dafür den Vergleich von Stimuli innerhalb eines visuellen Kontexts, was zu einer Vereinfachung der Aufgabe für die Versuchsperson sowie zur Vermeidung von Problemen geführt hat, welche mit Vergleichen über visuelle Kontextbedingungen hinweg in Verbindung gebracht werden. Zusätzlich hierzu, schien MLDS der Methode des asymmetrischen Vergleichs für die Bestimmung von Hel-

ligkeitskonstanz überlegen, da diese im Gegensatz zu MLDS nur ein indirektes Maß liefern kann und Wahrnehmungsskalen nicht direkt misst.

Der Zusammenhang zwischen MLDS und Diskriminanzverfahren wurde auch im Rahmen der Signalentdeckungstheorie untersucht, wie zuletzt von Devinck & Knoblauch (2012) vorgeschlagen. Simulationen haben gezeigt, dass MLDS effizienter ist und in der Sensitivitätsschätzung quantitativ ähnlich zu traditionellen *2-AFC* Diskriminanzverfahren, jedoch nur wenn alle Modellannahmen erfüllt waren. Diese Äquivalenz wurde zudem experimentell in einer *slant-from-texture* Aufgabe getestet, für die Sensitivitätsmaße bereits untersucht wurden. Ich fand unterschiedliche Abstufungen von Übereinstimmung, die entweder tatsächliche Unterschiede in der perzeptuellen Repräsentation oder eine Verletzung der Modellannahmen darstellen können.

Zusammen mit der Verwendung realistischer Stimuli bietet MLDS eine reliablere Methode zur Messung perzeptueller Dimensionen und ermöglicht so die Testung theoretischer Modelle perzeptueller Inferenz.

PUBLICATIONS

Parts of this thesis have been published in:

Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision, 17*(1):37, 1-18. doi:10.1167/17.1.37

Wiebel C.B.*, Aguilar G.*, Maertens M. (2017). Maximum likelihood difference scales represent perceptual magnitudes and predict appearance matches. *Journal of Vision, 17*(4):1, 1-14. doi:10.1167/17.4.1
*: *equal contribution*

Wichmann F. A., Janssen D. H. J., Geirhos R., Aguilar G., Schütt H. H., Maertens M., & Bethge M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, 2017*(14): 36-45. doi:10.2352/ISSN.2470-1173.2017.14.HVEI-113

Wahrlich es ist nicht das Wissen, sondern das Lernen,
nicht das Besitzen sondern das Erwerben,
nicht das Da-Seyn, sondern das Hinkommen,
was den grössten Genuss gewährt

It is not knowledge, but the act of learning,
not possession but the act of getting there,
which grants the greatest enjoyment.

— Carl Friedrich Gauss

ACKNOWLEDGMENTS

First I want to thank my advisor Marianne Maertens. I'm really thankful for all the time, patience, ideas and feedback that she provided, making the years on my Ph.D. the best I've had. She has taught me how good psychophysics is done, and how to avoid the traps of the hypes. Her contagious critical thinking made all this work possible.

I want to also thank Felix Wichmann, for the stimulating discussions and his advice on both projects here exposed; in particular for his critical feedback regarding the work on MLDS and signal detection theory. I also thank my colleagues Christiane Wiebel and David Higgins. With Christiane we had a fruitful collaboration that lead us to publish our study in co-authorship. And with Christiane and David we shared, more than an office, memorable times of work and laughs together.

I also thank the numerous reviewers of the manuscripts, who provided insightful comments on them and consequently also helped improve this thesis: Michael Landy, Kenneth Knoblauch, Bart Anderson, Richard Murray, Frank Jäkel, David Brainard, Michael Kubovy, Laurence Maloney and one other anonymous reviewer. Also I thank the funding resources that allowed these years of work: the Graduate Research Training 'Sensory Computation and Neural Systems' (GRK 1589/1-2), and Marianne Maertens's project (DFG MA5127/1-1), both from the German Research Foundation (DFG).

Finally, I would like to thank my family back in Chile, and to Ba, my partner in crime, for his company, patience, and encouragement in these last years. And to the rest of my new family in Germany: Karin Ludwig, Dirk Warnick, Rafaela Wahl, and the Berlin Bruisers' family. They inspire me to continue in this life journey of learning and discovery.

CONTENTS

ACRONYMS

MLDS Maximum Likelihood Difference Scaling

JNDs Just-noticeable differences

GLM Generalized linear model

2-AFC two-alternative forced-choice

2-IFC two-interval forced-choice

Part I

BACKGROUND

1

MEASURING PERCEPTION

An open question in vision research is how the visual system constructs perception given the ambiguous information arriving to the retina. We normally have a stable and constant percept of the world's attributes, despite the changes in viewing contexts that produce a radical change in the actual stimulation impinging our sensory organs. This phenomenon is known as perceptual constancy, and it can be defined as the ability of a perceptual system to provide constant representation of stimulus attributes despite the significant change in viewing conditions or context. The human visual system is constant to object's attributes such as size and surface reflectance (i.e. color), among others.

Perceptual constancy

Figure 1.1 illustrates the issue in the domain of lightness perception. We consider an illumination source, like the sun, that emits a flux of light in all directions. We also consider an object, like the cube in Figure 1.1 where its surface reflects part of the light it receives. The *reflectance* of the object is the proportion of light reflected and it is a physical quality of the object's material. The flux of reflected light or *luminance*, measured in cd/m^2, arrives to the retina and it depends on a combination of both the illumination and the object reflectance. Multiple combinations of illumination and reflectance can produce the same luminance value; luminance is thus ambiguous with respect to both of its components. Luminance is the *proximal* stimulus and the product of the image formation process that occurs when the *distal* stimuli (i.e. illumination, reflectance, and eventually intervening media) are projected into the retina. *Lightness* is the counterpart in the perceptual dimension of the distal attribute, in this case lightness is the perceived surface reflectance, and it is the outcome of the process of perceptual inference.

The ambiguity of the retinal input can be best illustrated with the 'Adelson checker-shadow illusion' shown in Figure 1.2A. In this image two checks, one inside and one outside the shadow (in 'plain view') are identical in luminance

Adelson checkerboard

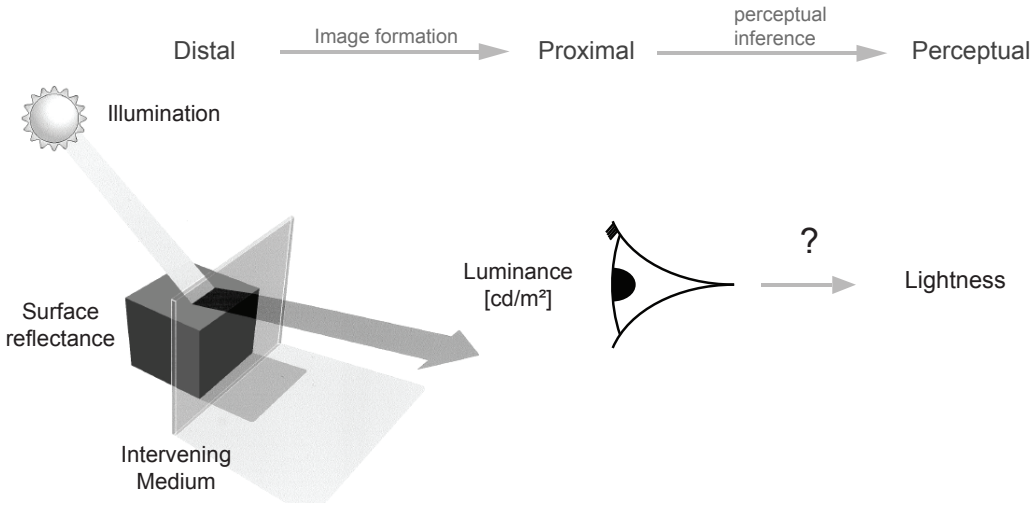


Figure 1.1: Distal, proximal and perceptual dimensions in the domain of lightness perception.

but different in reflectance. The retinal input (or grayvalue in the picture) is identical, however lightness follows closely the actual reflectance of the checks, the one in shadow perceived as lighter than the one outside of it.

To understand how the visual system can accomplish perceptual constancy it is necessary the reliable measurement of perception, i.e. the perceptual dimension evoked by a stimulus. After more than a century of psychophysical research it is still a challenge to do measurements in a reliable way. This challenge is twofold: choosing the type of stimuli to probe the visual system, and choosing the type of psychophysical method (Figure 1.3).

Stimulus naturalism

Stimuli can be broadly classified according to their complexity or naturalism, ranging from simple, well controlled but artificial stimuli, to more naturalistic and complex stimuli such as the Adelson's checker-shadow illusion. Simple stimuli are usually flat two-dimensional arrays, such as Gabor patches or sinusoidal gratings. In the case of lightness perception, examples are the simultaneous brightness contrast display, or disk-and-annulus stimuli (e.g. Wallach, 1948; Rudd & Zemach, 2007). Figure 1.2B shows the simultaneous brightness contrast display, where two equiluminant squares are embedded in surrounds of differ-

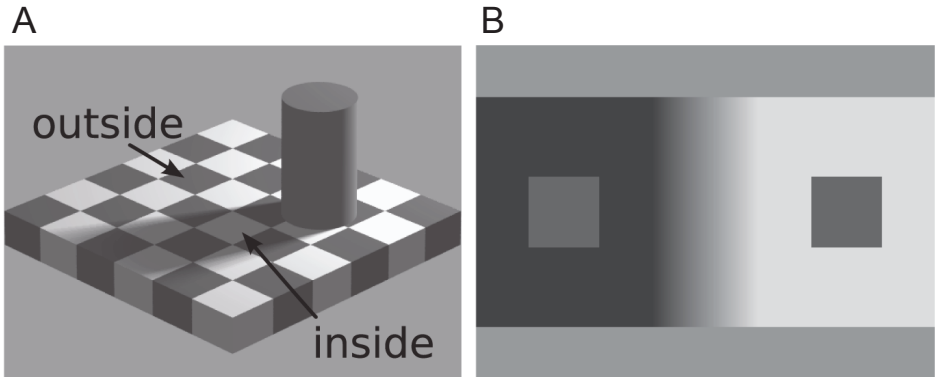


Figure 1.2: A. Adelson's checker-shadow illusion. The *luminance* of two checks (inside and outside the shadow, arrows) are identical, however they are perceived as having different surface reflectance, i.e they differ in *lightness*. B. Simultaneous brightness contrast stimulus. Two equiluminant squares are surrounded by backgrounds of different luminance, appearing different in lightness. The luminance values of the checks and their average surround are equivalent between panels A and B. Adapted from Maertens et al. (2015).

ent luminance. The squares are equiluminant and appear different in lightness: the square embedded in a dark surround appears lighter than the one embedded in a light surround.

The direction of the effect in the simultaneous brightness contrast display is similar to the Adelson checker-shadow illusion (panel A of same Figure). A critical difference, however, is that the simultaneous contrast display is composed by a flat two-dimensional array of luminance values, and thus it has not a clear distal source. Between panels A and B in Figure 1.2 the luminance values of the equiluminant checks and squares have been equated, as well as the surround where they are embedded. Maertens et al. (2015) used these stimuli and quantified the magnitude of the illusion effect, i.e. the lightness difference between the two equiluminant checks (or squares). They found that the effect is bigger for the Adelson's checker-shadow illusion than for the simultaneous brightness display, despite that the targets were equiluminant and they were surrounded by similar luminance configuration. They concluded that the difference in the effect

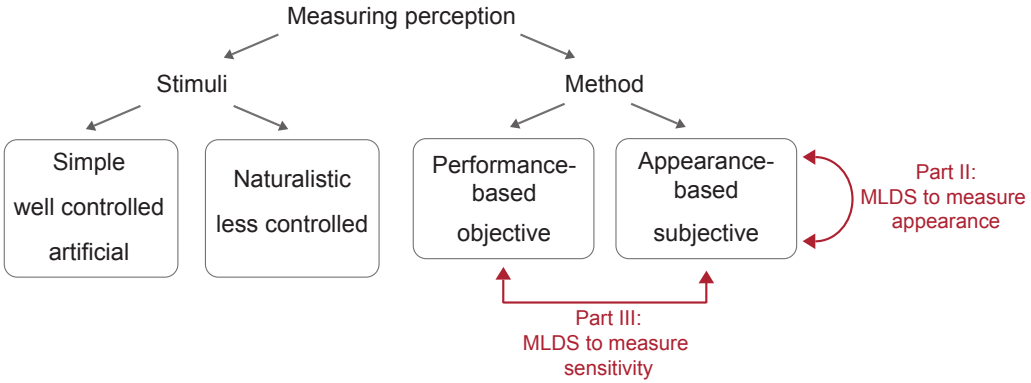


Figure 1.3: The challenges of measuring perception are two-fold: the choice of stimulus and the type of psychophysical method. Part II and III refer to the sections on this thesis.

magnitude is likely due to the fact that the Adelson's checker-shadow illusion is more realistic than the simultaneous brightness contrast, probing the visual system in a more appropriate way and this being reflected in a higher perceptual constancy.

In this thesis we chose stimuli of intermediate complexity and realism in the domain of lightness perception. Stimulus realism is not the main topic of this dissertation, but I will further explore and discuss the issue of how to incorporate more realism to stimuli at the end of this thesis (Chapter 13).

1.1 TWO TYPES OF METHODS

The second challenge in the measurement of perception is the choice of psychophysical method, and is the main focus of this thesis. Historically, there has been a division between two major approaches (Figure 1.3): the measurement of performance on stimulus discriminability, and the measurement of stimulus appearance (Kingdom & Prins, 2010). It is said that performance is measured when observers are asked to tell two stimuli apart, and appearance when observers are asked to report how stimuli 'look like'.

Performance methods

Performance methods estimate sensitivity to stimulus differences, or 'just noticeable differences', that usually requires the discrimination of stimuli that are very similar and close together in the stimulus continuum. In these type of tasks judgments are correct or incorrect, as defined by the actual physical stimulus attribute, and therefore a performance measure can be calculated.

Following a reductionist approach and drawing from the methods in classical Physics, performance methods have been extensively used in the study of the visual system. Performance-based methods focus on the measurement of sensitivity, i.e. thresholds, and use tasks that are usually difficult and unintuitive because they probe discrimination for small, near-threshold stimulus differences. Examples of these tasks are yes/no or forced-choice procedures. Commonly, performance-based methods use simple stimuli.

Although widely used, the reductionist approach in performance-based methods can be problematic. It is assumed that by studying the visual system with simple stimuli will eventually lead to explain the function of the visual system when confronted with naturally-occurring stimuli. In the natural environment the visual system is however always confronted with complex scenes that have clear distal sources and with a multitude of cues (as described above).

Appearance methods

Alternatively, appearance methods, such as scaling or matching methods, aim to measure how stimuli 'look like' by letting the observers adjust a probe that matches their perception (matching), or by asking them to judge set of stimuli that vary in some dimension of interest (scaling). In appearance methods there is no correct or incorrect response, as the whole point of measurement is to establish the mapping between the stimulus attribute and its perceptual dimension, i.e. a perceptual scale. An example of an appearance method is the method of paired comparisons (in Thurstonian scaling), where pairs of stimuli are presented and the observer judges their similarity along some perceptual dimension of interest. By analyzing the similarity judgments a perceptual scale can be constructed (Torgerson, 1958)

Traditionally these two different kind of measurement have had distinct procedures, tasks and statistical analysis tools. The division has been a source of controversy in the literature that expands until today (e.g. Luce & Krumhansl, 1988; Gescheider, 1997; Kingdom & Prins, 2010). Intuitively one would expect

that the ability to discriminate two stimuli has to depend on how they appear, because discrimination must rely on common mechanisms of perceptual representation from where also appearance judgments are drawn. Some researchers agree to this equivalency, while others argue that the two types of measurement evoke distinct perceptual mechanisms and therefore cannot be easily equated.

For some domains such as size perception, performance and appearance measurements differ significantly (Ross, 1997). In lightness perception, Whittle (1994) showed that sensitivity to luminance increments (or decrements) follow closely the judgments of appearance in a partition scaling task (reviewed also by Kingdom, 2016). The distinction between methods of appearance and discrimination is critical because models of perceptual inference provide different predictions if judgments are based on one or the other. Taking the results of Whittle (1994) into account, discrimination thresholds in the Adelson's checkerboard for increments (or decrements) measured on equiluminant checks – in shadow and plain view – should be different if they depend on perceived lightness. Contrarily, thresholds should be equal if they depend on luminance instead of lightness.

It is still an open question in the field which type of method, performance-based or appearance-based, is better at probing successfully the perceptual dimension. As suggested above, it would seem that appearance-based methods could be better because they can provide tasks that are not as difficult as discrimination of near-threshold stimulus differences. In this thesis I focus on the use of appearance-based methods for studying perception, as it will be discussed in the following.

1.2 MEASUREMENT PROBLEMS

As pointed out correctly by Runeson (1977) "the fact that subjects do judge a certain variable does not prove that they possess perceptual mechanisms of an appropriate kind. When the task does not fit the perceptual mechanisms we must expect the subject to try to compensate by using intellectual abilities, and such results will not be relevant to the study of perception". In fact, there have been reports that call the validity of some standard psychophysical measurement

methods into question. The following two examples illustrate that the problem that equally affects performance-based and appearance-based methods.

Ekroll and Faul (2013) asked observers to match a target color in an asymmetric matching experiment using simultaneous color contrast stimuli. The stimuli were conceived in a 3-D color space, but in order to accomplish satisfying matches observers resorted to a fourth transparency dimension. Thus, unintended by the experimenters, the dimensionality of the perceptual space exceeded that of the stimulus space (see also Logvinenko & Maloney, 2006). If transparency had not been included as an adjustable dimension, observers would have set unsatisfactory adjustments not matching in appearance, and the experimenter could have interpreted these results as a failure in constancy.

Todd, Christensen, and Guckes (2010), employing a discrimination procedure, measured apparent slant for textured surfaces that were slanted at different angles. The textured surfaces inevitably contain two dimensional (2-D) cues such as foreshortening or change of texture density that vary systematically with slant. It was assumed that the 2-D cues are used by the visual system to compute perceived slant, and that observers compare the stimuli with respect to perceived slant. This was however not the case; instead, the results from Todd et al. (2010) revealed that "observers' judgments were completely unaffected by whether or not the displays actually appeared slanted" and that observers were doing the judgments by directly comparing 2-D cues. Thus, these studies call for a revision of the current methods used to measure perceptual phenomena.

1.3 MOTIVATION

It is still an open question which methods are better for a reliable measurement of perception, methods that could avoid the aforementioned problems. A recently introduced appearance-based method seems more suitable and may provide a better choice: Maximum Likelihood Difference Scaling (MLDS).

MLDS has been recently introduced by Maloney and Yang (2003) and is a scaling method aims to produce reliable and efficient estimates of perceptual scales. It can be used with stimuli that are more complex and with stimulus differences

that are clearly visible (supra-threshold). In this thesis I propose the use of MLDS for the reliable measurement of perception.

First, I review the available scaling methods with an emphasis on their known shortcomings, which are relevant for the evaluation of MLDS (Chapter 2). Then, I present MLDS in detail with its mathematical formulation and statistical methods (Chapter 3).

The accuracy and precision of MLDS as a statistical tool has not yet been studied in detail, and this was needed before any experimental application could be done with MLDS. This thesis provides analyses of accuracy, precision and the effect of violations in MLDS model assumptions (Chapter 4), which was developed simultaneously to experimental testing[1].

Then, the validity of MLDS was tested experimentally in a study of lightness perception, presented in its entirety in Part II (Wiebel et al., 2017). This study shows how MLDS can be used to estimate perceptual scales in a scenario of perceptual constancy. So far MLDS has not been widely adopted by the visual perception community, likely because of the reluctance to commit to the various assumptions that are required by MLDS in order to statistically estimate perceptual scales. In this study we show, however, that other appearance-based procedures such as asymmetric matching also assume the presence of internal scales which are hidden and their shape can not be inferred from observers' matches.

It has recently been proposed that MLDS could be also used to derive measures of discriminability, as an alternative to performance-based methods (Devinck & Knoblauch, 2012). This possibility is investigated in-depth in the study presented in its entirety in Part III (Aguilar et al., 2017). This study provides the theoretical framework for relating MLDS with performance-based (discrimination) methods, by formulating MLDS in a framework of classical signal detection theory (Green & Swets, 1966). It tests whether MLDS could be used for measuring sensitivity, first using simulations and later experimentally using a slant-from-texture task. Although slant-from-texture stimuli can be classified as simple and artificial, it

1 These analyses have been published mostly as Appendices and Supplementary Material in the two studies presented in this thesis, Aguilar, Wichmann, and Maertens (2017) and Wiebel, Aguilar, and Maertens (2017), as well as in Wichmann et al. (2017). Here they are presented first for clarity.

was used in this study in order to compare the results with previous work (Rosas, Wichmann, & Wagemans, 2004).

Finally, Part IV provides a general discussion of the findings and an outlook on how to incorporate more realism to stimuli used in the study the visual system.

2 SCALING METHODS

The perception of a stimulus usually does not depend in a one-to-one relationship on the physical stimulus. For example, the perceived loudness of a sound doubles when the physical intensity increases ten-fold, and not two-fold[1]. In order to understand these type of relationships, we need to measure the physical stimulus as well as the perceived sensation that it produces. In that way we can study how stimuli are mapped into sensations, and ultimately develop models of how a perceptual system works.

The measurement of the physical stimulus is done by the use of physical instruments, such as a photometer, a sound pressure meter, or a ruler. However, the measurement of the magnitude of sensation is not as straightforward, and it has been the most difficult endeavor in perception research. Over the last century many procedures of measuring magnitude sensation have been devised, collectively called *scaling methods*. They aim to establish the specific relationship between the physical stimulus (x) and the perceived dimension ($\Psi(x)$), i.e. the estimation of a *psychophysical magnitude function*, *perceptual scale* or *transducer function*.

Perceptual scale

2.1 FECHNER AND EARLY SCALING METHODS

Scaling dates back to the origin of psychophysics. Fechner (1860) postulated that there must be a relationship between the physical stimulus and the internal sensation that produce in an observer that can be measured quantitatively. Weber, a predecessor of Fechner in the study of perception, worked on weight perception and discovered that the amount of stimulus intensity (Δx) that must be added to a baseline stimulus (x) to be noticeable is proportional to the baseline stim-

1 An example from the decibel scale.

ulus itself. Fechner formalized this relationship in the Weber's law, that can be expressed as

Weber's law

$$\Delta x = k \cdot x \tag{2.1}$$

where k is a proportionality constant – the Weber's constant, or Weber's fraction. Weber's law has been found to hold in the mid intensity range for many stimulus dimensions in vision and audition (Baird, 1978; Gescheider, 1997).

Fechner took Weber's law and, by postulating the existence of an internal dimension $\Psi(x)$, derived mathematically a relationship between physical stimulus x and its internal dimension. This relationship is not linear but logarithmic, and it can be expressed as

Fechner's law

$$\Psi(x) = C \cdot \ln(x/x_0)$$

where x_0 is the absolute threshold for the stimulus dimension of interest and C an arbitrary constant that depends on the experimental conditions (Baird, 1978). Experimenters are only able to measure the physical stimulus dimension x directly, with appropriate physical instruments. How can we then measure the internal, perception component $\Psi(x)$?

Fechnerian scaling

Fechner proposed that a sensation scale can be constructed by measuring and summing Just-noticeable differences (JNDs) sequentially. A JND can be defined as the difference in the stimulus dimension Δx that can be minimally distinguished for some performance criterion, for example on 75 % of the time in a 2-AFC task. Fechnerian scaling (also called discrimination scaling) constructs a sensation scale by measuring JNDs sequentially, starting at an absolute threshold, and assigning equal steps in the sensation scale. The procedure is repeated until the whole *discrimination scale* is measured (Gescheider, 1997).

Fechnerian integration has been criticized, theoretically as well as for lack of consistent experimental evidence to support it. A key assumption in Fechnerian integration is that Weber's law must hold, however evidence for many different stimulus domain show that Weber's law does not hold in many experimental scenarios such as extreme stimulus intensities (Torgerson, 1958; Baird, 1978). Another criticism stands for the method itself. The measurement of sensation is

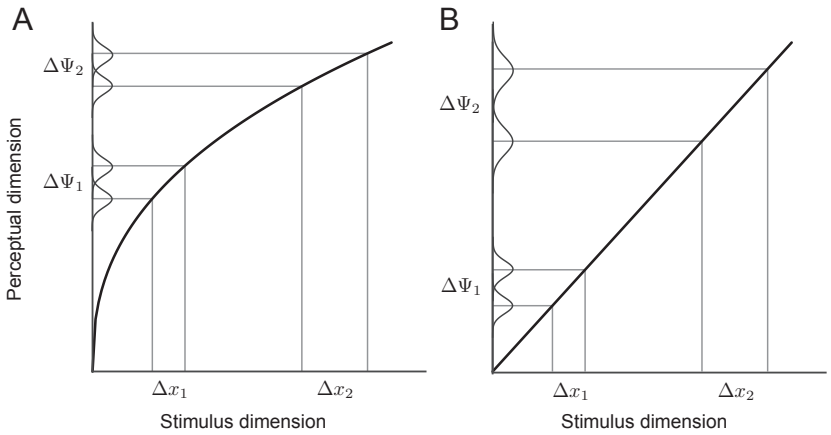


Figure 2.1: JNDs are not informative about the shape of the scale without assumptions of the internal noise. Two different JNDs (Δx_1 and Δx_2) can originate from a non-linear perceptual scale with constant noise variance (additive noise, left), or from a linear perceptual scale with increasing noise variance (multiplicative noise, right). Figure adapted from Kingdom and Prins (2010).

intrinsically noisy, and if done sequentially, measurement error could accumulate, giving less reliable estimates for increasing scale values.

JND dependency on the noise distribution

Another criticism concerns the assumed equality of JNDs in the sensation scale. In Fechnerian integration it is assumed that when JNDs are sequentially measured and summed, the same step is evoked in the sensation scale, i.e. all subjective just-noticeable differences ($\Delta\Psi$) are equal. These assumptions has been challenged by many authors (e.g. Gescheider, 1997; Kingdom & Prins, 2010), because JNDs critically depend not only on the shape of the scale but on the distribution of its noise. This criticism is illustrated in Figure 2.1, where two different functions are shown: one non-linear described by $\Psi(x) = x^{0.5}$ in panel A, and one strictly linear described by $\Psi(x) = x$ in panel B. For the non-linear function the noise is constant along the sensation dimension, i.e. it is of equal-variance, or 'additive' noise. For the linear function the noise increases along the sensation dimension, i.e. it is of unequal-variance, or 'multiplicative' noise. JNDs measured at two levels of the stimulus dimension (Δx_1 and Δx_2) are however identical for

both functions. It follows that JNDs are by themselves not informative for constructing a perceptual scale without considering the noise distribution, which by itself cannot be tested experimentally.

Due to these reasons the use of Fechnerian integration for deriving scales has been a topic of strong controversy (Krueger, 1989; Gescheider, 1997) and it is nowadays largely unused (Kingdom & Prins, 2010). However, the interest of studying sensation magnitude (or appearance) and relating it with JND-style discrimination methods (measuring sensitivity) has not decreased, because both types of measurement provide insights into the underlying perceptual mechanisms at work (e.g. Ross, 1997; Hillis & Brainard, 2005, 2007b; Maertens & Wichmann, 2013; Devinck & Knoblauch, 2012).

2.2 STEVENS AND DIRECT SCALING

A completely different approach of measuring perception was developed by Stevens (1957, 1975). Unlike Fechner, he postulated that the sensation scale can be probed directly by coupling it with a numerical response scale, by assuming that there is a direct, one-to-one mapping. By asking observers to estimate the magnitude of their sensations, the experimenter could construct a scale 'directly'. He called these methods *direct*, as oppose to *indirect* methods relying on discrimination, such as Fechnerian scaling. He devised several methods, mostly importantly the method of magnitude estimation. In this method, observers would be presented with a initial stimulus which would be explicitly anchored to some numerical value, say 50. Then, observers would be asked to rate the subsequent stimuli according to that initial anchor. For example a stimulus that elicits the 'double' sensation of the first would be assigned the number 100, a stimulus half the intensity would be assigned 25, *etc.* Thus, it is assumed that observers have direct access to the sensation scale, that they can judge their sensation magnitude and that they can assign numbers to them. It is however unclear if observers can reliable do this. Ratio or magnitude estimation can have strong unwanted confounds, such as cognitive strategies or bias intrinsic to the use of numbers (e.g. tendency to give rounded numbers) to name a few (reviewed in Gescheider, 1988; Marks & Algom, 1998).

Magnitude estimation

More fundamentally, direct scaling methods has been criticized as not providing more insights into the perceptual dimension than Fechnerian scaling. Instead, it is said that experimental data from direct scaling only is used to confirm *ad-hoc* definitions of the psychophysical magnitude function, i.e. Stevens' power law. Other definition of functions that map stimulus with sensation and responses are also possible, and direct scaling does not resolve among the different possibilities of mapping (reviewed in detail in Ch. 12 in Gescheider, 1997).

Cross-modality matching

To deal with these difficulties, Stevens developed cross-modality matching, a method that seeks to avoid these problems. In this method the observer would be presented two stimulus, one for a different sensory modality, e.g. a bright spot and a tone. The task is to adjust one of the stimuli (the light spot) in brightness until it matches in sensation magnitude with the loudness of the tone. As both sensation scales are assumed to map into the same numerical response scale, and observers are using this scale internally to match one modality with the other, it is thought that potential biases are bypassed. However, cross-modality matching across sensory modalities is at least odd and unnatural. Humans would not naturally adjust the brightness of a lamp with the loudness of a tone. Observers left to ambiguous or odd tasks can rely on countless different strategies, and the result of such experiment does not improve our knowledge of the perceptual system.

Direct scaling methods – magnitude estimation, cross-modality matching, and others – have been found to be subject of strong 'contextual effects' and a high inter-individual differences. These effects can be true differences in experimental conditions and observers, but they could also be confounds (Treisman, 1964a, 1964b; Gescheider, 1988; Marks & Algom, 1998). Although direct scaling was important for the establishment of the basic relationships of some modalities, such as pitch and loudness perception, direct scaling has fallen into disuse in the later decades (Kingdom & Prins, 2010).

2.3 THURSTONIAN SCALING

Thurstone (1927a, 1927b) proposed to study the perceptual dimension without having (necessarily) a clear mapping with the stimulus dimension. He aimed

to study abstract sensations such as artwork beauty, for which it is difficult or maybe impossible to define the physical stimulus dimension. For this aim he proposed the 'Law of Comparative Judgment' and several methods that allow the estimation of a perceptual scale without a stimulus definition.

Law of Comparative Judgment

Thurstonian scaling works by making observers judge the similarity of stimulus pairs along some perceptual dimension of interest (Torgerson, 1958). Importantly in Thurstonian scaling is the notion that perceptual representation is noisy, that is, a stimulus produces not a fixed but a variable representation in the perceptual dimension that is governed by random fluctuation. This notion is analogous and preceded signal detection theory (Green & Swets, 1966). By having observers judge pairs of stimuli repeatedly, and which are close to each other in the perceptual dimension, a measure of performance (e.g. percentage correct) can be derived. By assuming a normed dimension, e.g. equal-variance, Gaussian noise, performance for different stimulus pairs can be transformed into distances in that dimension using a z-score calculation and thus stimuli can be located in a perceptual scale. The specific methods and procedures are reviewed in detail in the scaling literature (Torgerson, 1958; McNicol, 1972; Marks & Gescheider, 2002).

The major drawback of Thurstonian scaling is the requirement of judgments of stimulus close to each other and on many repetitions. Needing a high amount of data can be problematic when multiple conditions and stimulus dimensions are of interest, especially when deriving scales on an individual basis. More fundamentally, Thurstonian scaling needs judgments of very closed spaced stimulus, which can be considered near-threshold performance. Scaling done from near-threshold performance has been problematic, because it has been historically argued that discrimination and appearance judgments belong to and can evoke different perceptual mechanisms (see Chapter 1). However, Thurstonian scaling and its notion of noisy representation, together with signal detection theory, have provided the theoretical basis of the current efforts in modern psychophysical scaling (Maloney & Yang, 2003).

2.4 PARTITION SCALING

Partition scaling results when the perceptual dimension is probed using the method of adjustment (Fechner, 1860). The simplest procedure in partition scaling is the bisection task, in which observers are given two stimulus anchors, a maximum and a minimum, and they are asked to adjust a middle stimulus so that it bisects the interval perceptually. The task can be repeated with many adjustment levels, either sequentially or simultaneously, in order to obtain a finer resolution of a scale (Gescheider, 1997). In the core, partition scaling – like Thurstonian scaling and MLDS – relies on the judgment of intervals differences, i.e. perceptual distances. The difference in partition scaling is that intervals are not fixed and presented to the observer, as in Thurstonian scaling, but rather is the observer who adjust the intervals to be perceptually equal.

Munsell scale

Partition scaling has been used in a variety of stimulus domains, importantly by Munsell, Sloan, and Godlove (1933) to derive the classical Munsell neutral value scale. This scale is used as a standard for equal steps in lightness perception for neutral gray values. More recently, Whittle (1994) also used partition scaling in the study of lightness and its relationship with discrimination thresholds.

Partition scales are not explored in detail in this thesis, however they have potential to be used in conjunction with other scaling methods.

2.5 SUMMARY

Ideally a method should provide a reliable and efficient estimation of the perceptual scale. It should not rely on JND measurement because of the reasons exposed in Section 2.1, and it should use a task that is intuitive and easy, with comparisons of clearly visible stimuli (supra-threshold), thus avoiding confounds due to task difficulty (Chapter 1). Under these criteria Table 2.1 provides a comparison overview of scaling methods, including MLDS which is reviewed in detail in the next chapter.

Thurstonian and Fechnerian scaling rely on comparison of near-threshold stimulus differences, making the task difficult and inefficient. Magnitude estimation can have strong confounding effects due to the nature of its task of verbal

	relies on JNDs	task type	task easiness	efficiency
Fechnerian scaling	yes	near-threshold comparison	+	low
Magnitude estimation	no	verbal estimation	+++	–
Thurstonian scaling	no	near-threshold comparison	+	low
Partition scaling	no	adjustment	++	–
MLDS	no	supra-threshold comparison	+++	high

Table 2.1: Comparison of scaling methods.

estimation, and it is thus nowadays largely unused. Partition scaling and MLDS provide better alternatives than all of the above, and MLDS seems suitable for the estimation of scales that is efficient, using an easy task of supra-threshold comparisons.

MLDS draws notions from Thurstonian scaling by assuming noise in the observer judgments. Judgments are made for the comparison of three or four stimulus exemplars, and it requires the definition of a stimulus dimension (unlike Thurstonian scaling) which reduces the amount of comparisons needed to construct a scale. It has been shown to be robust against different noise distributions, unlike Fechnerian integration, and it does not depend on the estimation of JNDs.

The next chapter presents MLDS, its mathematical definitions, estimation procedures, and assumptions. Then, Chapter 4 presents the results of simulations aimed to measure the accuracy and precision of MLDS in preparation for experimental testing, as well as the effect of model violations on the estimation results. These analysis were needed in order to determine MLDS performance and validity before any experimental application. The experiments using MLDS are presented in the two studies in Parts II and III of this thesis.

3

MAXIMUM LIKELIHOOD DIFFERENCE SCALING

MLDS is a scaling method developed by Maloney and Yang (2003) that consist of the construction of interval scales based on the judgment of interval differences. Contrary for other scaling methods, MLDS recognizes that observers' responses are stochastic, and so it takes into account that judgments of interval differences are noisy.

As all scaling methods, MLDS aimed to measure the psychophysical magnitude function, which maps the physical stimulus variable (x) with an internal sensation variable $\Psi(x)$ (also noted Ψ_x for simplicity). It is common to assume a power function (after Stevens, see Chapter 2) although other functions are allowed. A power function can be described by the formula

$$\Psi(x) = x^e \tag{3.1}$$ *Power function*

where e is the exponent parameter and controls the curvature of the function. In MLDS the observer is presented with three (or four) stimulus exemplars (x_i) that elicit discrete perceptual responses ($\Psi(x_i)$). An ensemble of three stimulus exemplars is called a triad (x_1, x_2, x_3), and an ensemble of four exemplars is called a quadruple (x_1, x_2, x_3, x_4).

In the method of triads, the task of the observer is to judge which pair of stimulus elicits a bigger difference, either the pair (x_1, x_2) or the pair (x_2, x_3). The observer is thus comparing the interval $[\Psi_{x_3} - \Psi_{x_2}]$ with the interval $[\Psi_{x_2} - \Psi_{x_1}]$. A decision variable can be written as the difference between these two intervals

$$\Delta = |\Psi_{x_3} - \Psi_{x_2}| - |\Psi_{x_2} - \Psi_{x_1}| + \epsilon \tag{3.2}$$ *Decision variable*

The content of this chapter has been partly published in Wichmann et al. (2017), and as Appendices and Supplementary Material in Aguilar et al. (2017) and in Wiebel et al. (2017).

MLDS includes stochasticity in the decision variable with the term ϵ, which is Gaussian distributed noise, with zero-mean and variance σ^2, i.e. $\epsilon \sim N(0, \sigma^2)$. Using this decision variable, the model selects the stimulus pair (x_2, x_3) as larger when $\Delta > 0$, otherwise it selects the pair (x_1, x_2).

Alternatively, MLDS can be used with the method of quadruples, in which case the decision variable is

$$\Delta = |\Psi_{x_4} - \Psi_{x_3}| - |\Psi_{x_2} - \Psi_{x_1}| + \epsilon \tag{3.3}$$

with same noise distribution and decision rule than for the method of triads. This thesis uses only the method of triads.

Giving observer judgments to triads (or quadruples), the goal of MLDS is to obtain an interval scale that reflects the underlying internal sensory function $\Psi(x)$ and an estimate of the variance of the noise parameter in the model $\hat{\sigma}$.

3.1 CONSTRUCTION OF TRIADS

During the design of an MLDS experiment the number of stimulus exemplars (x_i) needs to be decided in order to construct the triads. First, the range of the stimulus dimension under study must be decided. For example, if we are interested in the lightness domain, the stimulus dimension is physical reflectance, an unitless dimension that can range from 0 (no reflectance, perceived as black) and 1 (full reflectance, perceived as white). On this range, p different stimuli can be placed uniformly. The triads are then constructed by selecting all possible non-overlapping intervals using p discrete stimulus values. As an example, if $p = 4$, the complete set of triads are $(1,2,3), (1,2,4), (1,3,4), (2,3,4)$. The triplet $(3,2,4)$ is not a valid triad for MLDS because its intervals overlap. The total number of possible triads can be calculated as

$$n = \binom{p}{3} = \frac{p!}{(p-3)! \times 3!}$$

These complete set of triads must be repeated many times (r) to obtain reliable estimates. Thus, a full MLDS experiment consist of an observer judging $N = n \times r$

$p =$	7	8	9	10	11	12	13	14	15	20
$n =$	35	56	84	120	165	220	286	364	455	1140

Table 3.1: The number of possible triads (N) grows rapidly with increasing stimulus number (p).

trials, which can be done in a blocked design, with r blocks containing n trials each[1]. The order of the stimulus in the triad, with either ascending or descending values, must be randomized so to avoid systematic spatial or temporal arrangement of the stimuli that could lead to confounds.

3.2 ESTIMATION USING A GENERALIZED LINEAR MODEL

Once stimulus triads are presented to the observer and his/her answers collected, we can proceed to estimate the scale. MLDS has been implemented with two different algorithms: using direct optimization (Maloney & Yang, 2003) or as a Generalized linear model (GLM) for logistic regression (Knoblauch & Maloney, 2008, 2012). I focus on the GLM implementation as it is the newest and most used one.

For the GLM implementation we must further assume that the sensory function is monotonically increasing, thus avoiding the absolute value operation in Equation 3.2 and leading to a decision variable that can be rewritten as a linear combination of the sensory variables

$$\begin{aligned}\Delta &= (\Psi_{x_3} - \Psi_{x_2}) - (\Psi_{x_2} - \Psi_{x_1}) + \epsilon \\ &= \Psi_{x_3} - 2\Psi_{x_2} + \Psi_{x_1} + \epsilon\end{aligned} \tag{3.4}$$

1 In this work we chose to use full repetitions of the complete set of triads. Previous work with MLDS have instead use a procedure that randomly draws a triad (from the complete set) until the final number of desired trials is reached.

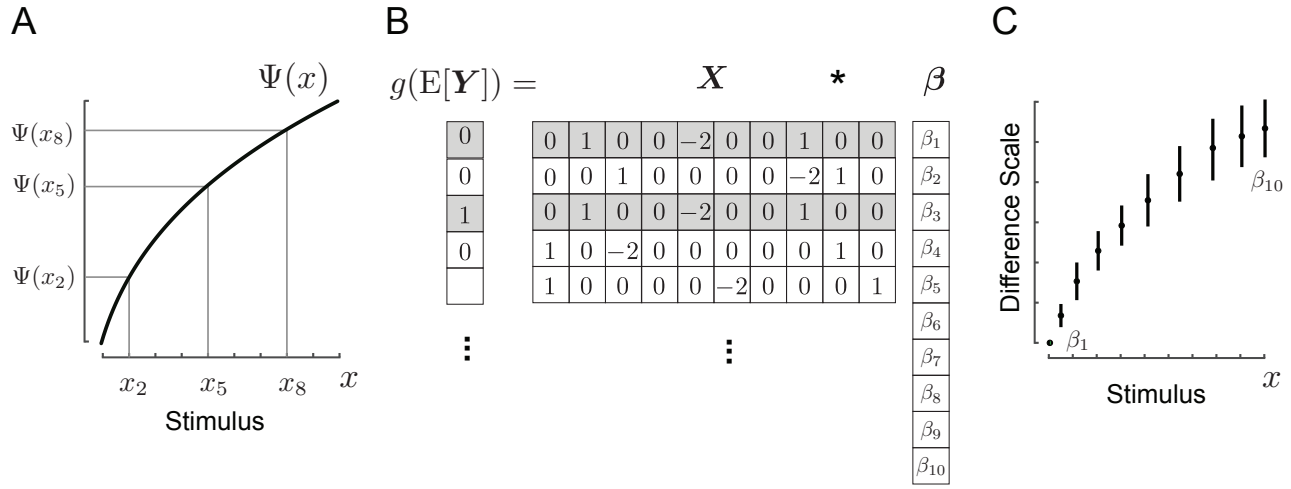


Figure 3.1: Estimation of scales using MLDS with triads. An example triad (x_2, x_5, x_8) evokes perceptual experiences $\Psi(x)$ on an hypothetical sensory function (A). The decision model for this example triad is $\Delta = \Psi(x_8) - 2\Psi(x_5) + \Psi(x_2)$. (B) GLM construction. A design matrix (X) is constructed by setting the entries in the corresponding columns as the weights of each $\Psi(x)$ from the decision model. Shaded rows indicate two repetitions of the same triad. Y is the binomial response variable. (C) The estimated scale is obtained by solving the GLM and finding the coefficients β, which correspond to the scale values at different levels of the physical variable. After Wiebel et al. (2017).

Under this assumption, MLDS can be reformulated as a GLM (Figure 3.1). The GLM is set up by taking the responses of an observer and assigning them to the response variable ($\mathbf{Y}$), and by constructing a design matrix ($\mathbf{X}$) that represents the weights of each $\Psi(x_i)$ component on the decision variable in a single-trial basis. The goal is to find the coefficients $\boldsymbol{\beta}$ that best account for the data ($\mathbf{Y}$) given $\mathbf{X}$. Formally, the GLM is described by

$$g(E[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta} \tag{3.5}$$

GLM

where $\mathbf{Y}$ is a vector of length n with entries 0 or 1, indicating the observer's response (for first vs. second pair, respectively). $\mathbf{X}$ is the design matrix of size $n \times p$, whereby n is the total number of triads and p is the number of stimulus levels sampled as well as the number of estimated points on the perceptual scale. Each row in matrix $\mathbf{X}$ contains non-zero entries (1,-2, 1) in the columns corresponding to the stimulus values for the presented triad values (x_1, x_2, x_3), and zero entries in the remaining $p-3$ columns. The coefficient vector $\hat{\boldsymbol{\beta}}$ is of length p and it contains the scale estimates (Figure 3.1C).

Design matrix

The link function $g()$ is required to establish the relationship between the *linear predictors* $\mathbf{X}\boldsymbol{\beta}$ and the mean response variable $E[\mathbf{Y}]$, where $\mathbf{Y}$ is binomially distributed with n=1 (also known as a Bernoulli process). The default link function for MLDS is the inverse of the Gaussian cumulative distribution function (Φ^{-1}), as it has been shown to be robust against different noise distributions (Maloney & Yang, 2003), and it was used throughout this work.

To make the model solvable, the first row of the design matrix is dropped, which effectively anchors the scale at a minimum of zero, $\beta_1 = 0$. The rest of the coefficients $\hat{\beta}_2 \cdots \hat{\beta}_p$ are estimated by maximum likelihood using standard GLM solvers. The coefficient vector $\hat{\boldsymbol{\beta}}$ is the estimated difference scale, and it is a interval scale (Figure 3.1C). The estimation using GLM produces scales where its maximum is related inversely to the estimated noise parameter of the model $\hat{\sigma}$

$$\hat{\beta}_p = \frac{1}{\hat{\sigma}} \tag{3.6}$$

These type of sales are called 'unconstrained' scales in MLDS literature (Knoblauch & Maloney, 2012).

3.3 MLDS AND SIGNAL DETECTION THEORY

The decision model underlying MLDS can be also framed in terms of signal detection theory. Figure 3.2 shows the equivalence between MLDS on its original formulation and forced-choice procedures used for signal detection theory. The equivalence has been suggested by Devinck and Knoblauch (2012) and is explained in the following.

An 'unconstrained scale' in MLDS can be converted to a normed scale defined in units of d', when the following assumptions are met. First, it is assumed that the decision process is not stochastic but deterministic. This would attribute all of the observed noise to the sensory representation, $\psi(x)$, which is a Gaussian random variable with mean $\Psi(x)$. Second, it is assumed that the noise is constant, i.e. independent of the stimulus level. Finally, it is assumed that the sensory representations are independent of each other. It follows from these assumptions that the $\psi(x)$ are independent Gaussian random variables with equal variance[2]. Then, the noise parameters can be 'carried' to the sensory representation, by rewriting the decision model (Equation 3.4) in this way

MLDS in signal detection theory formulation

$$\psi_{x_i} \sim \mathcal{N}\left(\Psi_{x_i}, \frac{\sigma^2}{4}\right) \tag{3.7}$$

$$\Delta = (\psi_{x_3} - \psi_{x_2}) - (\psi_{x_2} - \psi_{x_1}) \tag{3.8}$$

The variance of the decision variable Δ is σ^2 (Equation 3.4). When rewriting the model equations, the variance of each sensory representation ψ_x must be

2 This restricted case can be derived from the MLDS model only because: (i) the decision variable after the differencing is assumed to be Gaussian, for which the simplest case is when it is produced by underlying Gaussian distributed representations; and (ii) equal-variance is the simplest case to relate the variance of each representation with the variance of the decision variable. Other models (e.g. unequal-variance) cannot be derived easily as they would be underconstrained.

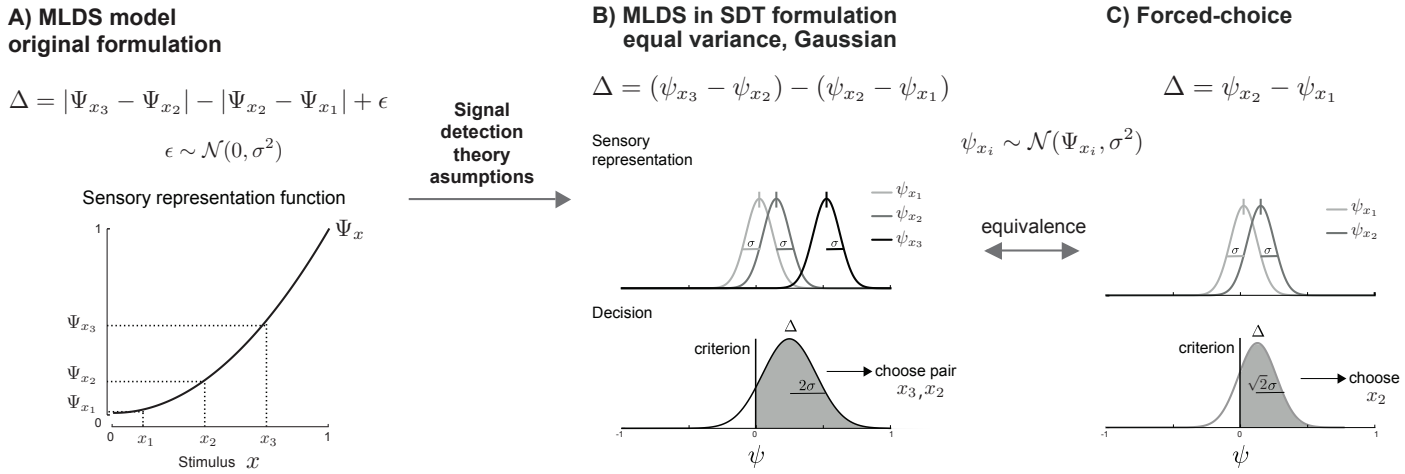


Figure 3.2: MLDS in the signal detection framework. (A) In its original formulation the decision variable (Δ) in MLDS is defined as the difference between intervals ($|\Psi_{s_3}, \Psi_{s_2}|$ and $|\Psi_{s_2}, \Psi_{s_1}|$) and this difference is corrupted by Gaussian noise (ϵ). (B) In the signal detection formulation of MLDS the noise originates only from the sensory representations (ψ_s) which are assumed to be independent Gaussian random variables with equal variance. In the signal detection version of MLDS the model is equivalent to forced-choice methods (C) at the level of the sensory representation. See text for details. Adapted from Aguilar et al. (2017).

adjusted so that Equation 3.4 still holds. Because the decision variable Δ is computed as a linear combination of four independent, Gaussian random variables, its variance is four times the variance of each individual variable ψ_{x_i}. Therefore each individual variance in the sensory representation needs to be 'corrected' by a factor of $1/4$.

Yet, MLDS provides the noise estimate $\hat{\sigma}$ (Equation 3.6) as an estimate of parameter σ of the decision variable and not of the sensory representation directly. By knowing the above explained relationship between the variance in the sensory representation and in the decision variable, the difference scale can be adjusted so as to represent the variance in the sensory representation. The $\hat{\sigma}$ estimated by MLDS corresponds to four times the variance present in the sensory representation ψ_x (Equation 3.7). Thus, the conversion is accomplished by multiplying the original scale by a factor of two (as done also by Devinck & Knoblauch, 2012). Formally, the new transformed scale maximum is two times the maximum of the original scale

d′ scales

$$\hat{\beta'_p} = \frac{1}{\frac{\hat{\sigma}}{2}} = 2\frac{1}{\hat{\sigma}} = 2\hat{\beta_p} \tag{3.9}$$

This new scale $\hat{\beta}'$ is in "d′ units", i.e. an interval difference of one in the scale dimension should represent a performance of d′ of one, when all assumptions are met. The parametrization in units of d′ can be used to derive sensitivity (see Chapter 10).

3.4 ESTIMATION ERROR

The variability of the scale estimation is determined using bootstrapping techniques (Knoblauch & Maloney, 2008, 2012). The goal is to estimate the variability of the coefficients $\hat{\boldsymbol{\beta}}$. For that purpose, Equation 3.5 is rearranged in order to compute the mean response probability for each triad from the fitted data

$$E[\mathbf{Y}] = g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{3.10}$$

The obtained vector $E[\mathbf{Y}]$ contains the expected probability of a Bernoulli variable (Y) for each triad, in other words, the mean probability of binary responses given the presented stimulus values in each triad. These probabilities are used to simulate a Bernoulli response in each triad, which is in turn used to estimate a new set of coefficients $\hat{\beta}^*_j$, $j = 1..p$ using the same GLM procedure. The coefficients $\hat{\beta}^*_{ij}$, $j = 1..p$ are the *i-th* bootstrap sample, and many bootstrap samples are drawn by repeating the simulation procedure many times ($N_s = 10000$). A matrix $\mathbf{S}$ with all the samples can be constructed with $N_s \times p$ entries.

$$S_{i,j} = \begin{pmatrix} \hat{\beta}^*_{1,1} & \cdots & \hat{\beta}^*_{1,p} \\ \vdots & \ddots & \vdots \\ \hat{\beta}^*_{N_s,1} & \cdots & \hat{\beta}^*_{N_s,p} \end{pmatrix} \tag{3.11}$$

From this matrix the confidence intervals for each scale estimate can be obtained from the distribution of bootstrap samples at a confidence of $(1-2\alpha)$ (e.g. $\alpha = 0.025$ for 95 % CIs).

Boostrap confidence intervals

There are multiple ways to obtain confidence intervals from bootstrap samples. The simplest and straightforward method is the 'percentile' method, in which the confidence intervals are drawn directly from the (α) and $(1-\alpha)$ percentiles of the bootstrap samples distribution. The percentile method often produce confidence intervals that are too small and do not represent the underlying variability, specially when the sample distribution violates the normality assumption (Efron & Tibshirani, 1993).

To avoid these problems, Efron and Tibshirani (1993, pp. 184-188) proposed the calculation of 'bias-corrected and accelerated' (BCa) confidence intervals. This method is robust against skewed distributions and is recommended for standard use. It is also the method of choice in estimation of psychometric functions using bootstrap (Wichmann & Hill, 2001). BCa confidence intervals are used throughout this thesis.

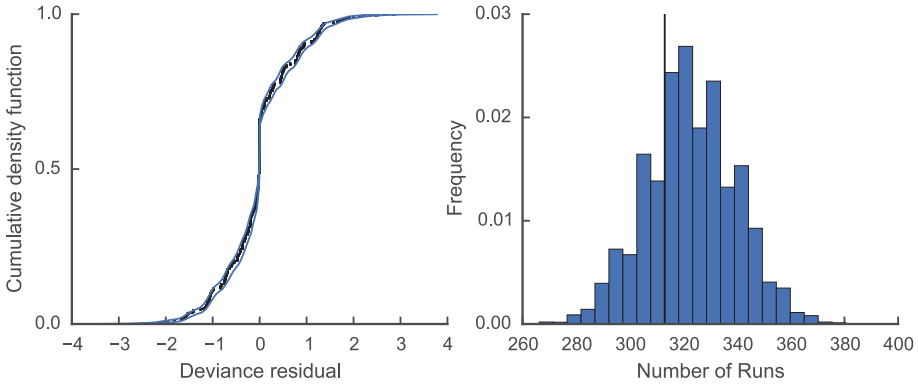


Figure 3.3: Goodness of fit diagnostics for MLDS (as provided by the R package) for a simulated experiment. (A) Cumulative distribution function of the linear predictors $X\beta$. The observed data is depicted in the markers, and its 95 % confidence interval obtained by bootstrap as lines. (B) Test for independence of deviance residuals. Histogram of 'number of runs' obtained by bootstrap, and the position of the observed number or runs (vertical line). See text for details.

3.5 GOODNESS OF FIT

Unlike for linear models, the goodness of fit of a binomial GLM can be difficult to evaluate. Two generic measures of GLM can be used. The first is the Akaike information criterion (AIC), a general metric of the likelihood weighted by the number of parameters, and that is commonly used for model comparison. The second is the 'deviance accounted for' (DAF), which is the reduction of deviance by the model fit with respect to the deviance of the null model, i.e. a model with only an intercept (Wood, 2006; Knoblauch & Maloney, 2012). The DAF is conceptually similar to the R^2 measure in linear regression, interpreted as the percentage of variance explained.

Goodness of fit diagnostics

However, the evaluation of the residuals of these type of GLM can be problematic, given that they distribute binomially (Knoblauch & Maloney, 2008, 2012). For this reason Wood (2006) recommends to test binomial GLM using Monte Carlo methods. The procedure involves the simulation of binomial responses according to the expected probabilities $E[\mathbf{Y}]$ (Equation 3.10). For each simulation,

the deviance residuals are obtained and sorted in a cumulative distribution function (cdf). Figure 3.3A shows the experimental distribution of deviance residuals (markers), and the $(1-\alpha)\%$ confidence envelope obtained from all ($N_s = 10000$) simulated cdfs (blue lines). If the decision model in MLDS is appropriate, the experimental distribution should be inside the confidence envelope, which is indeed the case in Figure 3.3A. Additionally, the independence of the deviance residuals can also be tested using the same Monte Carlo method. For each simulation described above, the deviance residuals are sorted according to the linear predictors, and the number of times the sign in the deviance residual changes is counted ('number of runs' in Figure 3.3B). An independent distribution of deviance residuals should have random distribution of sign and no evident trend. The value obtained experimentally (vertical line in Figure 3.3B) is compared with the distribution obtained from the Monte Carlo simulations by calculating a p-value. This procedure produces a p-value that is *not significant* if the deviance residuals show independence.

3.6 MEASUREMENT ASSUMPTIONS

MLDS bases its difference model in classical measurement theory. Krantz, Luce, Suppes, and Tversky (1971, Ch. 4, pp. 247) defines a series of axioms that must be fulfilled to ensure the existence of a scale that can be derived from the judgment of stimulus differences, or the judgment of intervals. Importantly for MLDS are the transitivity axiom and the 'weak monotoncity' axiom (*Axioms 3 and 4* in Krantz et al. (1971)).

Transitivity axiom

The transitivity axiom states that if two intervals $(x_1, x_2), (x_2, x_3)$ exist in the set of all possible intervals (X^*), then the interval (x_1, x_3) must be bigger than each of them alone

$$\begin{aligned} &\text{if} \quad (x_1, x_2), (x_2, x_3) \in X^* \\ &\text{then} \quad (x_1, x_3) > (x_1, x_2) \text{ and } (x_1, x_3) > (x_2, x_3) \end{aligned}$$

where $>$ stands for the binary comparison operation 'bigger than'. In other words, this axiom states that the stimuli x_i belong to an ordered dimension, for

example length, luminance, brightness, *etc.* It translates experimentally into an observer that can judge $x_3 > x_2 > x_1$, and thus who is able to order the stimulus increasingly.

Weak monotonicity axiom

The 'weak monotonicity' axiom states that given two pairs of contiguous intervals, $(x_1, x_2), (x_2, x_3)$ and $(x_4, x_5), (x_5, x_6)$,

$$\begin{aligned} &\text{if } (x_1, x_2) > (x_4, x_5) \\ &\text{and if } (x_2, x_3) > (x_5, x_6) \\ &\text{then } (x_1, x_3) > (x_4, x_6) \end{aligned}$$

This axiom implies that in the measured dimension the intervals increase as the distance between the stimulus gets larger, in other words, the function is monotonically increasing. Clear examples are lightness, depth and distance[3]. This assumption can also be tested experimentally in MLDS with the method of quadruples, by analyzing directly if the observer's judgments comply with the inequality stated in the axiom. For the method of triads this is not possible due to how the triads themselves are constructed. Thus, when using the method of triads the monotonicity assumption is a requisite that must be assumed and cannot be tested experimentally.

3.7 SUMMARY OF MLDS

To summarize, the goal of MLDS is to obtain an estimation of the internal scales for a stimulus dimension of interest, by asking observers to judge stimulus differences in triads or quadruples. It does so by setting up a statistical model (a generalized linear model) that estimates the scale values. The variability on the estimation and the goodness of fit is evaluated using bootstrap methods.

MLDS assumes:

- a function that maps the physical stimulus variable x to an internal perceptual scale $\Psi(x)$.

3 A counterexample would be a function with a shape of an inverted U, with a peak in the middle of the stimulus range.

- a function $\Psi(x)$ that is monotonically increasing. This allows the GLM implementation and the compliance with axioms of measurement theory.
- the observer judges the difference between the stimulus exemplars using a difference model (Equation 3.2).
- this judgment is corrupted by independent, Gaussian distributed noise.
- the observer is able to order the stimulus increasingly.

MLDS can be also parametrized in a signal detection framework, for which the obtained difference scales are in units of d'.

In the next chapter I addressed some of these assumptions using simulations, as well as determining the accuracy and precision of MLDS as a statistical tool. To anticipate, I found that MLDS could reliably estimate the generative perceptual function with low bias, providing at the same time a reliable estimate of the noise in the decision model when the equal-variance assumption holds. Additionally, MLDS provided confidence intervals that were stable in precision and adequate in coverage. These results provided the basis for the following experiments that tested the performance of MLDS experimentally (Part II, starting at Chapter 5).

4 SIMULATIONS

Since its introduction by Maloney and Yang (2003), the bias and variability of the MLDS estimation itself has not been studied in detail, as well as the effect of violations of its model assumptions. In this chapter I explore these issues using numerical simulations. These simulations also provide the upper limit of precision when a limited amount of data is available, for the practical use case in psychophysical experiments.

4.1 SPECIFICATION

The sensory functions (or transducer functions) were power functions with three different exponents ($e = 0.5, 1.0$ and 2.0) that gives the function a different curvature. The stimulus dimension x was set to the normalized range $[0, 1]$, and $p = 8$ stimulus values were linearly spaced on this range. As detailed in Section 3.1 this spacing gives $n = 56$ unique triads to be presented.

The sensory functions were subjected to Gaussian-distributed noise of either (i) equal variance ('additive noise'), or (ii) unequal variance ('multiplicative' noise). For the equal-variance case, the draw of the function's response to the stimulus x_i can be expressed as

$$\psi_{x_i} \sim \mathcal{N}\left(\Psi_{x_i}, \sigma^2\right) \tag{4.1}$$

Equal-variance noise

where σ is the fixed level of the Gaussian-distributed noise, and it could have values of $\sigma = \{0.035, 0.07, 0.14\}$. The choice of these values was informed by the study reported in Aguilar et al. (2017) (and presented in part III of this

The content of this chapter has been partly published in Wichmann et al. (2017), and as Appendices and Supplementary Material in Aguilar et al. (2017) and in Wiebel et al. (2017).

dissertation): $\hat{\sigma} = 0.07$ was the average value of estimated noise found in eight observers, and 0.035 and 0.14 were the respective half and double of that average. The range $[0.035, 0.14]$ covered all values observed experimentally.

For the unequal variance case, the noise increased constantly with the stimulus dimension, from $\sigma_{min} = 0.035$ at $x = 0$ to $\sigma_{max} = 0.14$ at $x = 1$. Formally this case can be expressed as

Unequal-variance noise

$$\psi_{x_i} \sim \mathcal{N}\left(\Psi_{x_i}, \sigma^2(x)\right) \tag{4.2}$$

where the standard deviation of the Gaussian-distributed noise increase linearly with the stimulus value

$$\sigma^2(x) = [(\sigma_{max} - \sigma_{min}) \cdot x + \sigma_{min}]^2$$

The unequal-variance noise case correspond to the most studied case of unequal-variance: the noise increases with the stimulus dimension as obeying Weber's law.

The decision variable was as assumed by MLDS for the method of triads (Equation 3.8), rewritten here for clarity:

$$\Delta = (\psi_{x_3} - \psi_{x_2}) - (\psi_{x_2} - \psi_{x_1}) \tag{4.3}$$

As described in Section 3.1, the set of unique triads needs to be repeated many times, thus varying the total number of trials in the experiment. In these simulations the total number of trials was varied systematically with values of $N \in \{280, 560, 840, 1680, 2520, 3360\}$, which is the result of having the set of $n = 56$ triads repeated $r \in \{5, 10, 15, 30, 45, 60\}$ times.

Each simulation was fed into the MLDS analysis routine (Section 3.2) that included the estimation error using bootstrap (Section 3.4) and the goodness of fit analysis (Section 3.5). The results are presented in the following sections.

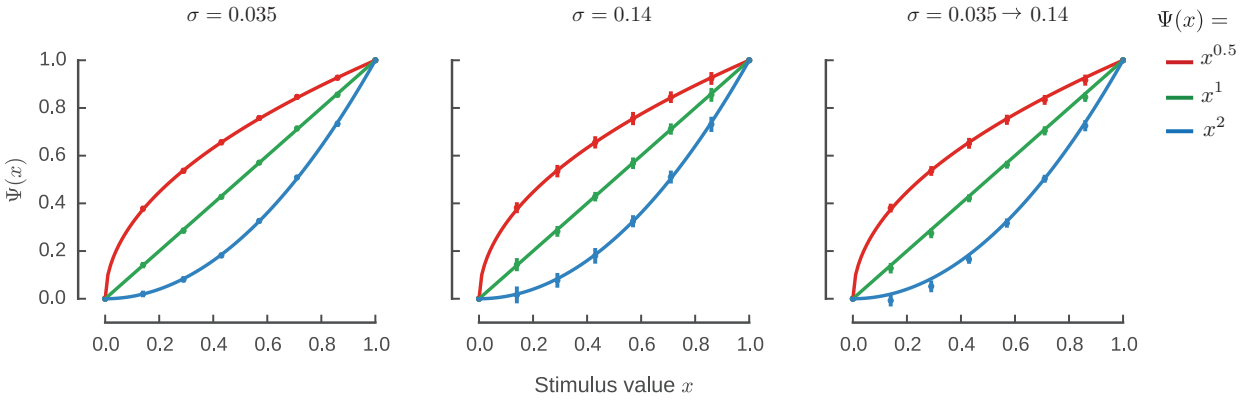


Figure 4.1: Scale estimation for three noise conditions and three different exponents for the sensory function $\Psi(x)$ ($M \pm 1S.D.$ across $N = 560$ simulated trials, scales normalized to their maximum.

4.2 BIAS OF SCALE AND NOISE ESTIMATES

Figure 4.1 shows an example of the scales estimated by MLDS for different exponents and different noise conditions. Scale values are normalized to the range [0,1] and depicted as errorbars showing the $M \pm 1S.D.$ across simulations. Ground truth functions are shown as continuous lines. The estimated values followed closely the ground truth functions for all the noise conditions studied, indicating low bias. These results replicate previous work for the equal-variance case (Maloney & Yang, 2003), and additionally they show how MLDS seems to be robust against the presence of unequal-variance noise (or 'multiplicative' noise, Figure 4.1 right). In addition, low bias was already present at relatively few number of trials: Figure 4.1 shows an example with $N = 560$ trials, which is equivalent to repeat $r = 10$ times the total number of unique triads. Thus, MLDS can estimate the underlying functions with a low bias, also in the presence of multiplicative noise.

The scales estimated in MLDS using the 'unconstrained' parametrization also provide an estimate of the noise parameter ($\hat{\sigma}$), because the scale maximum is

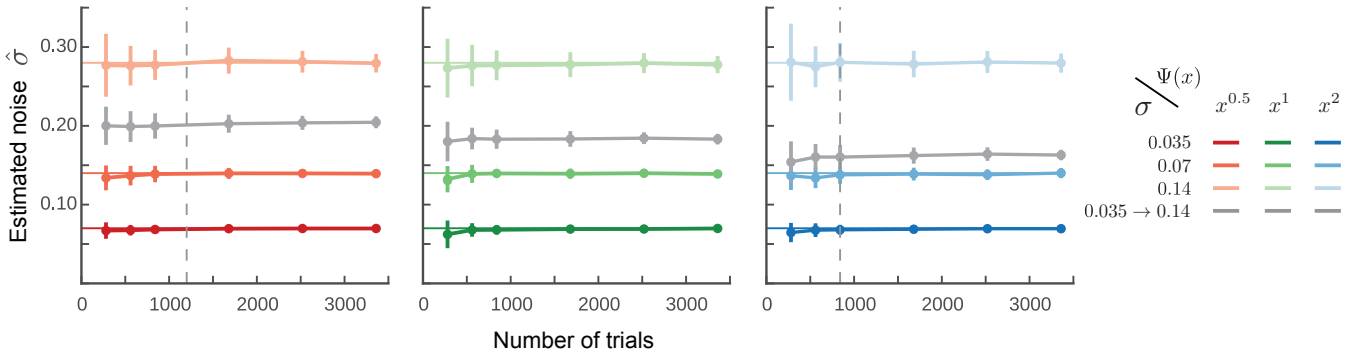


Figure 4.2: Noise estimation for all simulated conditions. Horizontal lines show the expected values, errorbars indicate M ± 1S.D. across simulations. The dashed vertical lines indicate the number of trials used in the experiments in Part III (rightmost, for $\Psi(x) = x^{2.0}$) and in Part II (leftmost, for $\Psi(x) = x^{0.5}$).

inversely related to the noise estimated by the model (Equation 3.6). Figure 4.2 shows the noise estimates as a function of the number of trials for all conditions. The estimated noise (errorbars) was in accordance with the expected values (horizontal lines) for all conditions when the equal-variance assumption held. For the unequal variance case ("0.035 → 0.14", continuous gray errobars) the estimated noise had values between the minimum and maximum as expected for the equal-variance case. For all conditions the estimated noise was stable around the expected value, showing low bias and low variability when more than approx. N = 1000 trials were used. Thus, MLDS can also reliably recover the noise in the decision model when the equal-variance assumption holds.

4.3 VARIABILITY AND COVERAGE OF CONFIDENCE INTERVALS

As detailed in Section 3.4, the error estimation in MLDS is calculated using bootstrap techniques. Bootstrap allows to calculate confidence intervals for the scale values, and the width of the confidence intervals are an indication of the variability in the estimation procedure. Figure 4.3 shows the width of the confidence

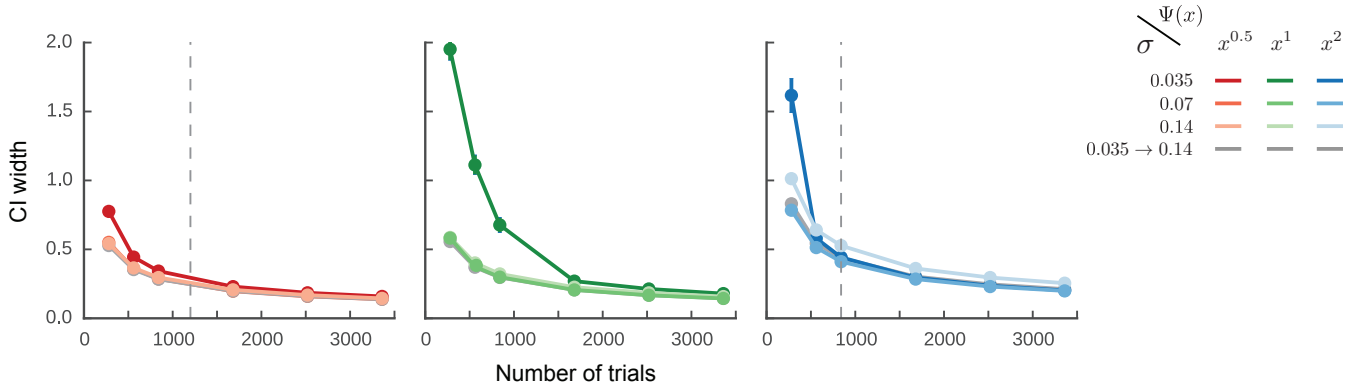


Figure 4.3: Width of the confidence intervals (normalized by the scale value itself) as a function of number of trials, for all simulated conditions. Example at a stimulus value $x = 0.51$. Vertical lines as in Figure 4.2

intervals (normalized by the scale value itself) as a function of the number of trials. It is expected that the estimation procedure gets more precise (less variable) as more trials are simulated and included in the analysis; this was indeed the case for all simulated conditions. The width of the confidence intervals decrease in an exponential way, with the biggest gain in precision on the first N = 1500 trials.

Coverage

Bootstrap techniques are known to produce confidence intervals that are too narrow, i.e. with low *coverage* (Wichmann & Hill, 2001). Coverage can be defined as the number of times the true value (as defined in the ground truth function) is included in the confidence interval across multiple simulations. Credible confidence intervals must have a coverage that is equal to the statistical confidence that is used to calculate them. In this way, coverage must be 95 % for confidence intervals calculated with 95 % statistical confidence level, i.e. the true value must be included in 95 % of the simulations. As MLDS uses bootstrap to calculate confidence intervals, it is relevant to check whether MLDS provides confidence intervals with adequate coverage. Figure 4.4 shows the coverage for all simulated conditions. Coverage approximated the expected 95 % as the number of trials is increased, and it was adequate for all conditions studied with N = 1000

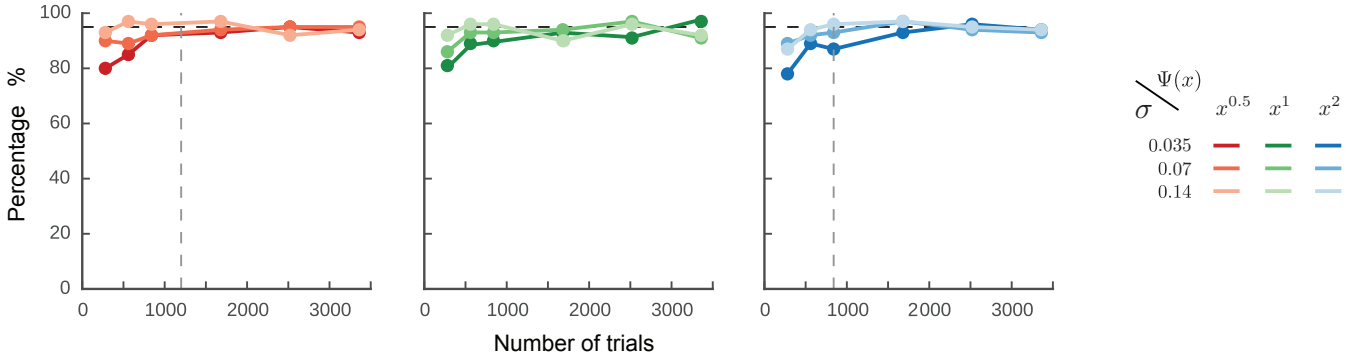


Figure 4.4: Coverage of confidence intervals as a function of number of trials. Horizontal dashed line depicts 95 % coverage; vertical lines as in Figure 4.3.

trials or more. Adequate coverage was largely independent of the underlying noise level.

4.4 GOODNESS OF FIT AND VIOLATION OF THE EQUAL-VARIANCE ASSUMPTION

Violations of some model assumptions are likely to occur in psychophysical experiments. It has been shown that some discrimination data is better fit when an unequal-variance model is assumed (e.g. Kingdom & Prins, 2010; Goris, Putzeys, Wagemans, & Wichmann, 2013). Thus, it is also desirable to study the effect of model violations in the outcome of MLDS, and determine whether the goodness of fit procedure in MLDS can detect these cases.

Figure 4.5 shows the percentage of simulations in which the goodness of fit was tagged as acceptable according to the p-value described in Section 3.5. For mostly all conditions the goodness of fit was acceptable in more than 95 % of the cases; the only exception was the unequal-variance noise condition with exponent of two at large number of trials (gray lines in Figure 4.5 right).

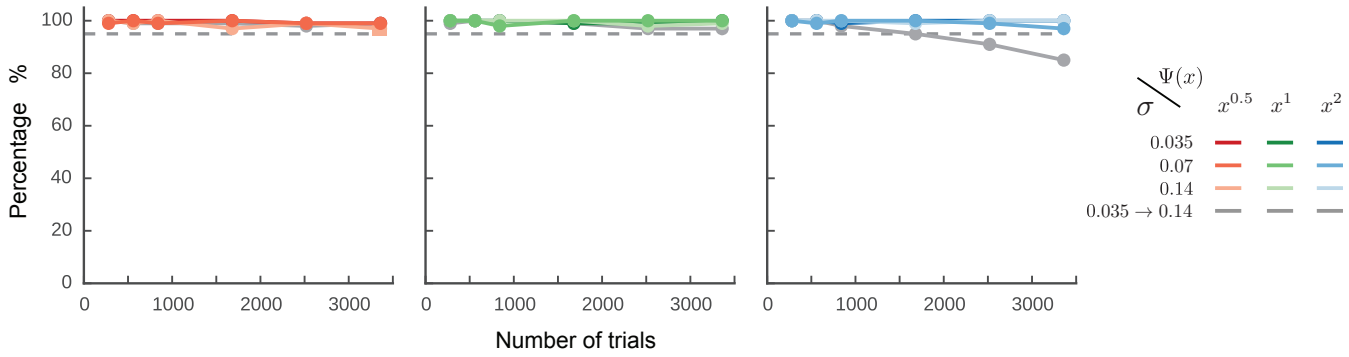


Figure 4.5: Percentage of simulations with acceptable goodness of fit, as described in Section 3.5. Horizontal dashed line indicates 95 %.

These results indicate that the goodness of fit procedure in MLDS mostly did not detect the violation of the equal-variance noise assumption. MLDS masked the underlying noise distribution when unknown, and therefore the outcome of an MLDS experiment when unequal-variance noise is present will be indistinguishable from an equal-variance case. This is a clear disadvantage in comparison to classical signal detection theory procedures, in which the unequal-variance case can be estimated from discrimination data (McNicol, 1972; Knoblauch & Maloney, 2008).

However, MLDS is robust in estimating the underlying function –for what it was designed for– in the presence of unequal-variance noise (Figure 4.1). In this respect MLDS overcomes the shortcomings of Fechnerian integration, recovering a scale that is independent of the noise distribution. As reviewed in Equation 2.1, scales derived by Fechnerian integration of JNDs cannot be guaranteed to recover the true function, as JNDs depend on the noise distribution and the distribution itself cannot be determined experimentally. MLDS avoids this issue by estimated the scale values independently of the noise distribution.

4.5 VIOLATION OF OTHER MODEL ASSUMPTIONS

Other assumptions in MLDS that could also be violated are reviewed here. It is relevant to know the effect of the violations on estimation, as these assumptions cannot be tested experimentally.

4.5.1 *Correlation*

MLDS in the signal detection theory form assumes that each draw from the sensory function is *independent* of each other (Section 3.3). In the method of triads the stimulus array is presented simultaneously, and therefore it is possible that correlation may exist among the sensory responses for a triad. In this case the independence assumption would be violated, and the estimates of the noise returned by MLDS may differ. Using simulations we can analyze quantitatively the differences that occur in the presence of correlated noise.

Correlated observer

To simulate correlated noise the observer model in Equation 4.1 must be modified by expressing it in a vectorized form. In this way, each triad is a vector $\vec{x} = (x_1, x_2, x_3)$ that is presented to the observer and evokes a vector of sensory responses ($\psi_{\vec{x}}$) that is drawn from a multivariate Gaussian process

$$\psi_{\vec{x}} \sim \mathcal{N}\left(\Psi(\vec{x}), \vec{\Sigma}\right)$$

with covariance matrix

$$\vec{\Sigma} = \begin{pmatrix} \sigma^2 & \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 & \sigma^2 \end{pmatrix}$$

The covariance matrix ($\vec{\Sigma}$) has diagonal components σ^2 and non-diagonal components σ_c^2. The diagonal values σ^2 are the assumed variance values for the non-correlated noise, and it was set to $\sigma^2 = 0.07^2$ as in previous simulations. The non-diagonal values σ_c^2 represent the covariance between the sensory variables

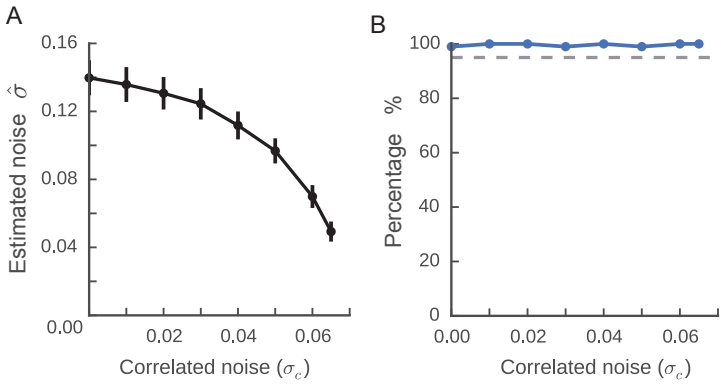


Figure 4.6: Results of simulated observers with added *correlated noise* (σ_c), violating the independence assumption. (A) Estimated noise level $\hat{\sigma}$ as a function of the correlated noise σ_c added to the simulation, for a fixed non-correlated noise level ($\sigma = 0.07$). Errorbars indicate M $\pm$ S.D. across N=100 simulations. (B) Percentage of simulations with acceptable goodness of fit from panel (A). Horizontal dashed line indicates 95 %.

in that triad, and correlation is introduced by setting $\sigma_c^2 > 0$. In principle each non-diagonal entry in the covariance matrix could have a different value, but for simplicity they were all set to a fixed value (σ_c^2). This "correlated observer" was simulated using the same procedure as previously described.

Figure 4.6A shows the simulation results by plotting the noise value estimated by MLDS ($\hat{\sigma}$, y-axis) as a function of increasing correlated noise (σ_c, x-axis). When no correlation is present ($\sigma_c = 0$) the estimated noise value matches the expected value ($\hat{\sigma} = 0.14$ for a simulated $\sigma = 0.07$, see Equation 3.9). When correlation is added, MLDS returns a noise estimate that is *lower* than the expected under the independence assumption. In addition, adding correlated noise was not detected by the goodness of fit procedure, as shown in Figure 4.6B. Thus, MLDS underestimate the noise when the independence assumption is violated, and this violation appears unnoticed to the goodness of fit diagnostics.

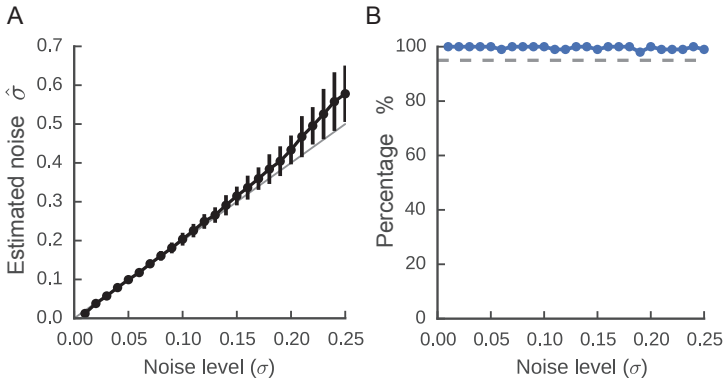


Figure 4.7: Results of simulated observers with an *absolute value decision rule*. (A) Estimated noise level $\hat{\sigma}$ as a function of the simulated noise σ. Errorbars indicate $M \pm$ S.D. across N=100 simulations. The gray line indicates the expected noise values. (B) Percentage of simulations with acceptable goodness of fit from panel (A). Horizontal dashed line indicates 95 %.

4.5.2 *Decision rule*

The decision rule for MLDS in its original formulation comprises the comparison of two perceptual intervals with an absolute value operation, and corrupted by Gaussian-noise (Equation 3.2). This rule is further simplified to a simple double difference rule by eliminating the absolute value operation (Equation 3.8), and thus allowing the scale estimation using a GLM (Section 3.2). However, in the presence of sensory representations that are noisy, the original decision rule could still be valid, representing the direct comparison of intervals (as comparing their absolute length)

Absolute value decision rule

$$\Delta = |\psi_{x_3} - \psi_{x_2}| - |\psi_{x_2} - \psi_{x_1}|$$

This rule is mathematically equivalent to Equation 3.8 when the function $\Psi(x)$ is deterministic and monotonically increasing, and thus differences inside the absolute value operation are always positive. However, since we modeled sensory functions that are draws from a random (Gaussian) distribution, the use

of the absolute value operation could have a different outcome. Differences can occur when the noise is large enough to produce intervals that are negative, i.e. when $\psi_{x_2} - \psi_{x_1} < 0$ and thus $\psi_{x_2} - \psi_{x_1} \neq |\psi_{x_2} - \psi_{x_1}|$. Using simulations we can quantify the effect of the change on the decision rule in MLDS estimation.

The results of these simulations are shown in Figure 4.7A. The estimated noise by MLDS ($\hat{\sigma}$, y-axis) is plotted against the noise introduced in the simulation (σ, x-axis) when the absolute value rule is applied. The estimated noise by MLDS (errorbars) was consistently higher than the expected values (gray line) at high noise levels. Thus, MLDS overestimate the noise in the presence of the absolute value rule, but only when at large noise values (approx. larger than $\hat{\sigma} = 0.4$). This upper limit is much higher than noise values observed experimentally in human observers. The goodness of fit diagnostics of MLDS also did not detect the change in the decision rule (Figure 4.7B).

4.6 SUMMARY

Before the testing of MLDS in experiments was carried out, the accuracy (bias) and precision (variability) of the MLDS estimation was quantified using simulations. MLDS could reliably estimate the generative perceptual function with low bias, providing at the same time a reliable estimate of the noise in the decision model when the equal-variance assumption holds. In addition, MLDS provided confidence intervals that are stable in precision after approx. $N = 1000$ trials, and with an adequate coverage that was independent of the simulated noise level.

In the case of unequal-variance noise distribution, MLDS could reliably recover the shape of the function. The distribution of the noise, e.g. equal- or unequal-variance, cannot be measured experimentally, therefore it is critical for MLDS to be robust against different noise distributions. The goodness of fit procedure did not detect this model violation, and the estimated noise is non-indicative of the underlying unequal-variance distribution.

Violations of other model assumptions were also studied, by introducing correlated noise, violating the independence assumption, and changing the decision rule operation from a simple difference to an absolute value operation. These vi-

olations also cannot be tested independently in experiments, and therefore it is relevant to show that MLDS could be robust against them.

The results of these simulations established the validity of MLDS for estimating perceptual scales. They also informed some practical decisions that must be taken during preparation of experiments, such as the stimulus spacing or number of trials, by considering the accuracy and precision of the method obtained from these simulations. In the part that follows (Part II) I present a study that shows how MLDS was used in the lightness domain, by measuring perceptual scales in a scenario of expected lightness constancy. In the study presented in Part III the possibility of using MLDS to derive sensitivity, traditionally done with performance-based methods, is explored in depth using simulations as well as experimental testing in a slant-from-texture task. Chapter 13 provides a general discussion of the results of both studies, and an overall evaluation of MLDS.

Part II

USING MLDS TO MEASURE APPEARANCE

This part has been published in:

Wiebel C.B.*, Aguilar G.*, Maertens M. (2017). Maximum likelihood difference scales represent perceptual magnitudes and predict appearance matches. *Journal of Vision, 17*(4):1, 1-14. doi:10.1167/17.4.1 (postprint)

*: *equal contribution*

5

INTRODUCTION

One major objective in the scientific study of perception is to understand how psychological experiences are linked to physical variables in the world (Fechner, 1860). Devising proper methods to quantify this relationship has turned out to be challenging, because psychological variables, contrary to physical ones, cannot be observed directly, but must be inferred from observers' responses to properly chosen stimuli (e.g. Gescheider, 1988). In the absence of a well-established measurement theory (Krantz et al., 1971), Fechner's simple method of adjustment (matching) is hard to beat and remains widely used (Koenderink, 2013).

To illustrate the problem let's say we are interested in the perceived lightness of the target check (Fig. 5.1A, red outline) presented behind a transparent medium. Introducing a transparent medium between a surface and the observer (Fig. 5.1B) changes the mapping between surface reflectance and retinal luminance in a characteristic way (Fig. 5.1C). The luminance range of surfaces seen through a transparent medium is substantially reduced and potentially shifted relative to the luminance range for surfaces seen in plain view. To be invariant against such changes the visual system has to "undo" these changes by appropriate computations (e.g. Singh & Anderson, 2002; Singh, 2004; Wiebel, Singh, & Maertens, 2016). This approximate invariance of perceived lightness across varying luminance is known as lightness constancy. We know from experience and from empirical studies that human observers are indeed largely invariant against such fluctuations in retinal luminance. However, we still lack a theoretical model of how the visual system accomplishes lightness constancy. To develop such a model, we must be able to measure the relationship between retinal luminance and perceived lightness in a reliable and comprehensive way. To that end, we ideally want to estimate the functions describing this relationship, which are known as transducer functions or perceptual scales (e.g. Kingdom & Prins, 2010).

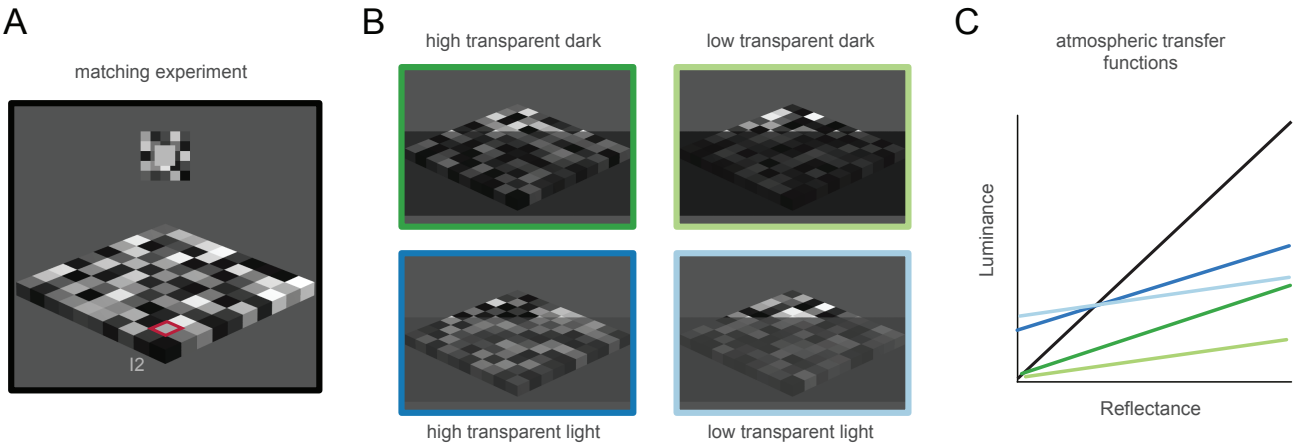


Figure 5.1: Experimental stimuli. A. The basic stimulus is a 10x10 checkerboard composed of checks with 13 possible reflectance values. In an asymmetric matching task observers adjust the luminance of an external test field so that is matches the perceived lightness of a specified target check (here $I2$). Observers are said to be lightness constant when their matches indicate the inversion of the various reflectance-to-luminance mappings that are introduced by different transparent media (see B and C). B. Checkerboards were also presented behind different transparent media that varied in reflectance (dark and light) and in transmittance (high and low). C. Atmospheric transfer functions (ATFs) relate target reflectance (x-axis) to target luminance (y-axis) (Adelson, 2000). The color scheme corresponds to the images in B. In the transparency conditions the luminance range is compressed and/or shifted with respect to plain view. This is reflected in corresponding slope and intercept changes of the respective ATFs.

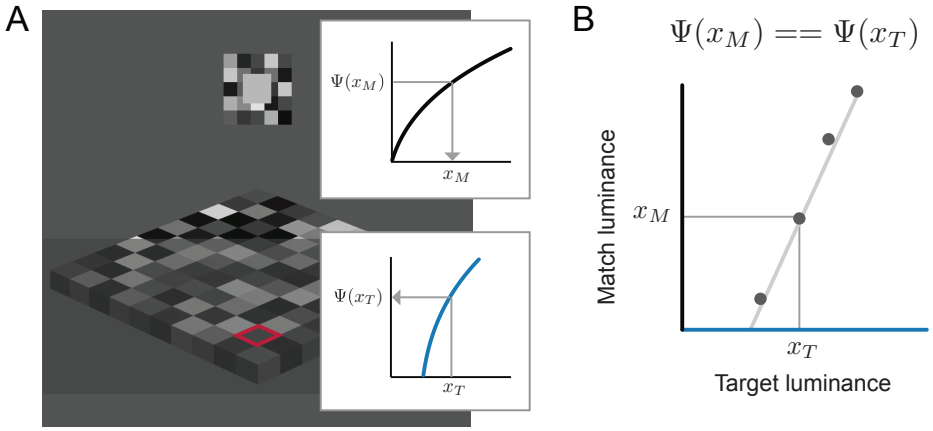


Figure 5.2: Perceptual processes underlying matching procedures. (A) At each position, match and target, there is a transducer function that relates retinal luminance (x_M, x_T) to perceived lightness ($\Psi(x_M), \Psi(x_T)$, insets on the stimulus). (B) What is measured in a matching procedure are the luminances x_M and x_T that correspond to equal perceived lightness at both positions ($\Psi(x_M) == \Psi(x_T)$). After Maertens and Wichmann (2013).

The most commonly used method for measuring this relationship is the method of adjustment, even though it does not provide a direct estimate of the transducer functions and it presumes a number of operations on the part of the observer. Figure 5.2 depicts the processes involved in adjustment or matching procedures for perceived lightness. An observer adjusts the intensity of a test stimulus so that it looks identical to a given standard. It is assumed that the observer internally compares magnitudes of perceived lightness for the target ($\Psi(x_T)$) and the match ($\Psi(x_M)$). What is being measured though, are not the transducer functions relating the two, but the corresponding luminances of the target and the match (x_T and x_M, Fig. 5.2B).

Another problem with the method arises when, as in the above case, test and match are presented in different contexts (asymmetric matching). In most cases researchers are interested in such asymmetric comparisons because they allow

one to quantify the degree of perceptual constancy. Such asymmetric comparisons become problematic, however, when the difference in context causes appearance differences that cannot be compensated along the dimension of the adjustment (Brainard, Brunt, & Speigle, 1997; Ekroll & Faul, 2013; Foster, 2003; Logvinenko & Maloney, 2006; Logvinenko, Petrini, & Maloney, 2008). The consequence would be an inaccurate or even invalid measurement that does not capture the perceptual representation of the stimulus.

Recently, there have been attempts to tackle the problems associated with matching (Logvinenko & Maloney, 2006; Logvinenko et al., 2008; Radonjic, Cottaris, & Brainard, 2015b; Radonjić & Brainard, 2016; Umbach, 2013). While it is widely accepted that observers are relatively lightness constant under natural viewing conditions, many experiments still find varying amounts of constancy for different viewing conditions, stimuli, task types or even instructions (Foster, 2011; Gilchrist et al., 1999). Such deviations might either be a consequence of methodological problems like the ones just outlined, or a meaningful deviation from constancy which then would need to be explained by any successful lightness model. Progress in revealing the underlying mechanisms for lightness perception is therefore tightly coupled with choosing appropriate and robust experimental methods that allow the comprehensive testing of theoretical models.

As an effort in this direction we address the limitations of matching procedures by adopting the following approach. We measure the transducer functions directly using Maximum-Likelihood Difference Scaling (MLDS, Maloney & Yang, 2003). MLDS is a scaling method that allows the efficient estimation of perceptual scales, i.e. the transducer functions relating retinal luminance and perceived lightness (Fig. 5.1). It has been used to study various perceptual dimensions (e.g. Obein, Knoblauch, & Viénot, 2004; Fleming, Jäkel, & Maloney, 2011). Furthermore, it is based on a signal detection model which potentially allows one to relate measurements of appearance with measurements of discriminability (Devinck & Knoblauch, 2012; Aguilar et al., 2017). Here we used MLDS to measure perceptual scales in different contexts using only within-context comparisons in order to avoid the procedural problems of asymmetric matching. The estimated scales are constructed from the judgment of perceived stimulus differences and not from the adjustment of a reference as in other scaling methods such as mag-

nitude estimation (Gescheider, 1988) or partition scaling (Whittle, 1994). MLDS requires a straightforward perceptual judgment and is thus less susceptible to strategic influences.

To scrutinize whether MLDS provides reliable perceptual scales of lightness we validate the scales empirically and theoretically. First, we use the estimated scales to predict perceptual matches and compare them to matches gathered in an independent asymmetric matching experiment. Second, we compare the predictive power of a contrast-based lightness model (Zeiner & Maertens, 2014; Wiebel et al., 2016) for scaling and matching data. To anticipate, we found that (a) the empirical perceptual scales for different contexts were consistent with lightness constancy, (b) matching data were well predicted by the perceptual scales, and (c) human lightness perception followed a difference scale that corresponds to a normalized contrast metric. The predictive power of the contrast-based lightness model was higher for the scaling than for the matching data, suggesting that estimating perceptual scales has the advantage of probing more directly the internal dimension under study.

6 METHODS

6.1 OBSERVERS

Ten naïve observers participated in the study, five of them were female. Observers' age ranged from 19 to 32 years. All observers had normal or corrected to normal visual ability and were reimbursed for participation. Informed written consent was given by all observers prior to the experiment.

6.2 STIMULI AND APPARATUS

Stimuli were presented on a linearized 21-inch Siemens SMM2106LS monitor (400 x 300mm, 1024 x 768px, 130Hz). Presentation was controlled by a DataPixx toolbox (Vpixx Technologies, Inc., Saint-Bruno, QC, Canada) and custom presentation software (`http://github.com/TUBvision/hrl`). Observers were seated 110 cm away from the screen in a dark experimental cabin. Observers' responses were registered with a ResponsePixx button-box (VPixxTechnologies, Inc., Saint-Bruno, QC, Canada).

The stimuli were images of customized checkerboards composed of 10 x 10 checks (Fig. 5.1). The images were rendered using Povray (Persistence of Vision Raytracer Pty. Ltd., Williamstown, Victoria, Australia, 2004). The position of the checkerboard, the light source and the camera were kept constant across all images. Checks were assigned one out of thirteen surface reflectance values according to the experimental design (see below). In the transparency conditions, a transparent layer was placed between the checkerboard and the camera (Fig. 5.1B). It was positioned so as to cover all target and their surrounding checks in both the MLDS and the matching experiment. The transparency was created using alpha blending (Metelli's episcotister model). The image luminances of the background B and the foreground F are combined according to some weighting

factor α so as to result in a new image luminance at the position of transparency $T = \alpha \times B + (1 - \alpha) \times F$. An α value of 0 corresponds to an opaque foreground $T = F$, α of 1 corresponds to a fully transparent foreground $T = B$. The transparent layer varied in transmittance and reflectance. The dark transparency had a value of 0.35 in *povray* reflectance units ($19cd/m^2$) and the light transparency of 2 ($110cd/m^2$). A value of $\alpha = 0.4$ and 0.2 was used in the high and low transmittance condition respectively. The rendered images were converted to grayscale images. The background luminance was $141cd/m^2$. Detailed values of luminance for each transparent medium can be found in Appendix Table A.4).

In the matching experiment, an adjustable test field was presented above the checkerboard to assess observers' lightness matches (Fig. 5.1A). The test field was embedded in a coplanar surround checkerboard that was composed of 5x5 checks. The size of the test field was 1.2x1.2 degrees visual angle and that of the surround checkerboard was 3x3 degrees. The luminances of the checks in the surround checkerboard were fixed throughout the experiment and the luminances were chosen so that two adjacent checks did not have the same luminance. The mean luminance of the surround checks was $178cd/m^2$, which is identical to the mean luminance of the 13 checks in the main checkerboard in plain view. The surround checkerboard was presented in four different spatial arrangements resulting from clockwise rotation of the original in steps of 90 degree. A configuration was assigned randomly to each trial.

6.3 DESIGN AND PROCEDURE

Perceptual scales and asymmetric matching functions were measured for five different viewing conditions, a plain view condition and four transparency conditions (Fig. 5.1).

6.3.1 *MLDS experiment*

We used MLDS with the methods of triads (Figure 6.1A, Knoblauch & Maloney, 2008, 2012). We used ten out of the 13 reflectance values to construct the triads.

The lowest and the two highest reflectance values were omitted to achieve a feasible number of trials. With $p = 10$ reflectance values the total number of unique triads was $n = p!/((p-3)! \times 3!) = 10!/(7! \times 3!) = 120$. Each triad contained three values that were selected so as to enclose non-overlapping intervals. They were presented in ascending ($x_1 < x_2 < x_3$) or descending ($x_1 > x_2 > x_3$) order (Knoblauch & Maloney, 2008). The reference, x_2 (check *I2* in Fig. 6.1A), was located between the two comparisons, x_1 and x_3 (checks *B2* and *I9* in Fig. 6.1A). In each trial observers judged which comparison check x_1 or x_3 was more different in lightness from the reference. Observers used a left or right response button to indicate their choices. No time limit was imposed.

To keep the local context comparable for the elements of a triad we controlled the luminances of the eight checks surrounding each triad element. The same eight luminance values were used for each triad element but they differed in spatial arrangement. Their mean luminance was $178 cd/m^2$ which was identical to the mean luminance of all checks seen in plain view. The remaining 73 checks were drawn randomly without replacement from a set consisting of six repeats of the 13 different reflectance values. This resulted in a slight variation of the mean luminance of those checks between trials (up to $6 cd/m^2$). The checks were positioned so that two neighboring checks did not have the same reflectance.

Each triad was repeated 10 times resulting in 1200 trials per viewing condition and 6000 trials in total. Trials were randomized across viewing condition, triad and target reflectance. The experiment was divided into several sessions. A new image was created for each trial.

6.3.2 *Matching experiment*

Target reflectances and viewing conditions were identical to those in the MLDS experiment. The target check was presented at the position of the reference (check *I2* in Figure 5.1) in the MLDS experiment. Observers adjusted the luminance of the external test field to match the perceived lightness of the target check. The luminance was adjusted by pressing one of four buttons, two of them for coarse adjustments ($\pm 10 cd/m^2$) and the other two for fine adjustments ($\pm 1 cd/m^2$). The maximum luminance of the monitor was $550 cd/m^2$. Satisfactory matches were

confirmed with a fifth button which initiated the next trial. No time limit was imposed on the adjustment procedure.

The eight checks surrounding the target were assigned in the same way as in the MLDS experiment. The remaining 91 check reflectances were drawn randomly without replacement from a set consisting of eight repeats of all 13 reflectance values. Thus the mean luminance across trials was comparable to that in the MLDS experiment. Again, neighboring checks had to have different reflectances.

Each combination of target reflectance and viewing condition was repeated 10 times resulting in a total of 500 trials. A new image was created for each trial and trials were randomized across experimental conditions. [1]

6.4 SIMULATION OF OBSERVER MODELS

We used an ideal observer analysis to test whether MLDS could distinguish between different generative models. In particular, we tested a lightness-constant against a luminance-based observer, two extremes of behavioral judgments. The model comparison is done as follows. We define internal scales for each of the two models (Fig. 6.1B). For a luminance-based observer the luminance-to-lightness mappings in different contexts coincide on a single function and differ only in the range of luminance values (Fig. 6.1B lower left panel). Formally, the sensory representation function was defined as

$$\Psi^{lum}(x) = a \cdot x + b$$

where x is luminance, and a, b are linear coefficients calculated to map the range of luminance in plain view $[L_{min}, L_{max}]$ to the range $[0, 1]$.

For a lightness-constant observer the mapping functions in different contexts should 'undo' the transformations of image formation in which equal surface reflectances are mapped onto different luminance ranges (Fig 5.1C). Thus, we model this observer by using internal mapping functions that are the inverse

1 Section "MLDS analysis" from the published paper was omitted here to avoid redundancy with Chapter 3 of this thesis.

functions of the ATFs shown in Figure 5.1C. Formally, the sensory representation function was defined as

$$\Psi^{light}(x) = a_i \cdot x + b_i \qquad i \in 1...5$$

where x is luminance, and a_i, b_i are linear coefficients calculated to map the range of luminance for each viewing condition to the range $[0, 1]$ (for simplicity we used linear functions, but power functions could be used as well and would not change our ideal observer results).

Each of the two observer models is used to generate responses in a 'mock' MLDS experiment that has the same number of triads and repetitions as the actual experiment. For each triad and repetition, the decision variable was calculated as

$$\Delta = [\Psi^*(x_3) - \Psi^*(x_2)] - [\Psi^*(x_2) - \Psi^*(x_1)] + \epsilon \tag{6.1}$$

with $\epsilon \sim N(0, \sigma^2)$, and Ψ^* is either Ψ^{lum} or Ψ^{light}. Simulated responses were generated choosing the triad (x_2, x_3) when $\Delta > 0$ and (x_1, x_2) otherwise. Finally, the simulated data were subjected to the MLDS analysis to obtain the coefficients β that constitute the scale values. Figure 6.1B shows the model perceptual scales (left) and the estimated scales (right), and it is evident that for the chosen noise level ($\sigma = 0.15$) the method recovers the underlying scale.

We repeated the ideal observer analysis for a range of different noise levels (σ, minimum = 0.01 and maximum 1.2, see Appendix A.1). The two observer models were distinguishable for a broad range of noise levels up to approximately 0.4. This upper-bound value was much higher than the noise levels that have been observed in previous experiments (Knoblauch & Maloney, 2008; Devinck & Knoblauch, 2012). We therefore concluded that MLDS could be used to derive meaningful scales because they would allow us to distinguish between these two different observer models.

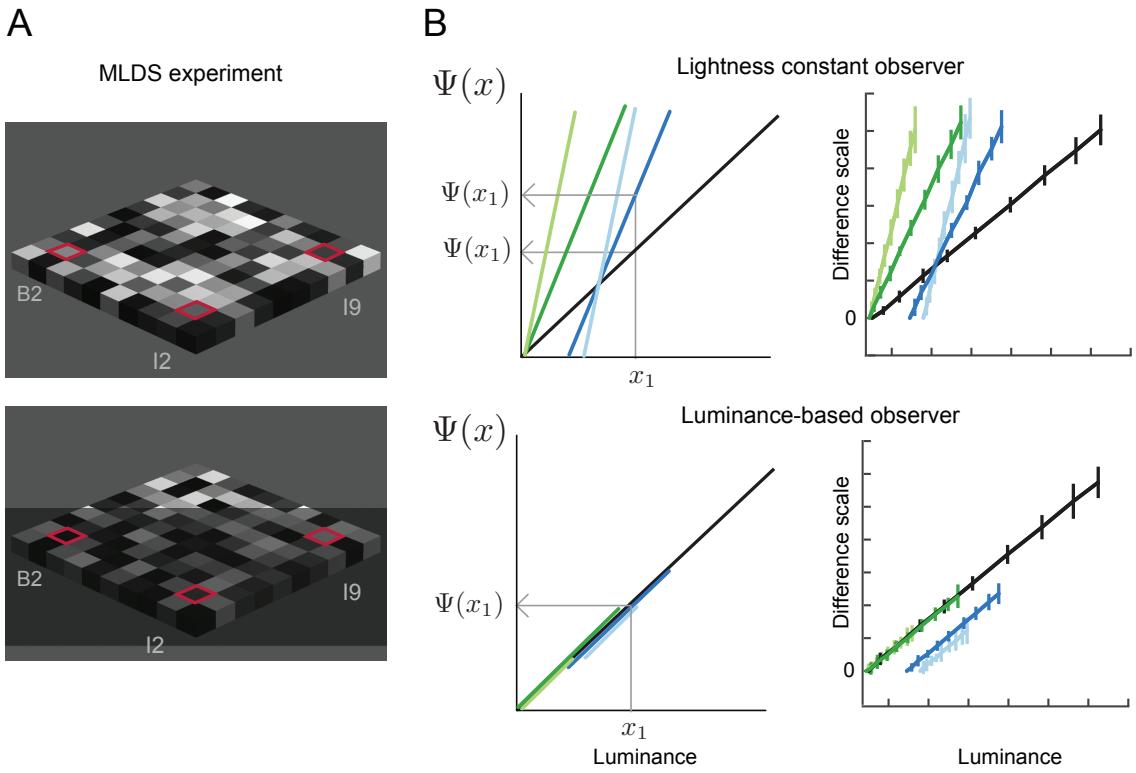


Figure 6.1: Method of triad procedure and observer models. A. In the triad comparison observers compared the lightness of three specified checks (*B2*, *I2* and *I9*, marked with a red outline). The upper panel shows a triad comparison in plain view, the lower panel a comparison behind one of the transparent media. B. Simulation for a lightness constant (upper panels) and a luminance-based observer (lower panels). For the lightness constant observer the perceptual scales (upper left panel) correspond to an inverse mapping of the atmospheric transfer functions (Fig. 5.1C). For the luminance-based observer (lower left panel) the luminance-to-lightness mappings in different contexts coincide on a single function. We generated data for each of the models in simulations, and the estimated perceptual scales are shown on the right panels. See text for details.

7

RESULTS

Figure 7.1 shows the perceptual scales measured in different viewing conditions aggregated across all observers. The scales are interval scales with the minimum anchored at zero and the maximum being inversely proportional to the estimated noise (in MLDS terminology referred to as 'unconstrained scales'; Knoblauch & Maloney, 2012).

The empirical scales are consistent with a lightness-constant observer and not with a luminance-based observer. This is evident from a comparison between the model predictions (Fig. 6.1B) and the observed result pattern (Fig. 7.1A). Although the estimated scales are not linear they share crucial features with the hypothetical scales. First, there is a difference in 'intercept' between perceptual scales in the light and dark transparency conditions (blue vs. green lines in Fig. 7.1A). Second, the scales are steeper for transparent media with lower transmittance than with higher transmittance (light vs dark colored lines in Fig. 7.1A). Figure 7.1A also plots the Munsell neutral value scale (Munsell et al., 1933) that would be predicted for our choice of luminances (dashed black line in Fig. 7.1A).

The Munsell scale represents the expected scale that relates equal steps in perceived lightness to luminance (Whittle, 1994). It was calculated by setting the highest luminance in the plain view stimulus as the white reference, i.e. to the maximum of one (Pauli, 1976). It is evident from Figure 7.1A that the Munsell scale is consistent with the perceptual scale estimated in our plain view condition. The typical nonlinear shape indicates higher sensitivity for differences between checks of low reflectances than for checks with high reflectances. This has indeed been reported in previous work (e.g. Chubb, Landy, & Econopouly, 2004). To aggregate scales across observers we normalized the scales of each individual observer relative to the maximum scale value in the plain view condition. The ranges of the scales differed between observers because different observers

have different noise levels. The data for individual observers are provided in Appendix Figure A.1.

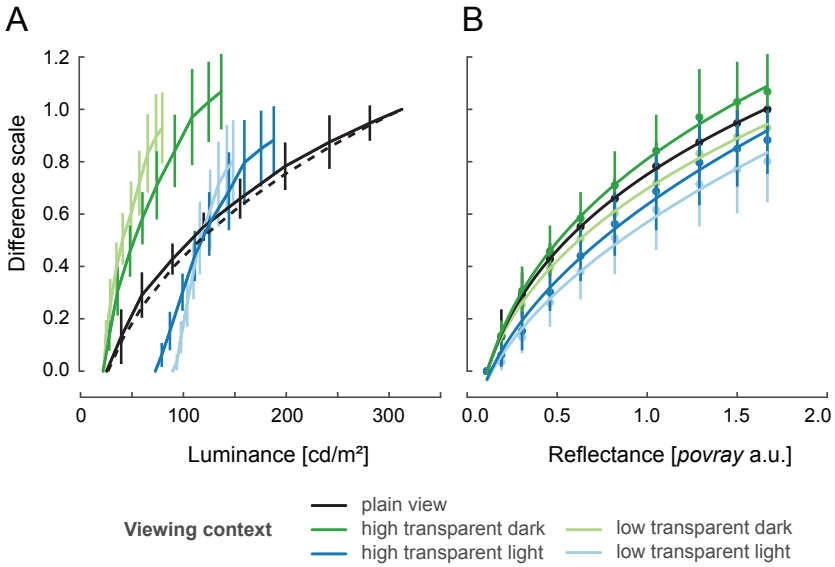


Figure 7.1: MLDS difference scales in different viewing conditions. A. Difference scales as a function of luminance. The functions depict the aggregated scales across observers (n=10). For each observer the scales were normalized with respect to plain view, and then aggregated. The dashed black line depicts the Munsell scale in plain view (see main text for a description of the Munsell scale). Error bars indicate M $\pm$ S.D. B. Same as in A. but scales are plotted as a function of reflectance. Scale values (markers, M $\pm$ S.D.) were fitted with a power function (lines) individually for each viewing condition.

7.1 SCALES AS A FUNCTION OF REFLECTANCE

To better illustrate the degree of lightness constancy across conditions we replaced luminance by reflectance at the x-axis of the perceptual scales. In such a perceived lightness vs. reflectance plot the scales of a lightness-constant observer

should coincide on a single function. Figure 7.1B shows that this was indeed that case.

To assess the agreement between scales in different conditions quantitatively we compared the functions that were fit in each condition against what we call a global fit in which the data from all conditions are fitted by a single function. If the data in different viewing conditions can be explained by one internal model then the global fit should account for the data as well as the individual fits for each viewing condition. We fitted the scale parameters in each condition and the global scale with a power function $\Psi(x) = ax^e + b$ using a nonlinear least squares method (Ritz & Streibig, 2008). To evaluate the goodness of fit we computed R^2 values for linear fits to the data. The average R^2 was already reasonably high (0.86). We then performed F-tests on nested models (power function vs. its linear submodel with $e = 1$) which revealed that the power functions fitted the data significantly better than the linear ones ($F_{min}(1, 97) = 15.6$, $p < 0.001$). From this we conclude that the power functions captured the data sufficiently well.

We used a general non-linear model to test whether applying single models to the data in the five different viewing conditions would result in better fits than applying a global model to all data. We compared the respective sum of squares for the global model with three parameters (a, b, e) and for the separate models with five times three parameters. There was a benefit for the separate model fits relative to the global model ($F(12, 497) = 18.57$, $p < 0.001$). To explore the cause for this difference we computed one-way repeated measures ANOVAs for each of the three parameters of the power functions. We found a significant difference between scales for the exponent parameter, e ($F(4, 36) = 16.6$, $p < 0.001$) which determines the curvature of the function. Posthoc tests on the exponents revealed significant differences between each of the light transparency conditions and the plain view and the dark transparency with high transmittance (Bonferroni corrected $p < 0.05$). The main difference between the light transparency conditions and the plain view and the dark transparency (high transmittance) conditions is the difference in curvature between these functions (Fig. 7.1B).

The light transparency conditions are special insofar as during image formation the reflectance-to-luminance mapping undergoes a range reduction and a range shift (see Fig. 5.1). This means that checks seen through a light transparent

medium undergo the greatest compression in its contrast range. The Michelson contrast for targets in plain view range from -0.84 to 0.4, whereas in the low transparent light condition range from -0.16 to 0.16 (the contrast is computed relative to the mean luminance in the region of transparency). Therefore, sensitivity might be lower for this range of the stimuli.

7.2 PERCEPTUAL SCALES AND MATCHING FUNCTIONS

We illustrated in Figure 5.2 how the data recorded in matching procedures are related to perceptual scales. Here we show to what extent the theoretical relationship can be corroborated by experimental data. To predict matching data from perceptual scales, one needs to first find the scale value $\Psi(x_T)$ that corresponds to a particular target luminance x_T in one of the transparency conditions. In the next step we need to find the luminance value x_M that corresponds to the scale value at the match position ($\Psi(x_M)$) assuming that observers match the lightness of the match region to that of the target region according to $\Psi(x_M) == \Psi(x_T)$. We did not measure a perceptual scale at the match position but instead adopt the plain view scale to represent the scale for the matches. In order to be able to read out x-values corresponding to any possible Ψ-value and *vice versa* we fitted the scales with power functions ($\psi(x) = ax^e + b$) using a non-linear least squares method. We derived the predicted matching data from the 'unconstrained' scales individually for each observer, and we then aggregated them in the same way as the empirical data obtained from the matching experiment.

In Figure 7.2 empirical and predicted matches are plotted next to each other (panels A and B, respectively) and it can be seen that they share some characteristic features. The matching functions, like the scales (Fig. 7.1A), differ in slope and intercept between the different transparency conditions. Differences in transmittance are accompanied by differences in slope and differences in reflectance are accompanied by differences in intercept. Unlike the scales the matching functions are linear.

For a quantitative evaluation of the degree of similarity between empirical and predicted matching data we computed linear regressions for each of the viewing conditions. We used within-subject t-tests to compare slopes and intercepts be-

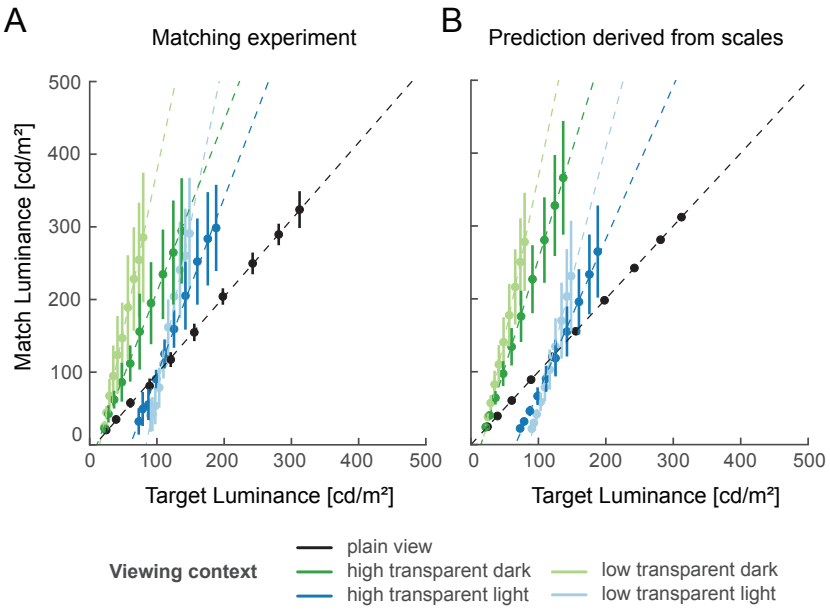


Figure 7.2: Empirical and predicted matching data. A. Results of the matching experiment. The luminance adjusted in the matching field (y-axis) is plotted as a function of target luminance (x-axis) in each viewing conditions. Data were aggregated across observers (n=10). Error bars indicate mean ± S.D. (B) Same as in (A) but for matches predicted from the estimated MLDS scales.

tween predicted and empirical functions. The average slope and intercept values are listed in Appendix Table A.3 together with the relevant test statistics. We found significant differences between the predicted and the empirical functions only for the dark transparent medium with a high transmittance.

7.3 PREDICTIVE POWER OF A CONTRAST-BASED MODEL

The estimated perceptual scales are an interesting test case for lightness models because they represent a more direct measurement of perceived lightness than the matching data. In particular we compared how well our previously suggested normalized-contrast model (Zeiner & Maertens, 2014) could account for both the scaling and the matching data.

The normalized contrast model was initially motivated by the observation that the introduction of a transparent medium leads to a systematic change in contrast range of the respective image region. It was suggested that this change in contrast range might serve as a cue to segregate the region from regions seen in plain view (Anderson, 1999; Singh & Anderson, 2002; Singh, 2004). It has been subsequently shown that the accompanying contrast statistics can be used to accurately predict perceived lightness (Singh & Anderson, 2002; Singh, 2004; Zeiner & Maertens, 2014; Wiebel et al., 2016). The normalized contrast model engages two processing steps: first the target intensity is normalized relative to its local surround by computing the Michelson contrast between target and surround. Second this target contrast is normalized relative to the contrast range in the region of the transparency which is subsequently mapped to the contrast range in plain view (for details of the normalized contrast model calculation see Appendix). The so derived normalized contrast predicts observers' lightness matches in contrast units.

Figure 7.3 shows the aggregated data of both experiments as a function of the model predictions. If the computed normalized contrast accounts well for differences in appearance then the functions should line up on top of each other and they should become more linear (see Knoblauch & Maloney, 2012 for a similar rationale underlying correlation perception). Transforming the x-axis into units of normalized Michelson contrast did indeed linearize the perceptual scales. To

test how well the normalized contrast model accounts for the variability between the different context conditions, we computed a global R^2 value. As described before we treat all data as if they were coming from one underlying model. The normalized contrast measure accounts for 98% of the variance in the scaling data and for 88% of the variance in the matching data. This indicates that the normalized contrast measure is a better predictor for the scales than to the matching data by explaining more variance.

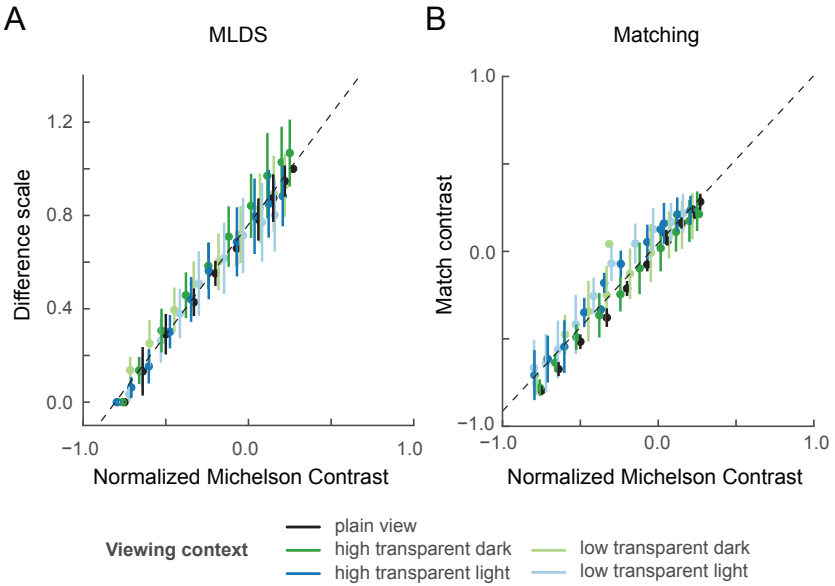


Figure 7.3: Perceptual scales (Fig. 7.1A) and matching data (Fig. 7.2A) plotted as a function of the Normalized Michelson contrast. Dashed lines indicate a linear fit to the data for all viewing contexts ($R^2 = 0.98$ for MLDS, $R^2 = 0.88$ for matching). Error bars indicate mean $\pm$ S.D. across observers.

8

DISCUSSION

The goal of this work was to better understand how psychological experiences are linked to physical variables. We studied the question in the domain of lightness perception but the observed principles equally apply to other domains of perceptual appearance. To make progress towards that goal we measured perceptual scales that link perceived lightness to image luminance using MLDS. Our results show that the estimated perceptual scales (a) are consistent with a lightness-constant observer model in all viewing contexts, (b) predict perceptual equality across different viewing contexts, (c)indicate that human lightness perception follows a difference scale that corresponds to a normalized contrast metric. The normalized contrast model accounted for more of the variance in the scaling (98%) than in the matching data (88%), suggesting that estimating perceptual scales has the advantage of probing more directly the internal lightness scale.

8.1 MLDS-BASED LIGHTNESS SCALES

The estimated perceptual scales were in close correspondence with each other (Fig. 7.1B), i.e. perceived lightness followed the actual check reflectances despite substantial variations in check luminance across viewing conditions. This reflects a high degree of lightness constancy. This was corroborated by the simulated observer models, because the empirical scales were consistent with the lightness-constant and not the luminance-based observer. The shape of the perceptual scales followed the shape of the classical Munsell scale. The perceptual scales are an estimation of the transducer functions which cannot be uncovered using matching (see Fig. 5.2).

In addition to the MLDS experiment we conducted a conventional asymmetric matching experiment. We tested to what extent the postulated relationship between perceptual scales and matching (Fig. 5.2) would be evident in the data.

Predicted and empirical matching functions were consistent with each other (Fig. 7.2). The high degree of consistency is noteworthy because triad comparisons and matching require different perceptual judgments. Asymmetric matching can be likened to measuring a rod of unknown length with a ruler whereas in triads rods of different lengths would be compared among each other. The consistency between both types of measurements indicates that the stimulus suitably constrains the perceptual response to judgments based on lightness, and not on luminance. This cannot be taken for granted (Arend & Goldstein, 1987; Radonjić & Brainard, 2016) in particular since observers were not explicitly told what dimension to judge.

A potential challenge when comparing perceptual scales measured in different contexts is the necessary assumption of how scales are anchored. Two perceptual scales might have the same shape but cover a different range, implying a different anchoring. Classical scaling experiments did not confront this problem because perceptual scales were measured in only one context, i.e. plain view. The default of MLDS is to anchor the perceptual scales at zero. This is an arbitrary choice and any linear transformation of the scale would be a valid outcome of the analysis. The good correspondence between the estimated scales and the matching data in the present case suggests that there was no substantial anchoring problem.

As described by Knoblauch and Maloney (2012), MLDS assumes that observers are stochastic in their judgments, with the noise originating at the decision level (as shown in Eq. 6.1). This assumption implies that observers are worse at judging interval differences that are small, i.e. when $[\Psi(x_3) - \Psi(x_2)] \sim [\Psi(x_2) - \Psi(x_1)]$. This critical assumption in MLDS is different than other scaling methods, such as Fechnerian scaling that uses integration of JNDs or other discrimination-based scaling methods (Baird, 1978). These scaling methods assume a noise source at an early sensory representation level, and not at a late decision level. Here we compared perceptual lightness scales that were measured in different viewing conditions, and hence could have been associated with different amounts of decision noise. This was not what we observed. Although individual observers differed in their overall noise level, all scales measured for one observer had comparable estimated noise levels. However, these assumptions must be consid-

ered carefully, and ultimately their validity must be addressed experimentally (Aguilar et al., 2017).

The estimated noise level is critical for the interpretation of scales, also with respect to the distinction between our two observer models (lightness constant vs. luminance-based). This is possible only up to a limit at which observers' noise is too large for the models to be distinguished. We established in simulation that this upper bound is at an estimated noise level $\hat{\sigma} = 0.4$ (Appendix A.1). In our observers the estimated $\hat{\sigma}$ values varied from 0.13 to 0.21 for observers O1 to O8, i.e. values below the upper limit of model discriminability. For observer O9 $\hat{\sigma} = 0.39$ was at the boundary of discriminability, and for observer O10 $\hat{\sigma} = 0.71$, was beyond the upper limit. Thus, the noise level of observer O10 did not allow a definite selection of either of the two models. The estimated noise level must be also considered carefully when comparing scales against ideal observer models.

8.2 ALTERNATIVES TO ASYMMETRIC MATCHING

Asymmetric matching has been criticized in the past for mainly two reasons: First, observers' matches reflect the underlying perceptual magnitudes only indirectly (Gescheider, 1988; Maertens & Wichmann, 2013). Second, observers' matches might not reflect perceptual identity but merely the best possible match (Brainard et al., 1997; Ekroll & Faul, 2013; Foster, 2003; Logvinenko & Maloney, 2006). In particular, the question whether lightness is represented by more than one dimension across different contexts has been tackled using different methods (Logvinenko & Maloney, 2006; Logvinenko et al., 2008; Umbach, 2013). Beyond methodological shortcomings, asymmetric matching tasks have also been criticized for their lack of realism, because in real life we rarely adjust the color of an object but rather select objects based on their color. In two recent studies Radonjic et al. (2015b); Radonjic, Cottaris, and Brainard (2015a) measured color constancy in a color selection paradigm where they asked observers to select which of two competitors was more similar to a given target.

Their task was analogous to the triad comparison used in MLDS, but the design was different from the standard MLDS design. A limited number of targets was presented as anchor for a respective set of competitors but these competitors

were not compared with each other. MLDS would involve triad comparisons of all possible combinations of targets and competitors. The data were analyzed with a customized version of MLDS. The crucial difference to our approach is that in their critical condition target and competitors were presented in different illuminations. As a consequence, observers' judgments were subject to the same comparison problem as in asymmetric matching. To estimate a perceptual scale it was assumed that target and competitors are represented on a common underlying dimension. In our way of thinking this means to skip the step of estimating the different transducer functions (scales), which map luminance to perceived lightness in different contexts (Fig. 6.1B), and to compare stimuli directly on the internal axis. As we have outlined above this assumption is valid only for a lightness constant observer, i.e. for observers whose perceptual scales in different viewing situations have comparable scale maxima. The authors reported moderately high color constancy indices which were comparable to asymmetric matches for the same type of stimuli (Radonjic et al., 2015a, 2015b). We suggest to include such cross-context comparisons only to validate predictions from the MLDS-based scales as we did here with the asymmetric matches.

8.3 MODELS OF LIGHTNESS PERCEPTION

We claim that perceptual scales are an important test case for models of lightness perception because they offer a direct estimate of the transducer functions that we are interested in. A successful model should be able to explain both characteristics of lightness appearance: perceptual equality across contexts as well as sensitivity differences manifested in the shape of perceptual scales (Hillis & Brainard, 2007a).

If we assume that the goal of the visual system is to accurately represent surface reflectance, then reflectance would be the best predictor of perceived surface lightness. Thus, for a perfectly lightness-constant observer the perceptual scales measured in different contexts should perfectly overlap when plotted against reflectance. Our empirical scales are consistent with a lightness-constant observer, however they reveal small deviations, especially for the two lighter transparent media (Fig. 7.1B). When we plotted the scales as a function of normalized con-

trast (Wiebel et al., 2016; Zeiner & Maertens, 2014) instead of reflectance, the differences between scales were substantially reduced (Fig. 7.3A). This means that the normalized contrast metric does not perfectly capture *veridical* surface reflectances but is rather tightly correlated with them. One might be tempted to conclude that the predictive power of the contrast-based model 'exceeds' that of physical surface reflectances, because it accounts for the deviations from lightness constancy that we observed in the data.

This finding is consistent with the idea that the visual system, instead of doing inverse optics (e.g. Barrow & Tenenbaum, 1978; D'Zmura & Iverson, 1993), might use a set of readily available but imperfect cues to infer stable properties of objects (e.g. Anderson, 2011; Fleming, 2014) . The involved computations might not always lead to a *veridical* percept with respect to the physical world, but to an overall reliable estimate of the appearance of objects (e.g. Marlow, Kim, & Anderson, 2012). The estimated scales were linearized by the transformation to contrast units, which implies that the model accounts for the sensitivity differences between low and high reflectances (e.g. Lu & Sperling, 2012), a feature which cannot be quantitatively captured with matching. The higher agreement between the model and the perceptual scales (compared to matching) supports the idea that the perceptual scales are a more direct and informative measure of the internal variable of lightness and subject to fewer sources of variability.

8.4 GENERAL CONCLUSIONS

In this paper we show that a scaling method is more powerful than matching in elucidating the perceptual representation of surface lightness. MLDS provides a direct estimate of the transducer functions that relate the physical dimension of reflectance to the psychological dimension of perceived lightness. In addition, MLDS avoids the practical difficulties associated with asymmetric matching tasks because all perceptual comparisons are made within the same viewing context. Observers confirmed that subjectively the triad comparison required by MLDS was a natural and straightforward task.

So why is it then that asymmetric matching remains the method of choice despite the obvious benefits of MLDS. We suspect that experimenters feel slightly

uneasy about explicitly making and commiting to the various assumptions that are required by MLDS in order to statistically estimate the perceptual scales. However, as we illustrate in Figure 5.2, asymmetric matching procedures also assume the presence of internal scales but they are hidden and their shape can not be inferred from observers' matches. We think that the present results are encouraging and advocate the estimation of scales, because they provide a more direct estimate of internal variables against which we can test our theoretical models of appearance.

Part III

USING MLDS TO MEASURE SENSITIVITY

This part has been published in:

Aguilar G., Wichmann F. A., Maertens M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision, 17*(1):37, 1-18. doi:10.1167/17.1.37 (postprint)

9

INTRODUCTION

Maximum likelihood difference scaling (MLDS) is a psychophysical method that allows the efficient characterization of perceptual scales (Maloney & Yang, 2003; Knoblauch & Maloney, 2012). Observers are asked to judge appearance differences for supra-threshold stimuli that vary along some dimension of interest, and a scale is constructed based on the reported differences in appearance. The method has been used to study appearance in a variety of visual domains such as color differences (Maloney & Yang, 2003), texture properties (Emrith, Chantler, Green, Maloney, & Clarke, 2010), surface glossiness (Obein et al., 2004), transparency (Fleming et al., 2011) and material properties (Paulun, Kawabe, Nishida, & Fleming, 2015), as well as for the assessment of perceived image quality in compression-degraded images (Charrier, Maloney, Cherifi, & Knoblauch, 2007).

Recently, MLDS has been used to link stimulus appearance with stimulus discriminability. Assuming an underlying signal detection model Devinck and Knoblauch (2012) have demonstrated a quantitative agreement between sensitivity estimates derived from perceptual scales (MLDS) and sensitivity estimates assessed with a traditional forced-choice procedure for the watercolor effect. Their finding is remarkable given the long effort in psychophysical research of relating discrimination and appearance in a unified framework.

Relating stimulus appearance — the stimulus' subjective magnitude — to discrimination — the ability to discriminate stimuli—dates back to the roots of psychophysical research. Fechner (1860) proposed that by summing equal subjective just-noticeable differences (JND) and assuming Weber's law, a function could be constructed which relates stimulus' subjective magnitude and physical magnitude (Baird, 1978). Soon Fechner's suggestion was criticized, theoretically as well as for lack of experimental evidence to support it (reviewed in detail in Krueger, 1989).

Stevens (1957, 1975) later proposed that subjective magnitude could be directly measured from observer responses to supra-threshold stimuli. He devised 'direct' methods to measure subjective magnitude and derived (power) functions that would relate subjective and physical magnitude (Stevens, 1975; Gescheider, 1997; but c.f. Treisman, 1964a, 1964b). However, Stevens' proposal was met with equal criticism, partly because of the scales' lack of predictive power for discriminability and partly because of the methodological concerns of asking observers to numerically estimate or provide ratings of perceived sensation (e.g. Baird, 1989). Although a considerable amount of work has been done trying to unify discrimination and appearance, so far the debate still continues and mixed experimental evidence has been found (e.g. Krueger, 1989; Ross, 1997; Hillis & Brainard, 2007b). Thus, the finding of Devinck and Knoblauch (2012) that appearance and sensitivity can be linked via MLDS is promising, because it suggests that a supra-threshold method like MLDS could be used to predict sensitivity to near-threshold stimulus differences. Apart from potential theoretical implications, Devinck and Knoblauch's finding may be beneficial from a purely methodological point of view, because MLDS requires a considerably smaller amount of data than traditional discrimination methods. Because of its efficiency MLDS could be used to identify experimental settings in which appearance and discrimination judgments are consistent, by comparing sensitivity measured in discrimination tasks (e.g. two-interval forced-choice) with sensitivity derived from MLDS. The goal of this work was to further explore - theoretically and empirically - the possibility to use MLDS to predict near-threshold discrimination performance using a slant-from-texture task.

9.1 SLANT-FROM-TEXTURE TASKS

We measure perceptual scales in a slant-from-texture experiment. The perceptual scale that relates apparent and physical slant in slant-from-texture tasks has a non-linear shape and it therefore provides an interesting test case for predicting sensitivity at different positions of the MLDS based scale. Slant-from-texture stimuli have been used extensively in the study of depth and surface perception (e.g. Knill, 1998; Rosas et al., 2004; Todd, Thaler, & Dijkstra, 2005; Velisavljević & Elder,

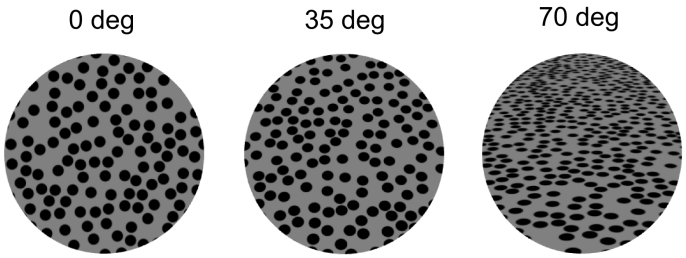


Figure 9.1: Example stimuli showing surfaces of different slants covered with the 'polka dots' texture. Here we used the method of triads for MLDS where observers judge which of the pairs exhibit a larger difference in perceived slant, the left-middle pair or the right-middle pair. Most observers would report that the right-middle pair (35, 70) contains the larger slant difference, although the physical slant difference is identical between the pairs: (0, 35) vs. (35, 70).

2006; Saunders & Backus, 2006), because texture gradients can evoke a strong impression of 3-D slant in the absence of other cues (Saunders, 2003). Stimuli are surfaces that are covered with a texture pattern such as randomly placed circular elements (or 'polka dots', Fig. 9.1). The surface is slanted at varying degrees relative to the fronto-parallel position resulting in characteristic changes in the polka dot patterns. The slanted texture is viewed through an aperture to isolate texture cues from other pictorial cues such as the shape and borders of the surface (Knill, 1998; Todd et al., 2010). Using this type of stimuli it has been found that sensitivity to slant is lower when the surface is close to the fronto-parallel position than when the surface is slanted away from it (Knill, 1998; Rosas et al., 2004). The difference in sensitivity between 0 and 70 deg can be up to ten-fold (Knill, 1998).

9.2 MLDS AND THE SIGNAL DETECTION MODEL

The decision model underlying the MLDS framework is depicted in Figure 3.2A. It is assumed that different stimulus levels x_i are associated with discrete perceptual responses Ψ_{x_i}, and that observers compare different stimuli by judging

the differences between the perceptual responses. The decision variable is assumed to be corrupted by decision noise, ϵ, which is assumed to be Gaussian distributed with zero mean and variance σ^2. MLDS estimates the perceptual scale together with the noise associated with the judgments (Maloney & Yang, 2003; Knoblauch & Maloney, 2008).

The same perceptual process can be rephrased in a signal detection framework by shifting the noise from the decision process to the sensory representation (see Figure 3.2B). In this way, the original MLDS scale can be transformed to a normed scale in which the units on the perceptual axis represent differences in units of d'. This transformation has been suggested by Devinck and Knoblauch (2012) to compare supra- and near-threshold judgments in the watercolor effect. A detailed description of the transformation and the MLDS model is provided in Section 3.3.

In order to apply this transformation the following assumptions are made: 1. The sensory representations associated with each stimulus level are Gaussian random variables with equal variance (σ^2). 2. They are independent. 3. The decision process is deterministic. 4. The sensory representation function is monotonically increasing. This produces only positive values of sensory response intervals so that the absolute value operation can be removed from the decision rule (Δ variable in Figs. 3.2A and B). An MLDS decision model with the above assumptions is equivalent to a signal detection model with equal-variance and Gaussian distributed sensory representations, as depicted in Fig.3.2C [1].

9.3 OBJECTIVES

We want to test whether and to what extent we can assume the equivalence of MLDS and forced-choice procedures for estimating sensitivity as it was reported by Devinck and Knoblauch (2012) for the Watercolor effect. We first examine the theoretical equivalence between both methods by means of simulations. We use a known observer model to generate sensitivity estimates for both methods. In

1 Which, in turn, is analogous to Thurstone's case V of the Law of Comparative Judgment (Thurstone, 1927b)

the present analysis we evaluate the adequacy of MLDS to predict sensitivity using the two-alternative forced-choice (2-AFC) method as the standard of reference as the latter has proven its usefulness in the estimation of sensitivity over time. We quantify the amount of agreement between the two methods in the presence of different violations in the assumptions underlying MLDS. We then test the empirical consistency between sensitivity estimates derived with MLDS and forced-choice procedures in two experiments with a slant-from-texture task. In Experiment 1 observers judge supra-threshold slant differences and perceptual scales are derived from the judgments using MLDS. From these scales we derive sensitivity estimates (thresholds) at different slant levels. In Experiment 2 observers judge near-threshold slant differences in a two-interval forced-choice (2-IFC) task. Sensitivity estimates (thresholds) are derived from psychometric functions for the same slant levels as in Experiment 1.

To anticipate, the amount of agreement between sensitivity estimates from the two methods varied substantially across observers. The simulations showed that disagreement between the method might be due to violations of the model assumptions underlying MLDS.

10 SIMULATIONS

The sensory representation was modelled as a power function, $\Psi(x) = x^e$, with exponent $e = 2.0$ (Fig. 3.2A). We used an exponent greater than one so that sensitivity would increase with stimulus intensity, which is the case for slant-from-texture (Knill, 1998). The sensory representation function was used to simulate responses of a model observer for the MLDS and the 2-IFC procedure. It was assumed to be a Gaussian random variable with the mean corresponding to $\Psi(x)$ and unique variance σ^2 (Figs. 3.2B-C). An example simulation is depicted in Figure 10.1. Thresholds were derived for a standard value of $st = 0.6$ from MLDS scales (panel A) and from psychometric functions in a 2-IFC task (panel B).

10.1 MLDS THRESHOLDS

We performed the MLDS experiment with the method of triads (Maloney & Yang, 2003; Knoblauch & Maloney, 2012). A triad consists of three stimuli, x_1, x_2 and x_3. To simulate a triad the generative model (Fig. 3.2A) assigns perceptual responses, Ψ_{x_i}, to each of the three stimuli, x_i. The simulated observer decides which of the pairs, (x_1, x_2) or (x_2, x_3), contains the bigger difference in perceived slant according to the decision model depicted in Figure 3.2B.

MLDS data (simulated and observed) were analyzed with the R package *MLDS*, available in CRAN (Knoblauch & Maloney, 2008) and with python routines based on *numpy* and *scipy* libraries. A python wrapper of the *MLDS* routines together with all subsequent analysis routines is available online (`http://github.com/TUBvision/mlds`).

We first estimated a perceptual scale from the simulated responses by employing the standard MLDS routines available in R (Knoblauch & Maloney, 2008) (see Section 3.2 for a detailed description of the estimation procedure). We then derived sensitivity estimates from the perceptual scale following the procedure

suggested by Devinck and Knoblauch (2012). To do this we re-parametrized the original unconstrained scale so that the scale values are expressed in units of d'. The details underlying the re-parametrization are explained in Section 3.3. In the simulation we derived sensitivity estimates for eight standard values (experiments were done with four standard values). Due to the non-linear shape of the perceptual scale the local slopes differed between different standard values and hence translated into different sensitivity levels along the stimulus dimension. For each standard we determined sensitivity at three performance levels ($d' = 0.5, 1$ and 2) above and below the standard. To derive the stimulus values that corresponded to each d' difference for a given standard, we interpolated between the sampled data points with a cubic spline fit ($\hat{\phi}(x)$, shown as solid dark gray line in Figure 10.1A). The scale value ($\hat{\phi}(st)$ in d' units) that corresponds to a particular standard stimulus (st) and performance level (d') was read from the fitted function. The readout can be described by

$$\hat{\phi}^{-1}\left(\hat{\phi}(st) \pm d'\right) = \hat{\theta}^{st}_{\pm d'}|\text{MLDS} \tag{10.1}$$

in which the $+$ ($-$) sign next to d' stands for comparison values above (below) the standard, and $\hat{\theta}^{st}_{\pm d'}|\text{MLDS}$ stands for a particular sensitivity value in stimulus units as estimated by MLDS.

10.2 2-IFC THRESHOLDS

The same generative model is used to simulate responses in the 2-IFC procedure. In each trial one response is generated for the standard and one for the comparison value. Perceptual responses are compared according to the decision model depicted in Figure 3.2C. We simulated the same number of trials that we ran in the behavioral experiments (see sections 11.1.3 and 11.1.4).

To allow the comparison of thresholds across different standard slants we report comparison values in terms of differences relative to each standard. We fitted separate psychometric functions for positive and negative comparison values (smaller and larger than the standard). Psychometric functions were Weibull functions (F) with the guess rate (γ) set to 50 % chance level. The lapse rate

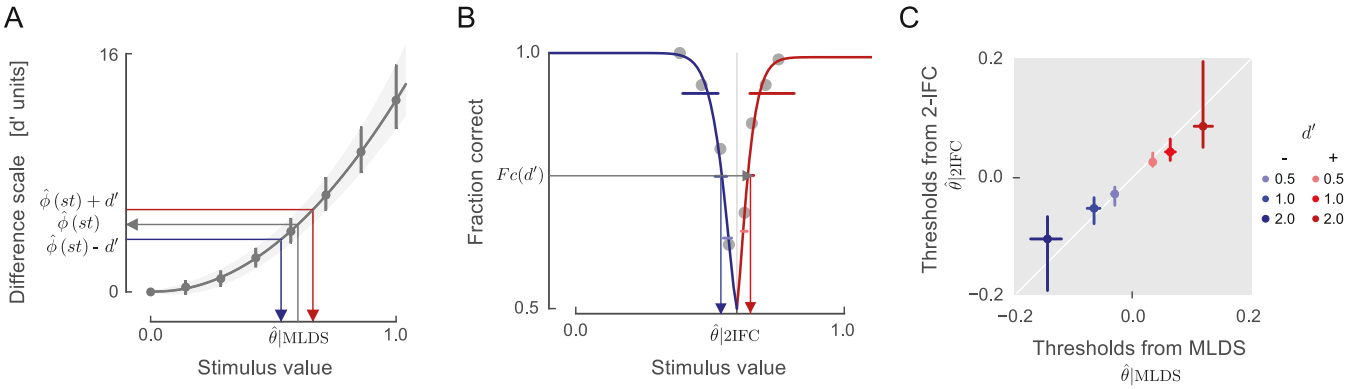


Figure 10.1: Comparison of MLDS and forced-choice thresholds. A. Difference scale for a simulated MLDS experiment using the sensory function depicted in Figure 3.2A. A cubic spline (dark solid line) is fitted to the scale values (circles). The procedure to read out thresholds is illustrated by arrows. Here we read out the threshold ($\hat{\theta}|_{\text{MLDS}}$) for a standard st $= 0.6$ (vertical gray line) at a performance level of d$' = 1$ for comparisons above (red arrow) and below (blue arrow) the standard. B. Psychometric functions from a simulated 2-IFC procedure with the same sensory function for comparisons above (red) and below (blue) the standard (vertical gray line). Thresholds ($\hat{\theta}|_{\text{2IFC}}$) were derived from the fraction correct corresponding to a performance level d$' = 1$ (Fc $= 0.76$). C. Thresholds derived with each methods are plotted against each other. They are expressed relative to the standard (st $= 0.6$) for comparisons above (red colors) and below (blue colors) the standard. The main diagonal white line indicates identity. Errorbars indicate 95 % C.I.

(λ), slope and position parameters of the psychometric function were estimated using Bayes inference (Kuss, Jäkel, & Wichmann, 2005). We used the *psignifit4* implementation (Schütt, Harmeling, Macke, & Wichmann, 2016) for function fitting, estimation of confidence intervals and analysis of goodness of fit. Each psychometric function was estimated from a total of 320 trials (4 comparison values x 80 repeats) as in the experiments.

An example psychometric function for one standard slant is shown in Figure 10.1B. Performance thresholds were obtained from each psychometric function by finding the stimulus value that produces a percentage correct corresponding to a desired d'. Assuming the equal variance Gaussian case of a signal detection model (Green & Swets, 1966), d' can be converted to percentage correct and vice versa, and the threshold can be read out by

$$\hat{F}^{-1}(Fc') = \hat{\theta}^{st}_{\pm d'}|2IFC$$

where $+$ $(-)$ indicate comparisons above (below) the standard, $Fc' = \{0.28, 0.52, 0.84\}$ are the unscaled fractions correct (range between 0 and 1) that correspond to the raw fractions correct $Fc = \{0.64, 0.76, 0.92\}$ (range between 0.5 and 1.0). These fraction correct values Fc correspond to the performance levels of d' = 0.5, 1 and 2, respectively, in a two-alternative forced-choice task (Green & Swets, 1966).

10.3 THRESHOLD COMPARISON

In Figure 10.1C the thresholds derived with each method are plotted against each other. They are expressed as differences relative to the standard value. Perfect agreement between the two methods is indicated by the main diagonal. To evaluate the statistical significance of the differences between thresholds we estimated the 95% confidence intervals for each of the thresholds using the bootstrap technique (for details see Section 10.5).

Thresholds were said to be in agreement when either one of the two confidence intervals of a data point (vertical or horizontal corresponding to 2-IFC and MLDS, respectively) crossed the unity line. This criterion ensures that the point estimate of one method is included in the 95% confidence interval of the other method.

In Figure 10.1C all data points coincided with the unity line resulting in a 100 % agreement.

We used this measure to quantify the degree to which the consistency between the thresholds. For eight different standard values we performed n=1000 simulations and Figure 10.2A shows a summary of the results for the average of the empirically observed noise level, $\sigma = 0.07$ (green lines). Thresholds agreed in more than 90 % of the cases, and the agreement was also high across a range of noise levels that we tested, from $\sigma = 0.035$ to $\sigma = 0.14$, which includes all the values of sensory noise observed in the experiments.

10.4 THRESHOLDS THAT COULD NOT BE OBTAINED

The estimation procedure in either of the methods sometimes failed when sensitivity was low. When the stimulus was in a range where the sensory function is too shallow, for example for values below 0.4 in the sensory function in Figure 3.2A, the interpolation of scale differences was not possible. Similarly, the psychometric function was sometimes so shallow that it did not allow the read out of a threshold at a given performance level. These 'failure' cases provide an additional test of consistency between the two methods, because when sensitivity is genuinely low both methods should fail to provide a threshold estimate. We counted the number of cases in which either one or both of the methods did not provide a threshold estimate for a given performance level. The results are shown in Figure 10.2A (gray lines). It can be read from the Figure that both methods did consistently fail to provide threshold estimates for standard values near zero.

10.4.1 *Model assumptions*

To test the effect of violations of some of the model assumptions on the agreements between thresholds we repeated the simulations with a modified generative model. We introduced sensory noise that was not independent of the stimulus level but instead increased with the stimulus value. We also tested a model

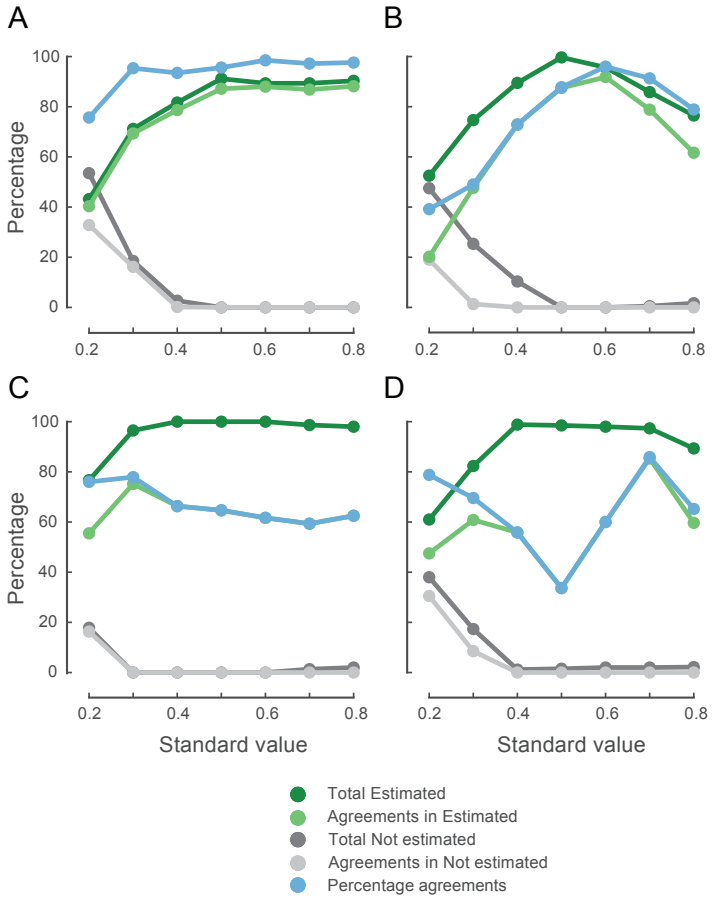


Figure 10.2: MLDS and forced-choice thresholds from simulations. At each standard level, thresholds for different performance levels could be successfully estimated (dark green) and have quantitative agreement between them (light green). There were cases in which thresholds could not be estimated (dark gray), from which an agreement occurred when both methods were unable to estimate it (light gray). The sum of agreement cases for estimated and not estimated thresholds is also shown (light blue). Percentage over 1000 simulations. (A) independent, equal-variance case with noise level $\sigma = 0.07$ (B) independent, unequal-variance case with increasing noise level from 0.035 to 0.14 (violation of equal-variance assumption), (C) same as (A) but with added uniform correlation in the sensory representation of $\rho = 0.8$ (violation of independence assumption), and (D) same as (B) but with added uniform correlation of $\rho = 0.8$ (both assumptions violated).

that included uniform correlations between the sensory representations (specific details in Chapter 4). These two modifications violate the assumptions of equal variance (assumption 1) and independence (assumption 2). As illustrated by Kingdom (2016) and tested in simulations by Maloney and Yang (2003), the scales themselves are insensitive to a violation of the equal variance assumption. However, violating the equal variance assumption did reduce the agreement between thresholds (Figure 10.2B) in particular for extreme standard values where the simulated noise was respectively lower or higher than in the equal-variance model. The reason for this is illustrated in Figure 10.1 which shows how threshold readout depends on noise on the sensory axis. Introducing correlations reduced the amount of agreement between thresholds independent of the standard value (Figure 10.2 C). We observed the smallest agreements when both assumptions are violated (Figure 10.2D).

10.5 VARIABILITY OF THRESHOLD ESTIMATES

To study the variability of threshold estimates we made use of the bootstrap samples that are already generated by MLDS to calculate confidence intervals (CIs) for the scale values (error bars in Figure 10.1A, Knoblauch & Maloney, 2012). Bootstrap samples are generated from the response probabilities that are observed for each triad. Each bootstrap sample is a new perceptual scale and by default MLDS generates 1000 of these bootstrapped scales. To derive the bootstrap samples for a particular threshold value we fitted a cubic spline to each bootstrap scale and determined the slant value corresponding to the threshold value (see Equation 10.1). From these bootstrap distributions we obtained the 95 % confidence intervals for each threshold (a detailed description of the procedure can be found in Section 3.4).

To compare the confidence intervals associated with each method Figure 10.3A plots the widths of the respective CIs against each other, for one example standard. The main diagonal indicates equal width in the confidence intervals, data points above the main diagonal indicate that the width of the CIs for thresholds from MLDS were smaller than the width of the CIs for thresholds from 2-IFC. For all standard levels (Figure 10.3) and for all tested noise levels (see Appendix) the

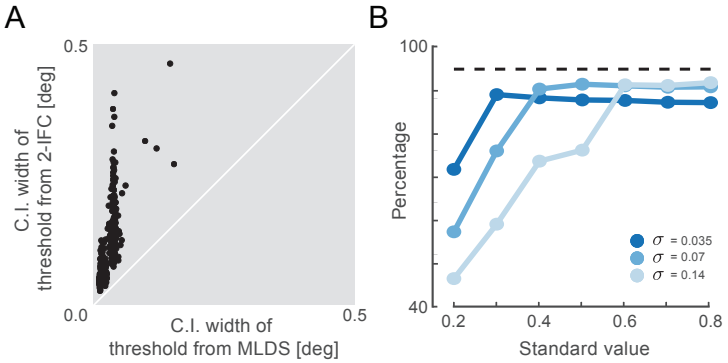


Figure 10.3: (A) Comparison of the variability in the threshold estimation. The width of the confidence intervals are plotted against each other for multiple simulations at one standard stimulus value $st = 0.4$ as example. (B) Coverage of threshold estimated at different standard levels, and for three different simulated noise levels (σ). Expected coverage of 95 % is shown as a black dashed line.

majority of confidence intervals (99 %) was smaller in MLDS than in 2-IFC. This is curious because MLDS requires a smaller amount of data than 2-IFC.

A smaller width in the confidence intervals could either be due to a truly more precise estimate (less underlying variability), or alternatively, it could result from an insufficient coverage of the confidence intervals. This is a common problem in derivations using bootstrap techniques (Wichmann & Hill, 2001) and we tested this with an analysis of the coverage of the scales and the derived thresholds. We calculated coverage by counting how many times the 'true' value (as defined in the generative model) was contained in the estimated confidence interval of a scale or threshold. For confidence intervals to be credible, coverage across multiple simulations should reflect the confidence in the confidence interval, i.e. coverage should be 95% over multiple simulations for 95% confidence intervals.

Coverage of the scale estimates was adequate for the range of noise levels studied (Figures 4.4 and A.7). MLDS thus provides credible confidence intervals for the scale estimates that it was designed for. However, coverage for the threshold

estimates was at best at 90 % for nominal values of 95% (Figure 10.3B), and for stimulus values at shallow portions of the sensory function (e.g. smaller than 0.4) coverage was as low as 50 - 60 %. These results indicate that the confidence intervals for the thresholds derived with MLDS were indeed too narrow. Threshold variability might hence be underestimated when derived from MLDS in the way described above. This is an important caveat when using MLDS to estimate thresholds.

Upon suggestion of the reviewers we performed a sanity check for the confidence intervals to test for their stability and bias when trial numbers are high. We repeated the scale and threshold estimation procedure and calculated bias and coverage for increasingly large trial numbers. For the scale estimation, the pattern of results indicates that coverage slightly improved with an increasing trial number (Figure 4.4). For the threshold estimation, coverage did not improve when the number of trials is tripled (Appendix Figure A.11). This result suggest that low coverage was not due to a small sample size.

10.6 SUMMARY AND DISCUSSION

We simulated an observer model with a known sensory representation function. We compared thresholds derived from MLD scales with thresholds derived from a 2-IFC procedure at different standard values and performance levels. We found a high degree of consistency between thresholds obtained with each method, when all the assumptions are met (Fig. 10.2A). The amount of agreement did not depend on the sensory noise level for the range of noise levels that was observed experimentally. The estimation procedure fails to obtain thresholds when sensitivity is low. In most of these cases both methods failed to estimate a threshold which is a further indication of consistency between them. The variability of threshold estimates, quantified as the width of their confidence intervals, is smaller for MLDS than for 2-IFC thresholds (Fig. 10.3A). This finding needs to be qualified by the coverage analysis which indicates that the bootstrapped confidence intervals for MLDS thresholds might be too small (Fig. 10.3B).

The agreement between threshold estimates did not amount to the theoretically expected 100 %. This might be due to the rather small number of simulated

trials. However, the simulations should capture actual psychophysical experiments where it is not practicable to collect large numbers of trials. In addition, they capture the software pipeline of estimation and statistical inference, which could be prone to different kind of problems (e.g. numerical). Thus, the simulation results establish an upper bound for the agreement that is expected for a realistic amount of collected data and estimation procedures.

Finally, we found that violating the equal-variance assumption by MLDS might lead to disagreement between estimated thresholds. The disagreement is relevant because unequal variance models might fit behavioral data better when the equal variance assumption is violated in real data (e.g. Goris et al., 2013).

11

EXPERIMENTS

11.1 METHODS

11.1.1 *Observers*

Six naïve (three male, age range between 23 and 29) and two experienced observers (observers "O3" and "O6") participated in the study. All observers had normal or corrected-to-normal visual acuity. The participation of naïve observers was voluntary and financially compensated. Informed written consent was given by all observers prior to the experiment.

11.1.2 *Stimuli and Apparatus*

Stimuli were planes textured with a 'polka dot' pattern and slanted about their horizontal axis. They were generated in two steps. First, the textures containing the 'polka dot' pattern were generated as 2500 x 500 pixel images. The 'polka dot' pattern is created using a *hard core point process*, which is a random spatial process that avoids dot superposition by applying an inhibition radius to each point. Using the R package *spatstat* (Baddeley & Turner, 2005), we generated fifteen samples of this process following specifications from previous work (Rosas et al., 2004). The textures consisted of black dots (0.4-0.6 cd/m^2, 12 pixels or 0.5 deg visual angle in diameter in the fronto-parallel plane) on a gray background area (48-52 cd/m^2, Figure 9.1).

In a second step the textured planes were rendered in 3-D using OpenGL (Shreiner, Woo, Neider, & Davis, 2005). The planes were slanted and perspectively projected into 2-D. The so-generated planes were viewed through simulated circular apertures that subtended 8.3 deg of visual angle and were added at the depth of the screen distance.

Stimuli were displayed on a 24.1-in. LCD monitor (Eizo CG243W 496x310mm, 1920x1200pixels, 60 Hz) located in a dark cabin. Observers viewed the stimuli monocularly with their dominant eye at a distance of 60cm. Eye dominance for each observer was determined with the Miles test (Miles, 1930) prior to the start of the experiment. The non-dominant eye was covered with an eye-patch and the head rested on a chin rest. Stimulus presentation was controlled by a computer (Apple Mac Pro QuadCore 2.66 with a graphic card Nvidia GeForce 7300GT) that was running custom-made software which was based on *python* and the visualization library *pyglet*. Observers' responses were registered via the keyboard.

11.1.3 *Procedure Experiment 1: MLDS*

In each trial, three stimulus exemplars that varied in slant were presented next to each other. Each of the slanted surfaces was rendered independently and viewed through a different circular aperture (see Figure 9.1). Slant values (s) varied between 0 (fronto-parallel) to 70deg in steps of 10deg. This spacing results in $p = 8$ possible slant values and a total number of $n = p!/((p-3)! \times 3!) = 56$ unique triads.

By design each triad consists of stimuli that are slanted so that the two intervals enclosed by the three stimuli do not overlap. The stimuli in a triad were presented in either ascending ($s_1 < s_2 < s_3$) or descending ($s_1 > s_2 > s_3$) order, and the order was randomized across trials. Observers were asked to report which of the pairs, (s_1, s_2) or (s_2, s_3), contained the bigger perceived difference in slant. Observers viewed the stimulus configuration with no time limit for their response. They indicated their choice by pressing a keyboard button and this triggered the next trial after a delay of one second. No feedback was given as to the correctness of the response.

The full set of unique triads was presented in one experimental block and 15 such blocks were presented within one session. In total, each observer judged 840 triads. This was the same amount of trials used in the simulations. Observers could pause after each block. Before the experiment observers were shown two to five examples of extreme triads ($(0, 10, 70)$ and $(0, 60, 70)$ deg), together with

the 'correct' answers and the corresponding keyboard presses. We employed this instruction method to ensure that observers understood the task. Comparing stimulus intervals is not an obvious task and in previous experiments we noted that, instead of reporting the pair with the biggest perceived *difference*, some observers reported the pair that included the most extreme slant.

11.1.4 *Procedure Experiment 2: 2-IFC*

A standard 2-IFC procedure was employed in Experiment 2. A trial started with a fixation cross that appeared for 1000 ms in the center of the screen. Then the first stimulus was presented for 200 ms. Its contrast ramped on and off from zero to full contrast and back to zero within the first and last 50 ms of presentation so that the stimulus was seen at full contrast for 100ms. After a blank inter-stimulus interval of 500 ms the second stimulus was presented with temporal parameters identical to those of the first. After stimulus offset observers had to report which of the two stimuli was more slanted using a keyboard button to indicate first or second. Observers did not receive feedback about their performance. Standard and comparison stimuli were randomly assigned to the first or the second interval.

Discrimination performance was measured for the same four standard slant values (26, 37, 53 and 64 degrees) for which MLDS thresholds were predicted. Each standard slant was compared with one of eight comparison slants (four below and four above the standard slant) in a method of constant stimuli procedure. In the first session the range of comparison stimuli for each standard slant was selected based on the point estimates corresponding to performance levels of $d' = 0.5, 1, 2$ and 3 that were derived from the MLD scale (section 10.1). After the first session the comparison values were adjusted so as to provide good coverage of the psychometric function (Wichmann & Hill, 2001). The full experimental design contained 4 standards x 8 comparison values (four above and four below the standard) x 80 repeats resulting in 2560 trials in total. This amount was the same as in the simulations. The presentation was randomized and the total number of trials was subdivided into 40 blocks of 64 trials each. Observers completed

all trials in three to four sessions of maximum one hour duration. Experiment 2 was run on a different day than Experiment 1 and subsequent to it.

There are obvious differences in stimulus spacing as well as in the number of trials between both methods and both factors might affect the shape of the respective fitted functions, scales or psychometric functions. However, there is no principled way to equate these aspects across the procedures and we would argue that they had little effect on the present results. We performed goodness-of-fit analyses for both procedures which showed that the fitted functions captured the data, and which also indicates that the stimulus choice was reasonable.

11.2 RESULTS

The objective of the experiments was to compare sensitivity estimates from a forced-choice and an MLDS procedure at different positions along the perceptual scale. Before we report these results we will show that the thresholds from the forced-choice task were comparable to those reported in earlier studies of slant-from-texture discrimination (Rosas et al., 2004).

The procedure in the forced-choice task (Experiment 2) was identical to that employed by Rosas et al. (2004). To capture sensitivity they computed an "area" measure, which was defined as the region between the two psychometric functions fitted separately for smaller and larger comparison values enclosed by the 60% and 80% percent performance levels (see Fig. 10.1 B). This 'area' is small when the psychometric functions are steep, i.e. when sensitivity to slant differences is high, and conversely, it is large when sensitivity is low. Thus, the calculated area measure is inversely related to the sensitivity at a particular standard slant.

We computed the area measure for each standard value and each observer. The results are shown in Figure 11.1. In order to average across observers the area measure was normalized relative to the highest value for each observer individually, because observers had different overall sensitivity to slant (inter-observer variability). In all observers the area measure was maximal for a standard slant of 26deg indicating lowest sensitivity. For comparison Figure 11.1 also shows the mean normalized area of the five observers reported in Rosas et al. (2004,

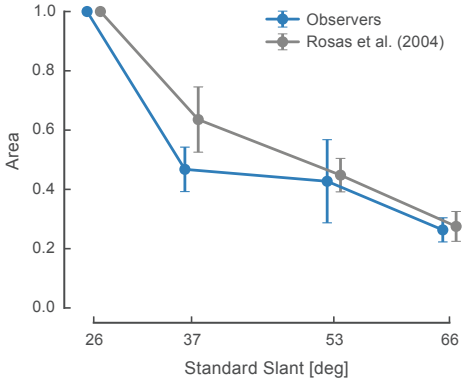


Figure 11.1: Sensitivity obtained from psychometric functions in Experiment 2. The 'area' enclosed by the two psychometric functions and the 60 % and 80% percentage correct (y-axis, see Figure 10.1B for a depiction) is plotted for the different standards (x-axis). Areas were normalized for each observer with respect to the maximum and aggregated across observers. Data from Rosas et al. (2004) is shown as reference (mean ± s.e.m.).

p.1523). Apart from the variability between observers, sensitivity increased with slant which is in accordance with the data reported by Rosas et al. (2004).

11.2.1 *Threshold comparison*

Thresholds for MLDS and 2-IFC were obtained in the same way as in the simulations. Figure 11.2 shows the data of one single observer. Thresholds from both methods are plotted against each other for performance levels of $d' = \pm 0.5, 1, 2$ and for the four standard values tested (panels). Data points lying on the main diagonal indicate a quantitative agreement between thresholds. This was observed for thresholds obtained at standard slants of 37, 53 and 66 deg. For a standard slant of 26 deg a correspondence between thresholds from both methods was observed for comparisons that were larger than the standard. For comparisons that were below the standard MLDS thresholds were smaller than 2-IFC thresholds. For

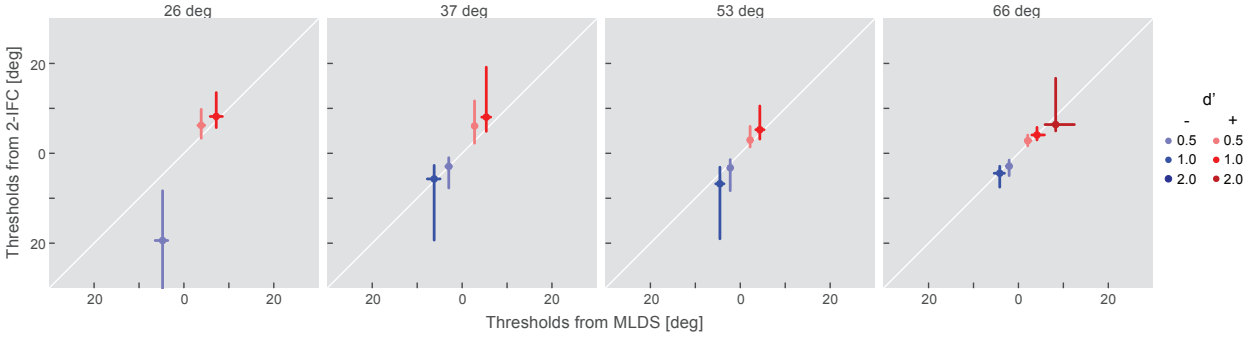


Figure 11.2: Threshold comparison for one observer (O1). Estimates of threshold from MLDS in Experiment 1 (x-axis) and from the psychometric functions obtained in a 2-IFC procedure in Experiment 2 (y-axis) are shown for each standard (different panels) and d′ performance level, for comparisons above (+, warm colors) and below (-, cold colors) the standard. Thresholds are expressed as relative values to the standard. Error-bars denote the 95% confidence interval of the point estimate.

some combinations of performance levels and standard values thresholds from either or both methods could not be calculated (see section 10.4).

As described for the simulations we classified thresholds to be in agreement when one of the confidence intervals of either method crossed the identity line (as in Fig. 10.1C). Observers differed substantially in their proportion of agreement between thresholds. We sorted them according to the amount of agreement in descending order (Fig. 11.3). There was agreement in 15 out of 16 data points (94 %) for observer O1 in Figure 11.2, 11 of 14 (79 %) for observer O2, 15 of 22 (68 %) for observer O3, 10 of 18 (56 %) for observer O4, 10 of 19 (53 %) for observer O5, 10 of 21 (48 %) for observer O6, 5 of 20 (25 %) for observer O7, and 2 of 14 (14 %) for observer O8. For observers O7 and O8, the data points fell above the diagonal line for comparisons above the standard (Figure 11.3, red markers), and below the diagonal line for comparisons below the standard (blue markers). This pattern of results indicates that for these two observers thresholds obtained with MLDS were consistently smaller than thresholds obtained with 2-IFC. In other words, MLDS estimated a higher sensitivity than the 2-IFC procedure; the opposite

case did not occur. Taking all observers and standard levels together, 78 out of 144 (54 %) estimated thresholds agreed between the two methods.

11.2.2 *Thresholds that could not be obtained*

Thresholds could not be obtained from either method for stimulus comparisons that involved the lowest standard value (26deg) and/or comparison slant values below 30deg. For example, for the observer depicted in Figure 11.2 thresholds from MLDS could not be obtained for performance levels of $d' = 1, 2$ for comparisons below the standard slant of 26deg. The reason for this discrepancy was a shallow slope in the scale reflecting low sensitivity at that particular stimulus level.

As in the simulations we counted the number of cases in which either one or both of the methods did not produce a threshold for our experimental results. A total of 22 cases occurred in which either one of both thresholds could not be obtained. Four out of the 22 cases were cases in which thresholds from MLDS were missing (at standard of 26deg and 37 deg), eleven were cases in which thresholds from 2-IFC were missing (standard 26 deg and 37deg) and seven were cases in which both thresholds were missing (all for a standard of 26deg). So the methods consistently estimated low sensitivity in 32% of the cases for which thresholds could not be obtained.

11.2.3 *Variability of threshold estimates*

We also derived the variability of the threshold estimates for the experimental data. We found that the variability was lower for MLDS than for 2-IFC, consistent with the simulations. Figure 11.4 shows the widths of the confidence intervals for the thresholds obtained with each method from all observers. As in Figure 10.3A confidence intervals were smaller for thresholds from MLDS than for thresholds from 2-IFC. Overall for 142 of the 144 threshold comparisons (98.6 %) the width of the confidence interval was smaller for thresholds from MLDS (separate comparisons for each observer can be found in Appendix Figure A.10).

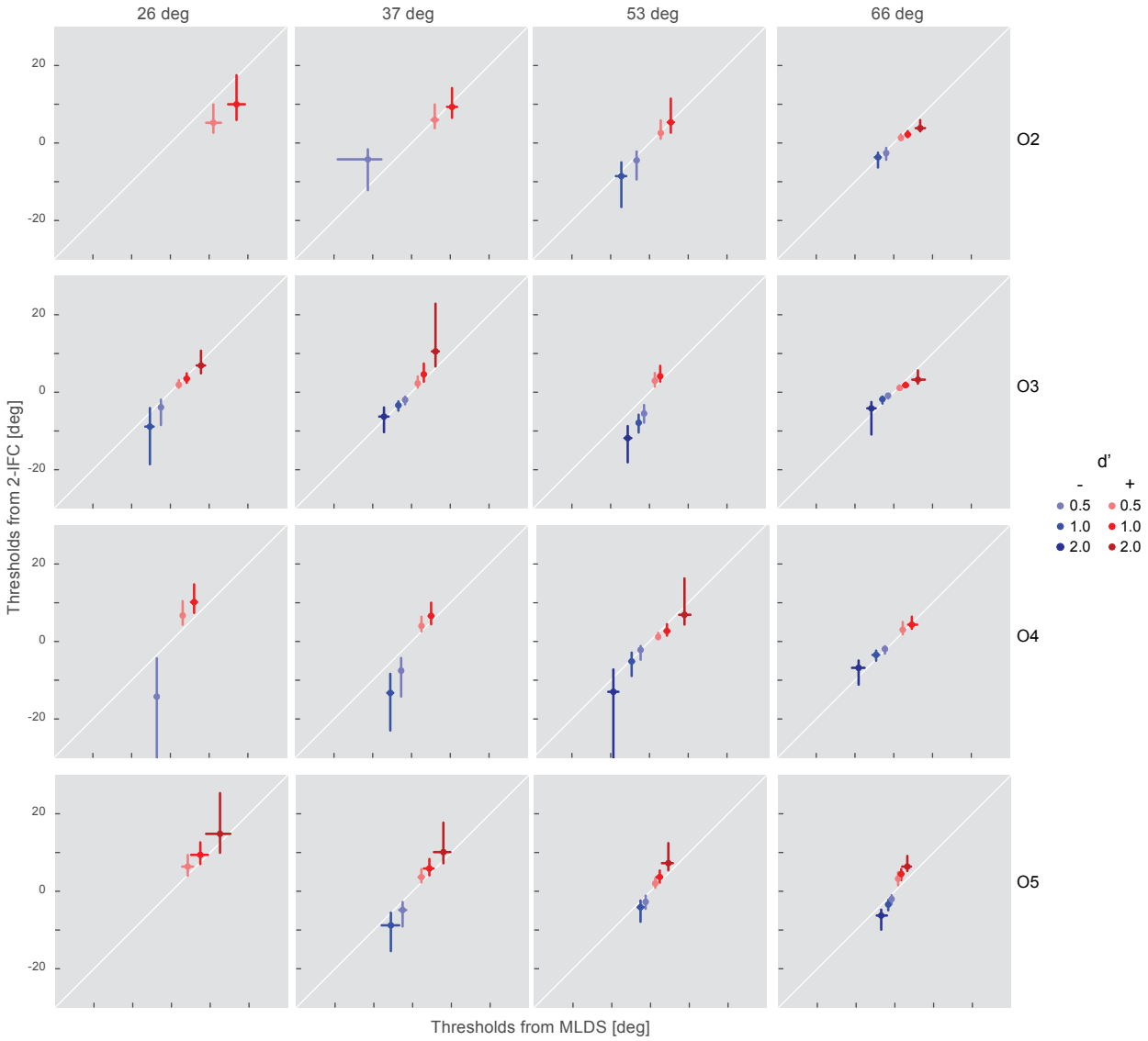
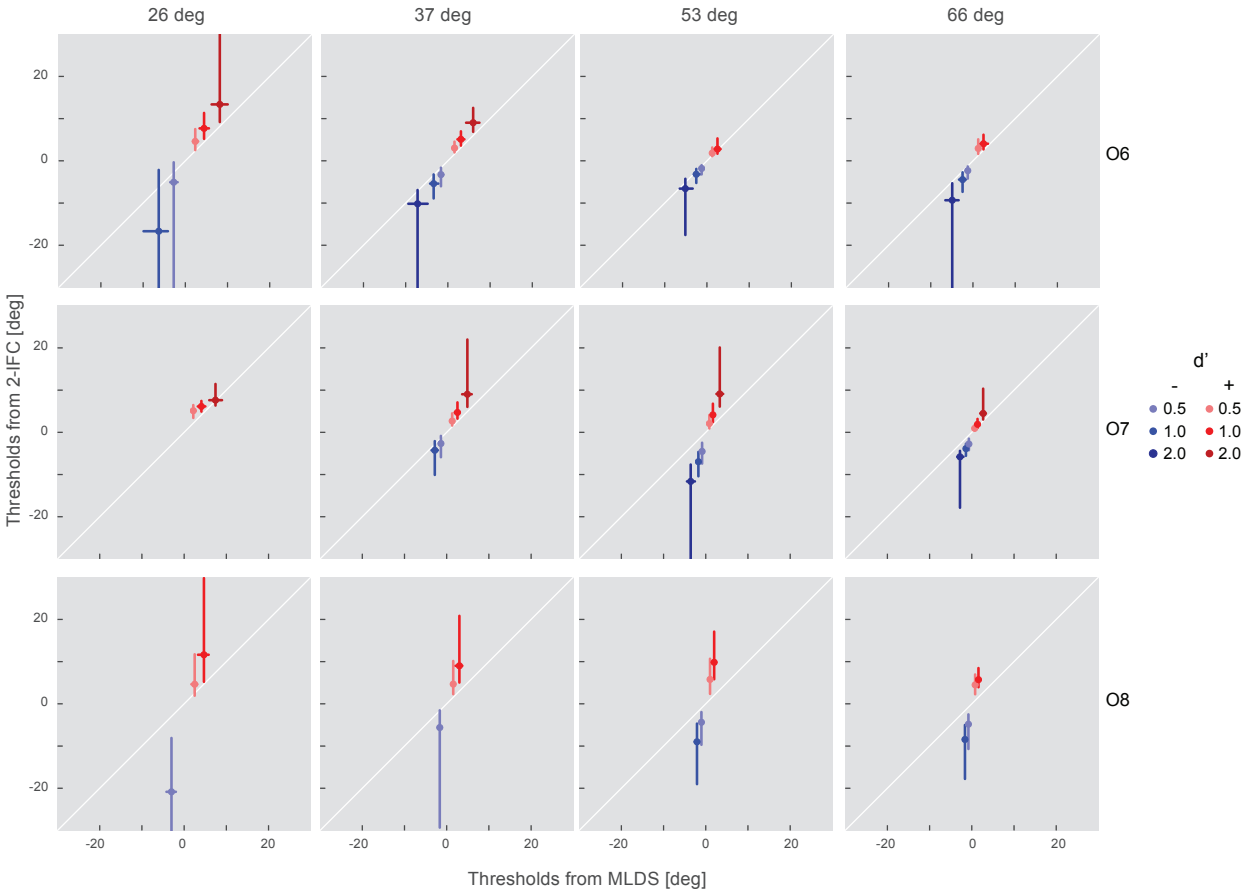


Figure 11.3: Threshold comparison for seven observers (O2-O8). Similar to Figure 11.2, thresholds relative to the standard obtained from the two methods are compared by observer (rows), standard (columns), and performance level (d′), for comparison values above (+, warm colors) and below (-, cold colors) the standard. Observers are sorted by percentage of threshold agreement between the two methods in descending order.


26 deg
37 deg
53 deg
66 deg
O6
O7
O8
Thresholds from 2-IFC [deg]
Thresholds from MLDS [deg]
20
0
-20
d'
-
+
0.5
1.0
2.0
0.5
1.0
2.0

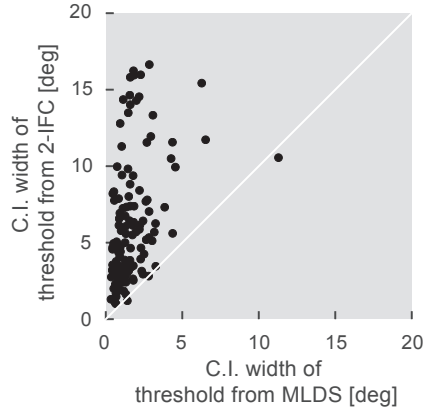


Figure 11.4: Comparison of the variability in the threshold estimation. The width of the confidence intervals from Figure 11.2 and Figure 11.3 are plotted against each other, for all observers and standard values.

12

DISCUSSION

The goal of the present study was to test whether judgments of stimulus appearance and judgments of stimulus discriminability are mutually consistent which would suggest that both types of judgments rely on a common perceptual representation of the stimulus dimension under study. The evidence on this question is mixed (e.g. Krueger, 1989; Ross, 1997; Hillis & Brainard, 2007b), but comparing supra-threshold judgments in a MLDS procedure and near-threshold judgments in a forced-choice procedure, Devinck and Knoblauch (2012) have reported that the two can be linked within a common signal detection framework. Using slant-from-texture stimuli we conducted two experiments that independently measured the sensitivity to differences in slant. In the first experiment observers judged supra-threshold stimulus differences and we derived thresholds from perceptual scales using the MLDS framework (Maloney & Yang, 2003). In the second experiment we measured sensitivity in a conventional two-alternative forced-choice procedure and we derived thresholds from psychometric functions. For some observers there was agreement between thresholds obtained with both methods but across observers the methods agreed in only 54 % of the cases. For two observers (O7 and O8) sensitivity estimates from the MLDS procedure were consistently higher than those from the forced choice procedure.

The observed lack of correspondence between the estimates could imply that the two tasks do indeed probe different perceptual representations of a stimulus. Alternatively, the lack of correspondence might result from violations of the model assumptions, and hence would not be informative about the relationship between appearance and discrimination tasks.

12.1 VIOLATIONS OF MODEL ASSUMPTIONS

The equivalence between MLD scales and the 2-IFC procedure used in the present work relies on a number of theoretical assumptions concerning the sensory representation and the decision model. In the following we will describe the effect of violations of one or more of these assumptions on the estimated scales and the consequences for the estimation procedure.

12.1.1 *Goodness of fit*

The MLDS framework provides goodness of fit procedures that test the plausibility of the data being produced by a difference scaling model (Knoblauch & Maloney, 2008, p. 219-222). In our data, the goodness of fit of the difference scales was insufficient for five out of eight observers when the default parameter values were used. We followed the refitting procedure suggested by Knoblauch and Maloney (2008, pp. 219-222) for these cases to modify the model specification. The procedure includes the estimation of 'guess' and 'lapse' rates, and a split of the raw data into two parts that were evaluated separately (detailed description of the goodness of fit procedure is provided in the Appendix A.2). After the refitting procedure we obtained an appropriate goodness of fit for all observers, and we derived thresholds from the refitted scales. The thresholds that were derived from the adjusted scales were not markedly different from the original ones. In particular, the disagreement between thresholds that we observed in three observers was present with or without the goodness of fit adjustment. Thus, the model violations that were detected by the goodness of fit routines did not have much of an effect on the shape of the scale, at least for the present data.

12.1.2 *Reconciling MLDS with 2-IFC thresholds.*

The assumption of independence between different levels of the sensory representation (assumption 2) is not and cannot be tested by the goodness of fit routine. If this assumption is violated it would affect the noise and it would require

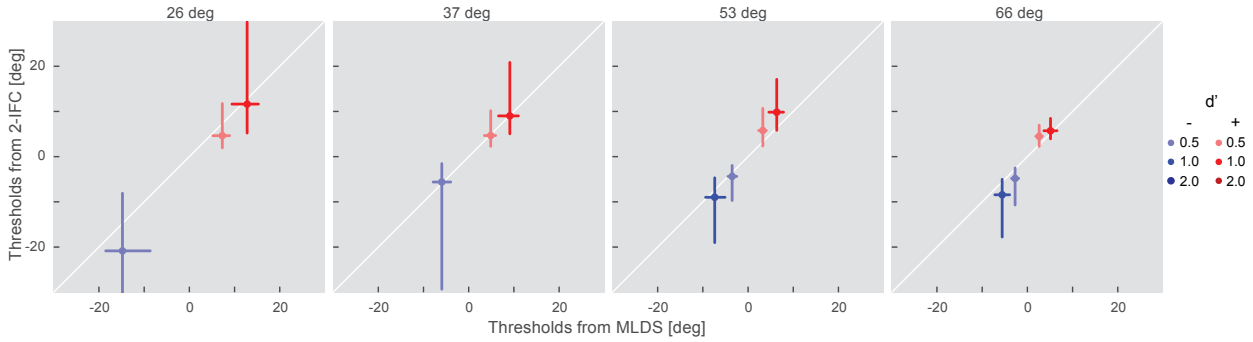


Figure 12.1: Threshold comparison for observer O8 when the difference scales are rescaled by a different factor to account for a possible dependence between different realizations of the random variable that characterizes the sensory representation (a factor of 0.6 corresponding to a correlation coefficient of 0.9).

an adjustment of the scaling factor that transforms the original MLDS scale into a scale in units of d′. The independence assumption would be violated when the sensory representations cannot be characterized as independent realizations of a Gaussian random variable but are instead correlated with each other. We tested the effect of these kind of correlations in the sensory representation in simulations. Correlated sensory variables do indeed affect the threshold estimates. To illustrate the effect we show that the magnitude of the correlation can be chosen so as to elicit a correspondence between thresholds derived from MLDS scales and from 2-IFC. Figure 12.1 shows the thresholds for observer O8 for a simulated case in which the sensory representations are highly correlated ($\rho = 0.9$). As a consequence of this correlation we give up the independence assumption and would have to rescale the perceptual scale by a factor of 0.6 (instead of the theoretical factor of two). In this scenario the resulting thresholds from MLDS correspond better with the thresholds from 2-IFC. Thus, an alternative transformation that accounts for a model violation can 'produce' a higher agreement between the two types of thresholds. We are not aware of any method to test the assumption

of independence empirically and it is therefore not possible to evaluate which of the many possible transformations is closest to the true sensory representation.

A similar issue arises when we scrutinize the effect of violating the assumed decision rule (assumption 4). Based on the assumption that the sensory representation function is monotonically increasing the decision rule can be expressed as a double difference operation (Δ variable in Figure 3.2B) instead of an absolute value operation (Δ variable in Figure 3.2A). This change from an absolute to a relative difference operation can have noticeable effects when the sensory representations are random variables (as assumed here) instead of fixed values. To explore the effect of the differencing operation we simulated an observer that judged the triads by using either one of the two decision rules. To analyze the effect on the estimated noise we applied MLDS to each of the two types of simulated responses, and found that the absolute difference operation produced higher noise estimates than the double difference operation. This difference increases progressively as the underlying sensory noise increases. Thus, the two decision rules can produce different results (see Section 4.5.1 and Section 4.5.2 for simulations and details).

It is not possible to determine empirically whether or when observers apply an absolute or a relative differencing rule, they might even change the rule with varying difficulty of the judgment. One should be aware that a deviation from the assumed decision rule or a violation of the independence assumption may both affect the noise estimate although in opposite directions. Thus, the combined contributions of both factors can produce various types of deviation from the true scaling factor and this affects the scale and the derived sensitivity estimate.

12.2 VARIABILITY OF THRESHOLDS ESTIMATED BY MLDS

As in the simulations we observed that the variability of thresholds derived from MLDS was smaller than the variability of thresholds derived from 2-IFC (Figure 10.3 and Figure 11.4). Again, this is counterintuitive because MLDS requires a smaller amount of data for the threshold derivation. However, the smaller variability must be interpreted with care, because our simulations revealed that the coverage of MLDS derived threshold might be insufficient and the width of the

confidence intervals might be underestimated. This should be considered for hypotheses tests as it may lead to Type-I errors.

However, apart from the coverage problem associated with our threshold derivation, MLDS provides an efficient method to acquire sensitivity estimates. In the present Experiment 1 we ran the MLDS procedure with 840 trials which took about 45 minutes per observer. In contrast, the 2-IFC procedure in Experiment 2 required 2560 trials and lasted three hours. Thus, the difference in both the amount of data and the required acquisition time might be up to three to four times more for 2-IFC than for MLDS. MLDS thus provides an efficient alternative to forced-choice procedures to obtain a rough estimate of sensitivity.

12.3 CONCLUSIONS

In the present experiment we investigated the question of equivalence of thresholds derived from an MLD scale and thresholds derived from a forced-choice procedure. Using simulations, we established upper bounds for a possible agreement considering the theoretical model assumptions, the finite amount of collected data and the necessary software pipeline. Experimentally, we found varying degrees of correspondence between the methods for different observers. Out of a total of 144 threshold estimates the methods' sensitivity estimates differed in 66 cases. We discuss that the equivalence of thresholds (or lack thereof) might either indicate a corresponding equivalence between the underlying perceptual representations (or lack thereof) as has been argued by Devinck and Knoblauch (2012), or alternatively, it might result from violations of the model assumptions.

An important point that has been made by one of the reviewers is that we gave the 2-AFC method the benefit of history. In the present analysis we used the 2-AFC method as a standard of reference against which we compared the sensitivity estimates derived with MLDS. Accordingly we tested the effect of model violations on threshold agreement only for the assumptions underlying MLDS. However, considering the present data and the numerous benefits associated with the experimental procedures of MLDS, it might be warranted to try to elaborate the first principles case of which of the two methods we would trust more if we started out *de novo*.

Our positive evaluation of the MLDS method is corroborated by recent results from Kingdom (2016) who used MLDS to decide between competing theories of internal noise in contrast transduction. In summary, we conclude that MLDS, as state-of-the-art scaling method, seems to have great potential to be used beyond the purpose that it was originally designed for.

Part IV

DISCUSSION AND OUTLOOK

13

GENERAL DISCUSSION AND OUTLOOK

In this thesis I propose the use of MLDS as a tool for measuring perception by estimating perceptual scales in an efficient and reliable way. MLDS – an appearance method – uses the judgment of clearly visible stimulus differences which avoids the shortcomings of Fechnerian, Thurstonian, and direct scaling methods (Chapter 2). Critically, MLDS uses the method of triads which provides an intuitive and easy task for the observer, avoiding the disadvantages of performance-based (discrimination) methods that use difficult and unintuitive tasks.

MLDS was used to reliably estimate scales in cases of perceptual constancy. MLDS can recover the underlying perceptual scales for different viewing contexts but requiring only judgments involving within-context comparisons (Chapter 5). In the case of lightness constancy, we showed how MLDS seems superior to asymmetric matching procedures as these, unlike MLDS, do not estimate the internal scales directly but rather their existence is assumed and cannot be infered from observers' matches (Chapter 6). Additionally, matching data derived from the MLDS experiment followed closely the data acquired in an asymmetric matching procedure (Chapter 7), which suggests that MLDS indeed probes the internal perceptual representation, i.e. lightness.

MLDS can be also used for predicting sensitivity by framing it in a signal detection theory formulation (Section 3.3). In simulations the estimation of sensitivity deployed with traditional performance methods – from psychometric functions in a 2-IFC procedure – was equivalent to the estimation of sensitivity from MLDS when all model assumptions were met (Chapter 10). Critically, MLDS was more efficient, requiring less data than the traditional 2-IFC procedure. In an experiment using a slant-from-texture task, we found varying degrees of agreement between the methods for different observers, which either indicates a lack of equivalence between the underlying perceptual representations, or alternatively, it may result from violations of the model assumptions (Chapter 11).

13.1 THE METHOD OF TRIADS

Historically, the method of paired comparisons was the first procedure used in scaling that involved comparison of stimuli (Thurstone, 1927a). The method of pair comparisons commonly requires a significant amount of data for a scale to be estimated. Later, the *method of tetrads* was introduced by Torgerson (1958) as a direct extension of the method of paired comparisons, and it consisted of four stimulus exemplars judged simultaneously. This procedure later derived into the *method of triadic combinations* that made acquisition more efficient (Torgerson, 1958) by presenting three stimulus exemplars in all possible triadic combinations. The method of triads in MLDS is an additional simplification of the method of triadic combinations by assuming an ordered and monotonically increasing dimension, which limits the number of combinations to only non-overlapping intervals (Section 3.1). Interestingly, most of the studies using MLDS have used the method of quadruples (e.g. Maloney and Yang (2003); Obein et al. (2004); Fleming et al. (2011); Paulun et al. (2015); Charrier et al. (2007), see Devinck and Knoblauch (2012) for an exception). In this thesis the method of triads was used instead of quadruples, because upon introspection the comparison of triads feels easier than of quadruples.

Decision variable in triads and quadruples

From a mathematical point of view the two methods are identical (Knoblauch & Maloney, 2012). In the quadruple case the decision variable is specified as $\Delta_Q = (\Psi_{x_4} - \Psi_{x_3}) - (\Psi_{x_2} - \Psi_{x_1})$ and in the triad case as $\Delta_T = (\Psi_{x_3} - \Psi_{x_2}) - (\Psi_{x_2} - \Psi_{x_1})$. It is assumed that for the triad comparison Ψ_{x_2} is sampled twice resulting in two independent realizations of a random variable (for MLDS in its signal detection theory formulation). Accordingly, the variances of Δ_Q and Δ_T are identical and so is the maximum of the respective unconstrained scales.

Subjectively, the triad comparison feels easier because the compared intervals ($[\Psi_{x_1}, \Psi_{x_2}]$ and $[\Psi_{x_2}, \Psi_{x_3}]$) are adjacent to each other with a common reference stimulus (Ψ_{x_2}) that can be used as anchor. This benefit would translate into a decision model for triad comparisons where Ψ_{x_1} and Ψ_{x_3} are compared to a single realization of Ψ_{x_2}. In this case, however, the variance of Δ_T will be larger than of Δ_Q, because the use of a single realization of Ψ_{x_2} adds covariance to the sum of otherwise independent Gaussian random variables. The consequence of

using a single realization of Ψ_{x_2} is hence the introduction of additional variability to the decision variable, which is counter-intuitive with the fact that the task for triads feels easier than for quadruples. Whether the equivalency between the method of triads and quadruples holds true is still an open question that needs to be addressed experimentally.

13.2 ANCHORING OF THE SCALES

MLDS estimates *interval* scales, and as such they provide representation of *interval differences*. For interval scales any linear transformation of the type $x' = a \cdot x + b$ is allowed, which keeps the representation of interval differences meaningful. Interval scales need – by definition – an arbitrary origin or zero value (Krantz et al., 1971). The GLM in MLDS is designed in a way that the origin is effectively defined at zero, making perceptual scales by default anchored at their minimum (Section 3.2).

Scales anchored at zero

The arbitrary nature of the scale's origin can be problematic when scales need to be compared with each other, such as in the present case (Wiebel et al., 2017). Here scales for different viewing contexts were measured independently, and a default origin at zero was assumed for all of them. In this study an equal origin translates into assuming that the perception of the lowest reflectance is equivalent among all viewing contexts, i.e. the darkest reflectance is perceived equally black. We discussed in Chapter 8 that, in this case, a problem of anchoring among the scales is unlikely, as the predicted matching data were consistent with the data acquired independently in the matching experiment. However, this equivalency may not apply to all experimental cases and it should be carefully considered when designing an experiment.

Other scaling methods, such as pair comparisons in Thurstonian scaling, are based on the judgment of interval differences as well, and therefore also produce interval scales with an arbitrary origin (with the exception of magnitude estimation that produces ratio scales). A potential problem of anchoring is therefore common to all scaling methods when scales are measured and compared among different contexts.

It may be possible to test the equivalency of the scales' origin with other method when in doubt, e.g. with an matching task (method of adjustment) for the lowest stimulus value. In Wiebel et al. (2017) we argued that the good correspondence between the estimated scales and the matching data suggests that there was no problem with the anchoring of the scales. However, it cannot be argued that MLDS is a better method than matching in probing the internal perceptual scale, if we justify the scales' anchor using the matching data; this argumentation would be circular.

It remains to be tested whether giving the scales other anchor(s) improves or worsens the predictions derived from them. In principle it is possible to modify the GLM in MLDS to allow an origin different than zero. It would require the definition of an alternative design matrix that drops not the first but any other column, setting the origin to that corresponding stimulus value (Section 3.2). This may be advantageous under theoretical considerations that would prefer another anchor for the scales, e.g. the maximum.

13.3 MLDS AND MULTIPLICATIVE NOISE

Scaling methods that rely on the measurement of JNDs, as Fechnerian scaling, fail to provide meaningful scales if the noise is multiplicative (Chapter 2). The fact that MLDS can estimate a perceptual scale in the presence of multiplicative (unequal-variance) noise is a clear methodological benefit (Chapter 4), given that the type of noise distribution cannot be tested experimentally.

Multiplicative noise

The multiplicative noise case tested in these simulations correspond to the most studied case of unequal-variance: the noise increases with the stimulus dimension as obeying Weber's law. In discrimination methods, signal detection theory does provides statistical tools to fit multiplicative noise models, e.g. in yes/no or 2-IFC tasks (DeCarlo, 1998; McNicol, 1972; Knoblauch & Maloney, 2008). MLDS in its current implementation does not allow the definition and estimation of such a statistical model. However, it is feasible to extend MLDS to include *heteroscedastic* (unequal-variance) noise.

The estimation would involve the use of modified link functions and a similar maximum likelihood estimation procedure. An extension could be developed by

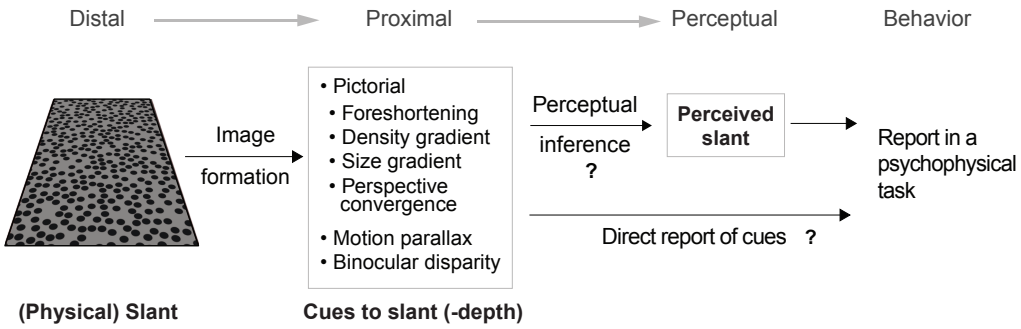


Figure 13.1: Distal, proximal and perceptual dimensions in slant-from-texture.

using existing packages in the *R* language environment (package *glmx*, Zeileis, Koenker, & Doebler, 2015), which implement binary GLMs with heteroscedastic noise. It remains to be tested whether this extension to MLDS can be used in the context of perceptual judgments of appearance, and whether it provides meaningful predictions of sensitivity.

13.4 MLDS AND OBSERVER MODELS

Ideal observer analysis is an useful framework for testing potential mechanisms of perceptual inference in the visual system (Geisler, 2011). In the study presented in Part II, ideal observer models were used to simulate two possible behaviors in lightness perception: one observer model that is fully lightness constant, and another that has only access to luminance information. These two models are extremes of possible behavior giving the ambiguous luminance signal in the retinal array. When these models were used to simulate responses to an 'mock' MLDS experiment, the scales predicted by MLDS were sufficiently different up to a certain noise level limit which was below what is commonly found experimentally (Section 6.4). MLDS could thus distinguish between these models because the mapping between the proximal and perceived dimensions was sufficiently different between them (Figure 6.1B).

Pictorial cues

The case of slant-from-texture in Part III is different. In slant-from-texture, multiple pictorial cues in the proximal retinal array are available to the visual system (Figure 13.1). These pictorial cues has been defined as: (i) foreshortening, the change in the aspect ratio of the texture elements; (ii) size gradient, the change in the size of the texture elements; (iii) density gradient, the change in the density of the texture elements; and (iv) perspective convergence, the tendency of the texture elements of converging towards the center due to linear perspective (Cutting & Millard, 1984; Saunders & Backus, 2006).

Figure 13.2A shows how some of the pictorial cues are calculated from a texture stimulus. Any pictorial cue could be used in isolation to infer slant by inverting the process of image formation. However, it is still an open question which cues the visual system uses in slant-from-texture (Knill, 1998; Saunders & Backus, 2006; Todd et al., 2010) For many pictorial cues the mapping between the proximal and the distal (slant) dimensions follow functions with similar curvature, as shown in Figure 13.2B.

Figure 13.2C shows the results of a pilot experiment using MLDS and slant-from-texture stimuli for one observer. The blue lines and errorbars show the (normalized) difference scale when the observer was asked to judge perceived slant, and the yellow, red, and orange lines show the scales from simulated observer models using the pictorial cues from Figure 13.2A. The scales were found not to correspond to any of the cues in isolation, for five different observers (Figure 13.2C). Rather, the results suggest that the scales actually follows a mixture of cues: a foreshortening cue at low slant values and a scaling contrast cue at higher slant values.

Scaling contrast

The 'scaling contrast' cue is calculated as the ratio difference between the widths of the elements in the uppermost and lowermost regions of the texture (Figure 13.2A), and it has been proposed to be a cue used by observers when judging slant-from-texture tasks (Todd, Thaler, Dijkstra, Koenderink, & Kappers, 2007; Todd et al., 2010). It can be seen in Figure 13.2C that the difference scale does not follow the scaling contrast cue. However, these results does not reject the possibility that scaling contrast is a vehicle for the judgment of slant which can be used in the visual system's computations, and that for observers it is simply not possible to report pictorial cues directly. Todd et al. (2010) used

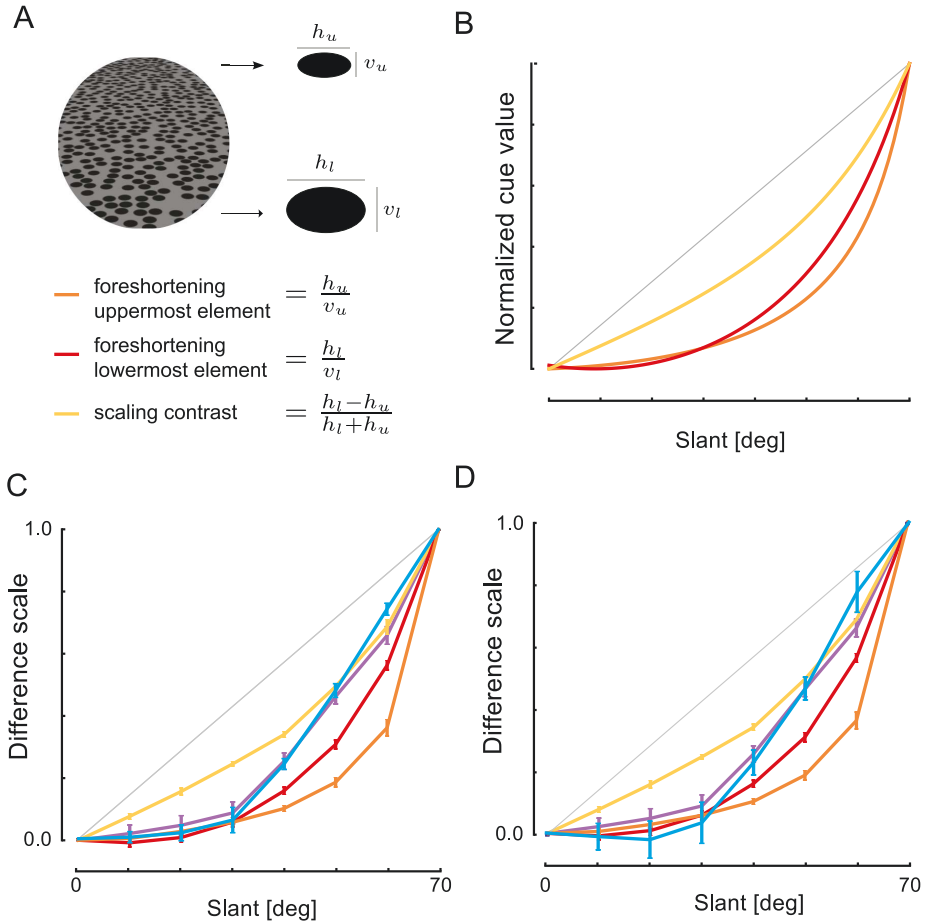


Figure 13.2: Pictorial cues in slant-from-texture and observer models. (A) Examples of two-dimensional pictorial cues in slant from texture. (B) Cues as a function of physical slant. (C) Difference scale from one observer judging perceived slant (blue), together with simulated observer models using the pictorial cues from (A). (D) Difference scale from the same observer as in (C) but being asked to judge the foreshortening of the uppermost elements.

a modified version of a partition scaling task that required the adjustment of perceptually equal intervals. The use of an adjustment task in their study may account for the differences found in our study.

When the same observer in Figure 13.2C was asked to judge the foreshortening of the uppermost elements, keeping stimuli and experimental procedure otherwise identical, the obtained scale was not different (Figure 13.2D). It is not possible to disentangle with a negative result whether the visual system is not using foreshortening or scaling contrast to infer slant, or it is simply not possible for the observer directly report proximal cues.

Thus, MLDS can be used with ideal observer models but only when the models are different enough with respect to the mapping among the distal, proximal and perceptual dimensions, as is the case for lightness. The development of these models depend on theoretical considerations and does not have to do *per se* with MLDS.

13.5 OUTLOOK: TOWARDS MORE REALISTIC STIMULI

This thesis evaluated the use of MLDS for the reliable measurement of perception, highlighting its advantages that have been already discussed. As critical as the psychophysical method chosen to measure perceptual phenomena is the type of stimulus that are used to probe the visual system (Figure 1.3). As outlined in Chapter 1, adequate and unambiguous stimuli may be more suitable for the study of vision by providing more realistic stimulation and thus avoiding unwanted strategies and confounds. The issue of realism in the stimulus has been lately a focus of attention in the field (Fleming, 2016; Brainard & Radonjic, 2016). In the following I discuss this issue, outlying some possible directions aiming to increase the realism of stimuli used in visual perception research.

13.6 SIMPLE VS. REALISTIC STIMULI

In lightness perception research, stimuli of diverse naturalness and complexity have been used. On one extreme, simple and flat stimuli have been used, such as

the simultaneous contrast display (Figure 1.2B) that only have luminance information without a clear distal structure as its origin. These type of stimulus are usually presented in flat monitor screens and show no cues to depth, conditions that are far away from naturally-occurring objects and surfaces. On the other extreme, lightness perception has been also studied using real physical stimuli, made of paper and cardboard, and real sources of illumination (e.g. Gilchrist, 1977, 1980). There is no doubt that these type of stimuli are realistic, however they are also impractical. They usually need complicated experimental apparatuses and the data collected are usually limited due to practical constraints on time.

In studies that find lightness constancy to be variable or even low, the finding is commonly attributed to the procedures or even the instructions used (Arend & Goldstein, 1987; Radonjić & Brainard, 2016), which would not constrain well responses based on lightness. Experimentally it may be possible that lightness constancy is truly low or variable, but it may as well be that the stimuli used are too simple and not realistic enough to stimulate the visual system in a naturally and appropriate way (Koenderink, 1999). Realistic stimuli are thus needed in order to reliably determine in which conditions lightness constancy is low, and develop successful models that must predict these deviations from constancy.

It seems to be difficult to design stimuli that are at the same time well-controlled and realistic. One solution is the use of computer-generated scenes by ray-tracing techniques (e.g. Heasly, Cottaris, Lichtman, Xiao, & Brainard, 2014). Ray-tracing rendering aims to simulate the behavior of the light realistically, providing well-controlled stimuli and efficient data collection. A high degree of constancy can be found in these type of experiments – as in the study in Part II – which tell us that these stimuli constrain well the perceptual response of judgments based on lightness.

Ray-tracing

However, ray-traced scenes are still presented in flat monitor screens, and some experimental evidence suggests that the use of monitors can be detrimental for perceived depth (Watt, Akeley, Ernst, & Banks, 2005; Hoffman, Girshick, Akeley, & Banks, 2008). In normal viewing of the natural world, our eyes' vergence is correctly aligned to a specific depth plane at the point of interest. When the eyes fixate at this depth plane, the lens accommodates its refractive power to

Depth cues conflict

match the plane, and consequently retinal blur occurs normally for depth planes that are farther or closer i.e. the eye's depth of field. These phenomena of vergence, accommodation of the lens, and retinal blur, collectively known as focus cues, dynamically signal the correct depth while our eyes look at the real world. On computer displays, however, focus cues signal flatness at a fixed distance to the screen. Therefore, focus cues are in conflict with experimentally added cues that will signal the desired depth plane (e.g. using pictorial or binocular disparity cues). Most of depth perception research has treated focus cues as negligible, however it has been shown that when cue conflict is avoided, depth and realism is improved (Hoffman et al., 2008).

In the following I propose two ways of increasing realism in experiments studying depth and lightness perception. These ideas were developed as research projects for this thesis, and are suggested here as an outlook on how to increase realism in psychophysical experimentation on depth and lightness.

13.6.1 *Increasing realism by adding motion parallax*

Motion parallax

Motion parallax is the relative motion of the retinal input due to head movements, and it signals relative depth. Computing relative depth can be accomplished by comparing the relative motion of objects in the retinal input (Rogers & Graham, 1979). Motion parallax is a strong cue to depth, and it is known that even small head movements can result in a detectable signal to be used to retrieve depth information (Rogers & Graham, 1979; Aytekin & Rucci, 2012). However, the contribution of motion parallax is omitted in almost all psychophysical experiments that work with depth. When inevitable small head movements occurs in standard psychophysical setups, these will signal flatness of the scene, hindering 3-D realism. Consequently, adding correct motion parallax to stimuli presented on screens will most likely increase the realism of the scene.

Rogers and Graham (1979) first studied motion parallax by dynamically adjusting the visual stimuli according to the observer's head position. The head rested on a chinrest that could move horizontally in front of the screen, and its position was used to update the stimuli accordingly, simulating correct perspective transformation (Rogers & Graham, 1979). This technique, also called head-joked

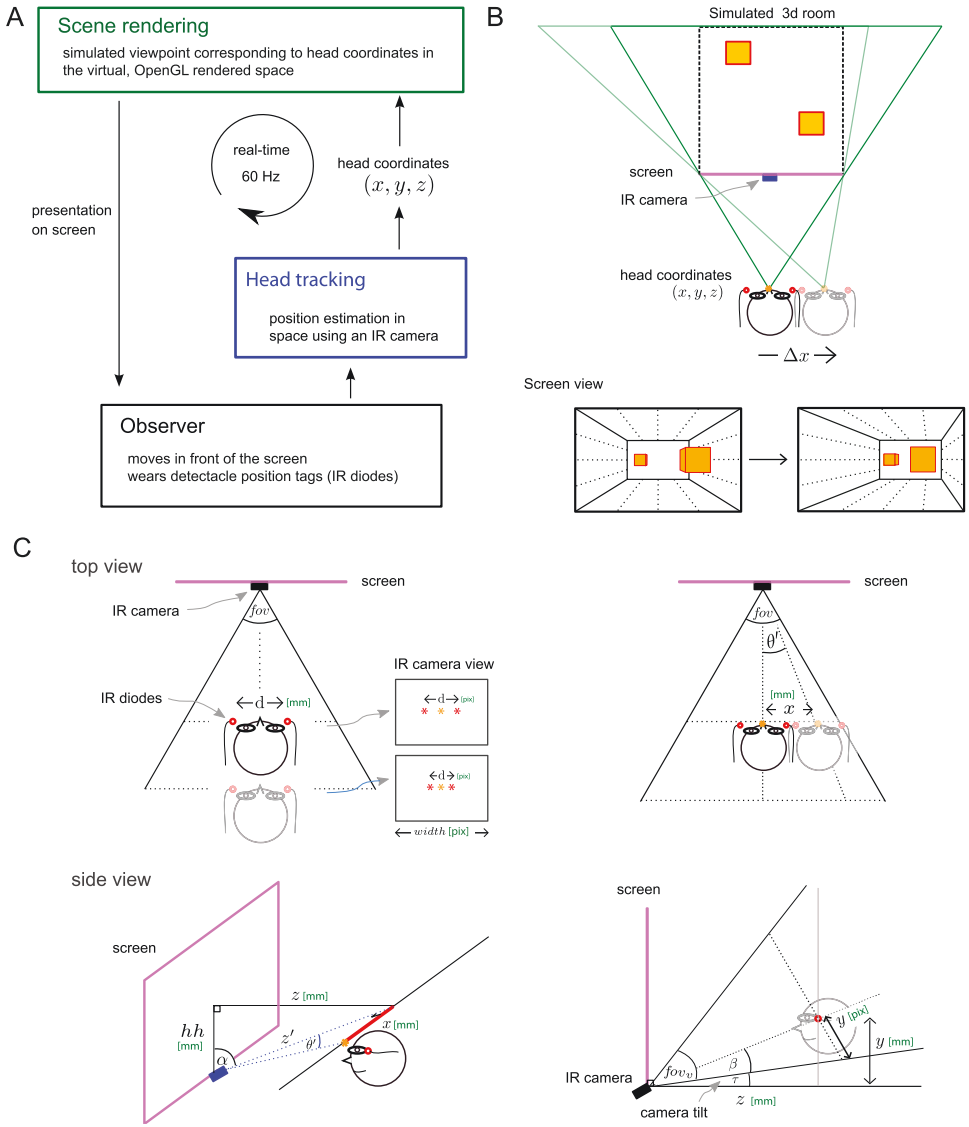


Figure 13.3: Adding motion parallax to stimulus presentation in a monitor screen. (A) Set-up sketch. The head position is tracked and used to update in real time the rendered scene. (B) The scene changes accordingly to the head position, e.g. when the observer moves to the right, the projection adapts as to simulate an identical movement in the camera position used for rendering. (C) Head tracking is accomplished by using an infrared camera placed at the bottom of the screen, and an observer that wears infrared diodes attached to a glasses' frame. The head position in space can be calculated using trigonometric rules.

method, allowed the experimenters to use motion parallax as an independent cue for depth perception. The proposed setup draws from this method, by tracking digitally the viewpoint of the observers and adjusting in real-time the perspective projection of the rendered scene according to the observer's viewpoint (Figure 13.3).

Head tracking

The first component required in the setup is head-tracking: we must track and record the position of the head and/or eyes of the observer. An easy and simple option for head-tracking is the use of infrared technology (other commercial technologies based on radio are also available). The hardware consist of an infrared (IR) camera and infrared emitting diodes mounted on a glasses' frame (see Figure 13.3B and C). The IR camera is placed centered below the screen pointing towards the observer, and the observer wears glasses with IR diodes attached to the sides. The distance between the diodes is known and fixed, as well as the camera's field of view. It is thus possible to estimate the head distance to the camera and consequently to the screen, and derive the coordinates in space using trigonometry and coordinate transformations (Figure 13.3C).

Dynamic rendering

The second component for this setup is the rendering of the simulated scene in 3-D with varying viewpoint that matches the head position (see Figure 13.3B). This can be accomplish using real-time rendering software, such as OpenGL or ray-traced pre-rendered scenes. The goal is to generate the illusion of a virtual scene that spans in depth, which is looked from the observer's point of view through the screen. The screen acts as a sort of 'window' to a simulated, virtual room (see Figure 13.3B). To accomplish this effect, the estimated position of the middle point between the IR diodes (which approximates to the middle point between the eyes) is used as the camera position or viewpoint in the rendering tool. When the observer moves, for example to the right as shown in Figure 13.3B, the perspective projection in the rendering changes accordingly, giving the impression of a real room in depth with its contents.

The setup for adding motion parallax was implemented during this thesis and a basic rendering example was tested. The example was a simulated room in 3D containing geometric volumes, with a visible framework for the room's walls that provided additional linear perspective cues (Figure 13.3B). The added motion parallax increased the realism of depth of this scene. However, it was

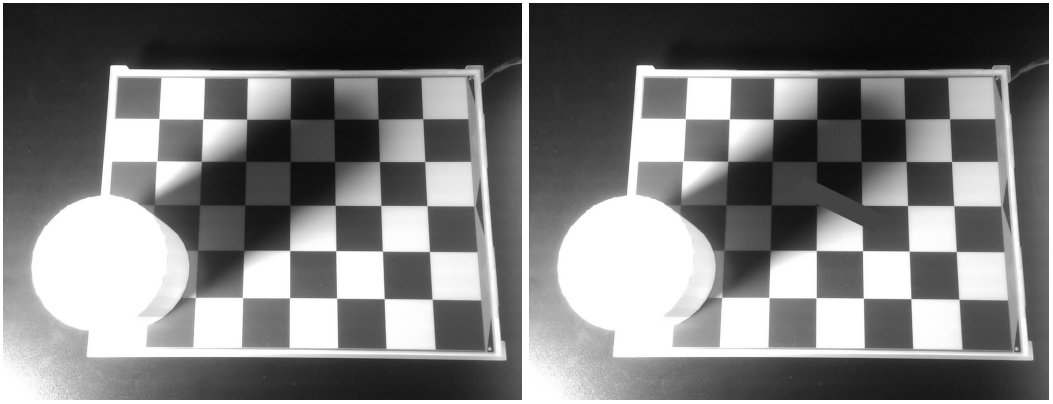


Figure 13.4: Adelson's checker-shadow illusion using a e-ink display device. Two checks, one in shadow and one in plain view, are equal in luminance but different in lightness. The right panel shows the image with a superimposed bar between the two checks; the left panel shows the image untouched.

not possible to increase the realism in the case of slant-from-texture, because, as Figure 9.1 shows, stimuli in slant-from-texture are seen through flat apertures that do not allow the presentation of the room's walls or any other perspective cues. It seems that the dynamic change of perspective according to the viewpoint is necessary for motion parallax to be useful as a depth cue. It remains to be tested whether the setup increases realism for more complex scenes, such as for the Adelson checker-shadow stimulus.

13.6.2 *Increasing realism by using an electronic ink display*

The fact that stimuli are presented in (flat) monitors may be critical for the study of lightness perception, because monitors, unlike real surfaces, *emit* light rather than *reflect* it. This could produce various confounds when lightness perception is probed. Koenderink (1999) critically expressed the concern that, with the use of computers, the modern study of visual perception deals more with computer graphics pipelines rather than with actual perceptual inference.

E-ink displays

An alternative approach is using real stimuli, but instead of using actual paper and cardboard, using electronic ink (e-ink) displays. These displays aim to resemble real paper, and they can be programmed to dynamically change their content. They do not emit light, rather, they contain microparticles filled with white or black ink that reflect light as normal matte paper does. These microparticles contain ink electrically charged, with different polarity for white and black ink. By varying the electrical field between the device's surface, the amount of black and white particles in each pixel can by controlled and manipulated, and thus the reflectance of the display.

The availability of these devices has grown in recent years, being now commercially available. During this thesis I explored the possibility of studying lightness perception using this type of device. Figure 13.4 shows the Adelson's checker-shadow illusion constructed with real stimuli. The checkerboard surface is an e-ink display (13.3' development kit, *Visionect Ltd*) that allows the change of its content digitally. The light source was a LED light coming from the left, and a white plastic cylinder was 3-D printed (Ultimaker 2, *Ultimaker B.V.*).

Adelson checkerboard on e-ink display

To construct the illusion, lookup tables were measured in order to determine the mapping between luminance and the device's reflectance, for check positions in shadow and plain view. These lookup tables depend critically on the geometry of the scene (light source position, cylinder position and other reflecting surfaces). Once measured, the checkerboard image was rendered with reflectance values that are thus known to produce equiluminant checks (20 cd/m^2). These two checks, although equiluminant, are perceived as light (shadow) and dark (plain view) in lightness. The left panel in Figure 13.4 shows a picture of the setup, and the right panel shows the same picture modified with a superimposed gray bar, indicating the equivalence of pixel grayvalues.

This setup promises to be useful in studying open questions regarding mechanisms of lightness perception such as assimilation or the effect of edges and contrast in lightness inference under naturalistic conditions One immediate application would be to measure the effect size in the Adelson's checker-shadow illusion using this setup, and comparing it to an analogous experiment using a ray-traced scene with same geometry and luminance specification. Subjectively, the effect in the Adelson checkerboard becomes more pronounced for the actual

real stimulus when compared to a picture of the same stimulus (Figure 13.4), which is consistent with the finding that the increase of realism makes the effect size more pronounced (Maertens et al., 2015). The magnitude of the effect change with this setup remains to be quantified; and more generally, it also remains to be determined whether the use of real stimuli indeed introduces more realism, and in this way a more reliable probing of the processes involved in lightness perception.

13.7 CONCLUSIONS

I propose in this doctoral dissertation the use of MLDS as a reliable tool for measuring perception. MLDS is a method based in appearance judgments of clearly visible stimulus differences, and it allows the estimation of perceptual scales in an efficient and reliable way, by using the method of triads which provides an intuitive and easy task for the observer. It avoids the shortcomings of performance-based (discrimination) methods that use judgments of small, near-threshold stimulus differences in difficult and unintuitive tasks.

In this thesis numerical simulations were first used to establish the accuracy and precision of MLDS, as well as the effects of violations on its model assumptions. MLDS was tested experimentally in the domain of lightness perception, where it was successful in estimating scales that reflected lightness constancy. The task required only within-context comparisons, and MLDS seems superior to asymmetric matching procedures as the latter provide only indirect measurements of the underlying scales. Additionally, matching data predicted by the perceptual scales followed closely the data acquired in an independent asymmetric matching procedure, further suggesting that MLDS indeed probed the internal representation of lightness.

MLDS was also framed in a signal detection theory formulation, and its performance for deriving sensitivity was established using simulations. MLDS was more efficient and quantitatively analogous than traditional performance methods in the estimation of sensitivity, but only when MLDS model assumptions were satisfied. This equivalency was tested experimentally in a slant-from-texture task, from which sensitivity has been previously studied in the literature. We found

varying degrees of agreement that could be due to truly differences in the perceptual representation, or alternatively violations of model assumptions.

The use of MLDS seems promising for the reliable measurement of perceptual phenomena. Together with the use of realistic stimuli, MLDS offers a method to measure the perceptual dimension, and in that way enabling the testing of theoretical models of perceptual inference.

A

APPENDIX

This appendix contains supplementary information, figures and tables from the studies presented in Part II and Part III.

A.1 APPENDIX TO PART II: USING MLDS TO MEASURE LIGHTNESS

A.1.1 *Goodness of fit of difference scales*

In general for all our observers the fitted models had acceptable goodness of fit in all viewing contexts (Table A.1). In only two out of 50 cases the goodness of fit was significantly different that the expected under the model assumptions (O7 and O8 for plain view context). The 'deviance accounted for' (DAF) varied greatly between observers, in the range between approx. 17 % to 63 % and it was directly related to the level of noise estimated for each observer. Observers were sorted according to their noise level, from lower to higher (as shown as the scale's height in Figure A.1), and from Table A.1 it is evident that the DAF on average decreases as the estimated noise of the observer is higher.

Observer	plain	high transp. dark	low transp. dark	high transp. light	low transp. light	mean
O1	63.0	66.6	59.6	61.4	62.0	62.5
O2	57.7	58.5	55.3	57.6	52.2	56.2
O3	53.6	58.8	49.5	54.3	50.9	53.4
O4	44.6	54.8	51.5	50.4	52.8	50.8
O5	48.9	50.9	45.3	43.6	46.8	47.1
O6	49.2	50.1	45.3	50.5	43.9	47.8
O7	45.8	48.6	44.1	48.7	45.3	46.5
O8	38.8	44.6	42.7	47.0	51.7	45.0
O9	40.6	38.9	32.6	40.4	30.7	36.7
O10	17.6	18.9	14.4	19.0	13.8	16.7

Table A.1: 'Deviance accounted for' (DAF) as a measure of goodness of fit of the difference scaling model, for each viewing context and observer. High and low indicate high and low transmittance, and dark and light indicate a lower and a higher reflectance of the transparent medium.

Observer	plain	high transp. dark	low transp. dark	high transp. light	low transp. light
O1	0.17	0.52	0.32	0.31	0.19
O2	0.06	0.37	0.35	0.09	0.01
O3	0.17	0.05	0.05	0.07	0.48
O4	0.02	0.40	0.11	0.25	0.59
O5	0.03	0.06	0.14	0.35	0.12
O6	0.05	0.49	0.06	0.34	0.05
O7	0.003 *	0.06	0.03	0.10	0.30
O8	0.002 *	0.06	0.17	0.03	0.25
O9	0.14	0.03	0.02	0.62	0.13
O10	0.30	0.38	0.07	0.65	0.16

Table A.2: P-value from Monte-Carlo simulations as a measure of goodness of fit of the difference scale model (as explained in Section 3.5) for each viewing context and observer. Asterisk indicate $p < 0.01$. High and low indicate high and low transmittance, and dark and light indicate a lower and a higher reflectance of the transparent medium.

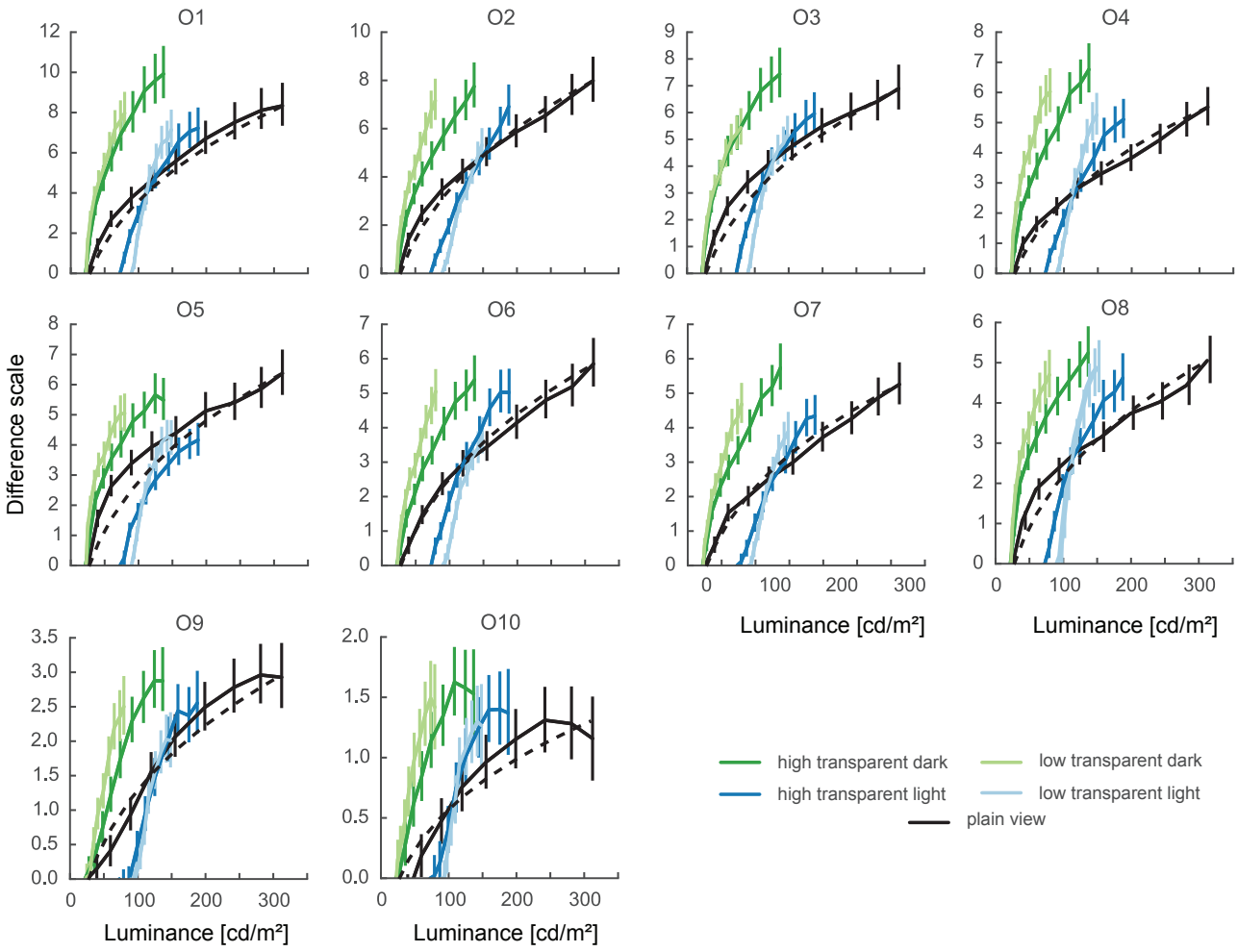


Figure A.1: Individual difference scales from all (n=10) observers. Difference scales were derived using MLDS for each viewing condition independently (colors). Despite that the viewing condition change the presented luminance range, the difference scales cover a similar range as in plain view. Error bars indicate 95 % confidence interval obtained using bootstrap techniques. Observers are sorted by the scales' maximum in decreasing order. The Munsell scale for plain view is also depicted (dashed black line).

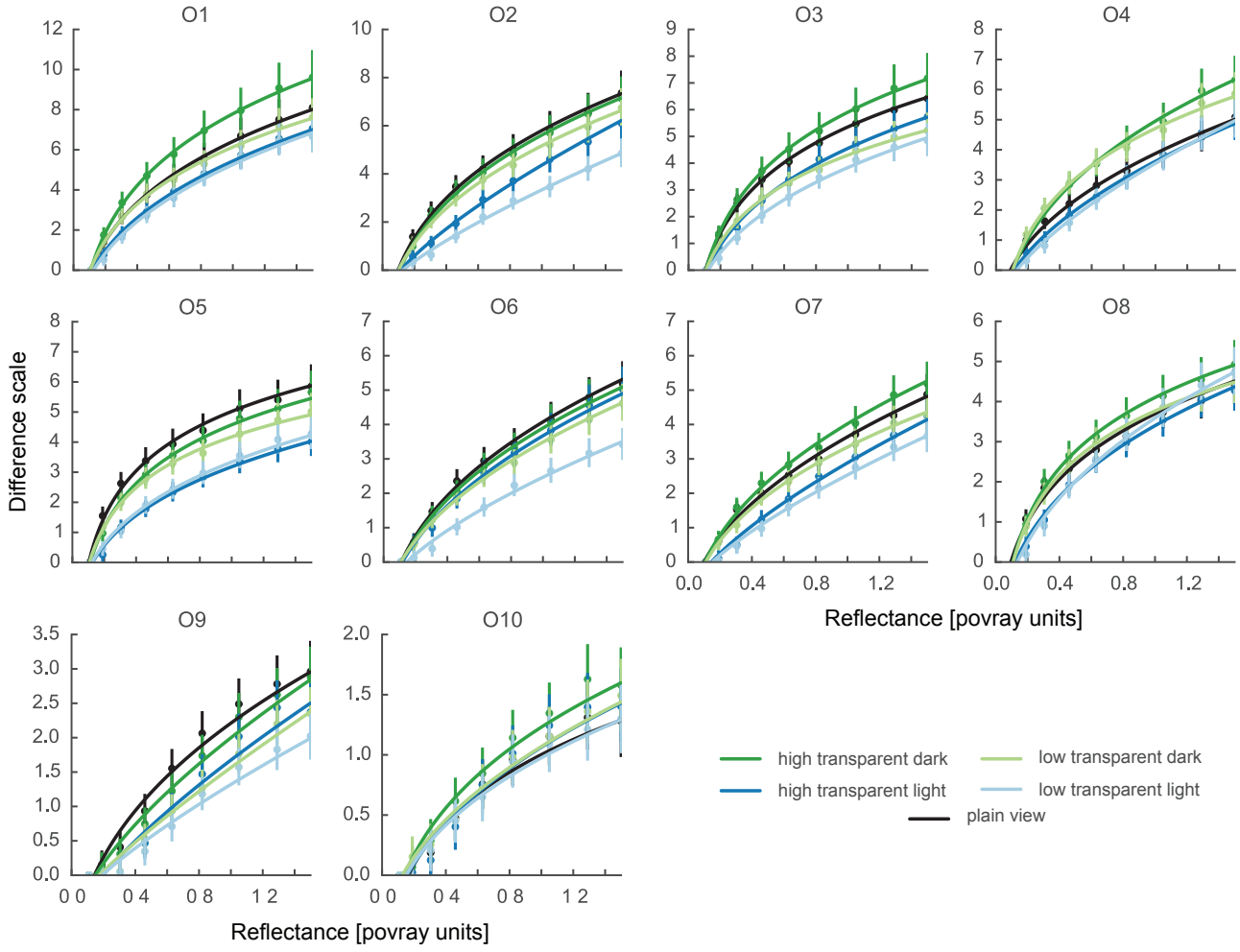


Figure A.2: Individual difference scales from all (n=10) observers. Same data from Figure A.1 but plotted as a function of reflectance.

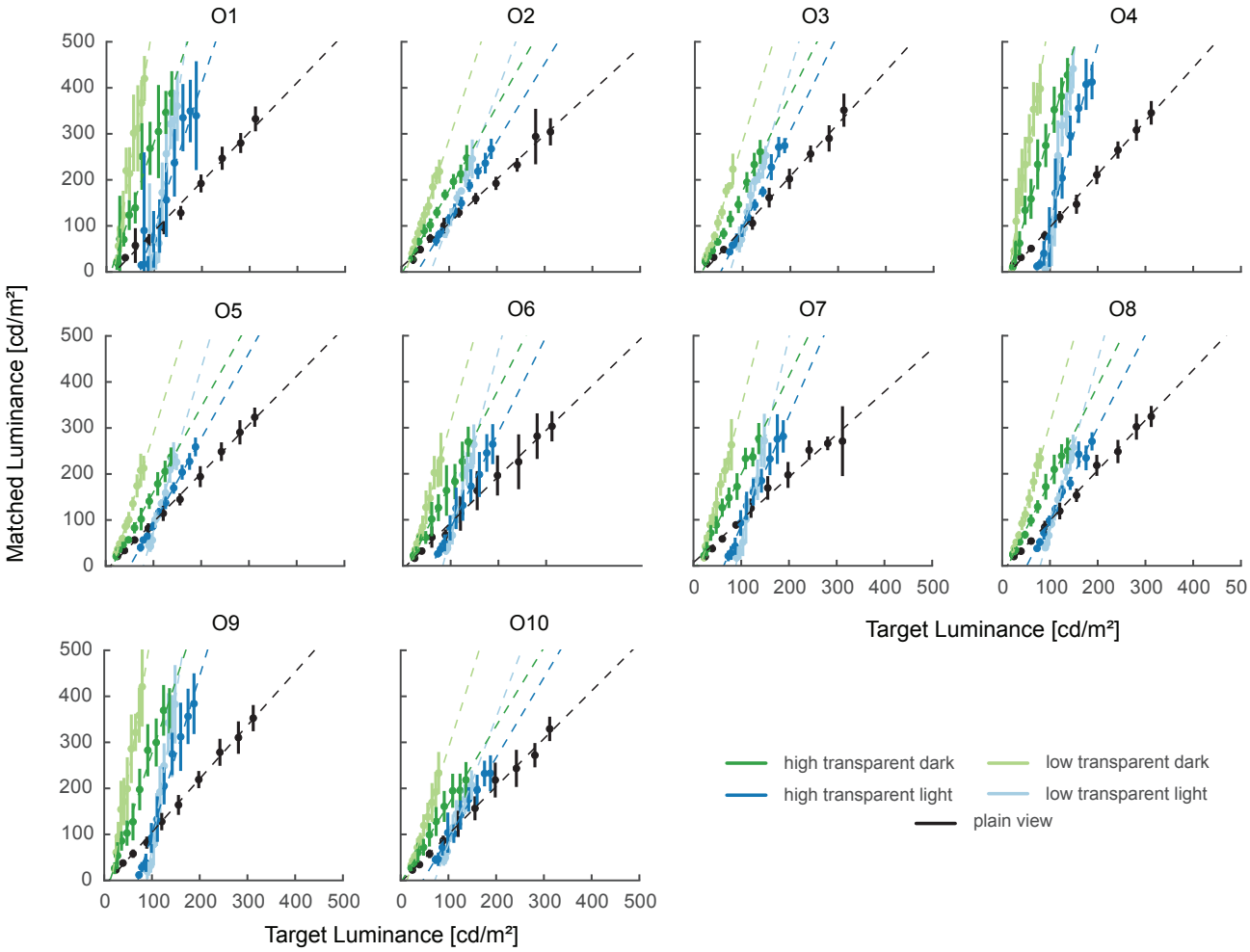


Figure A.3: Individual matching data of all (n=10) observers. Matched luminance is plotted as a function of target luminance for each viewing condition (colors). For each viewing condition a linear regression was computed and the fit is shown as a dash line.

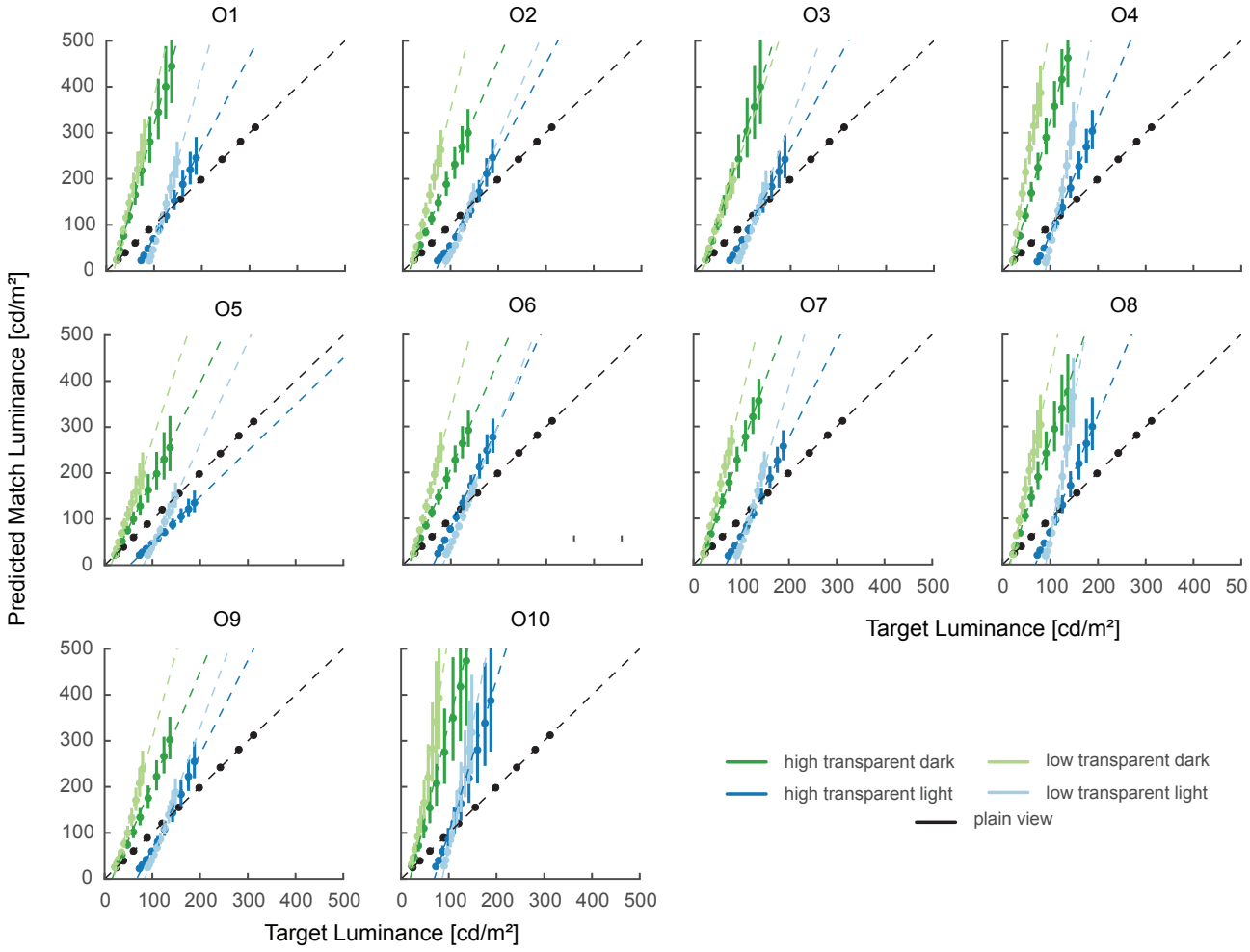


Figure A.4: Predicted matching data from MLDS of all (n=10) observers. Same layout and legend as in Figure A.3.

A.1.2 *Simulated observer models and noise*

We were interested in testing how good MLDS could distinguish the lightness constant observer model and the luminance-based observer. These two models represent two extremes of behavior for lightness judgments. We ran multiple simulations of the two observer models, and analyzed how often they could be told apart with increasing levels of noise. We aimed to establish an upper limit of noise at which the two models can reliably be separated using MLDS. We found that this limit of noise was considerably higher than the values found experimentally.

The simulated sensory functions were linear functions $\Psi(x) = ax + b$. For the lightness constant observer, there were five different sensory functions, one for each transparency condition, which were the inverse of the atmospheric transfer functions (ATFs, Figure 5.1C). Thus, the functions 'undo' the process of image formation by mapping the compressed range of luminance to the full range between 0 and 1 (Figure 6.1B top panels) For the luminance-based observer, a single sensory function was set which linearly mapped the full range of luminance values (x) to a range between 0 and 1 (Figure 6.1B bottom panels). The response to each of these functions was corrupted by Gaussian noise, with a fixed variance (σ^2).

The simulation proceeded using the same procedure as in the actual psychophysical experiment. The observer models responded to the method of triads, by drawing their response to each stimulus value in the triad. Triads were constructed as non-overlapping intervals using the ten reflectance values used in the experiment. In total $n = 120$ unique triads were simulated per block, and $n_b = 10$ blocks were simulated. The generated data was fed into the MLDS analysis procedure, and scales and their variability were estimated. Each simulation was repeated N=100 times per observer model, and the noise value σ was varied systematically in the range between 0.01 and 1.2.

We defined a measure of constancy in order to quality the type of output that simulations produced. For a given simulation we compared the scale's maximum derived for each transparency condition with the scale's maximum derived for plain view. When the 95 % confidence intervals of these two scale's maxima

overlapped, the simulation was qualified as with a 'lightness constant' result. This means that a lightness constant result is expected when the maxima of the scales coincide on their range in the y-axis. As the scales resulting from MLDS are - by design - anchored at zero on its minimum, only a comparison of the maximum is needed. In Figure 6.1B the overlap of scale's maxima is evident for the lightness constant model for all transparency conditions, and for none in the luminance-based observer.

The proportion of simulations that resulted in a constancy result (y-axis) is shown for varying levels of noise (x-axis) in Figure A.5. We found that as the noise level increases, the luminance-based model more frequently cannot be told apart from the lightness constant model. A complete separation between the two occurs when the estimated noise level is below $\hat{\sigma} = 0.4$. In this way, we established that when the scale's noise estimated with MLDS falls below this value, we can successfully distinguish between these two observer models.

A.1.3 *Normalized Contrast model*

In this work we applied the normalized contrast model introduced by Zeiner and Maertens (2014) and Wiebel et al. (2016). In brief, in a first step the Michelson contrast is computed between a target and its eight surround checks (x). That is, the target intensity X is normalized relative to its average local surround $S_{1,\dots,8}$ by

$$x = \frac{X - mean(S_{1,\dots,8})}{X + mean(S_{1,\dots,8})} \tag{A.1}$$

In a second step, the target Michelson contrast (x) is normalized relative to the contrast range in the region of transparency ($t_{max} - t_{min}$), and this range is subsequently mapped to the contrast range in plain view (p_{max} -p_{min}). The so normalized Michelson contrast (NMC) relates to the target contrast x according to the following equation:

$$NMC = \frac{x - t_{min}}{t_{max} - t_{min}} * (p_{max} - p_{min}) + p_{min} \tag{A.2}$$

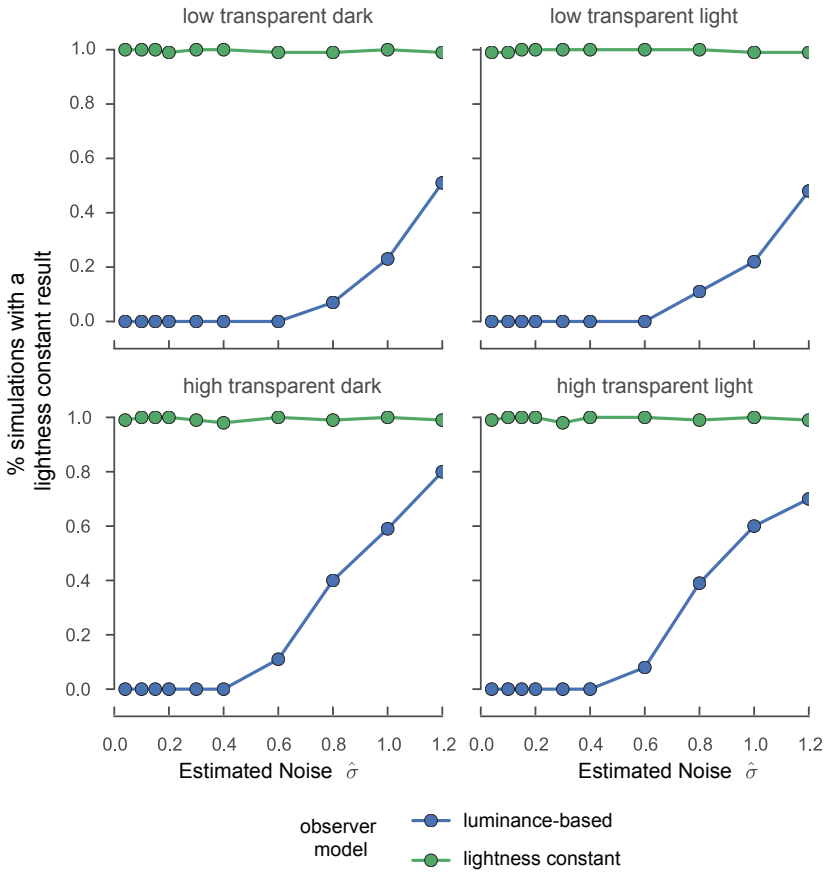


Figure A.5: Ability of distinguish between the two observer models. Percentage of simulations that produced a 'lightness-constant' output as a function of the estimated noise by MLDS and for both observer models.

Viewing context	Slope			
	empirical	predicted	t(9)	p
high transparent dark	2.4 ± 0.7	3.1 ± 0.7	-2.71	0.02
high transparent light	2.4 ± 0.8	2.2 ± 0.6	0.94	0.34
low transparent dark	4.4 ± 1.5	4.5 ± 0.1	-0.16	0.88
low transparent light	4.6 ± 1.8	3.7 ± 1.4	1.42	0.19

Viewing context	Intercept			
	empirical	predicted	t(9)	p
high transparent dark	-25.1 ± 15.6	-46.4 ± 17.6	3.08	0.01
high transparent light	-146.6 ± 75.7	-142.1 ± 44.9	-0.17	0.87
low transparent dark	-64.9 ± 23.8	-75.0 ± 27.2	0.90	0.39
low transparent light	-385.0 ± 191.5	-316.4 ± 128.1	-1.04	0.33

Table A.3: Linear regression results for data obtained in the matching experiment (empirical) in comparison to matching predictions derived using the estimated scales from the MLDS experiment (predicted). Mean $\pm$ S.D., p: p-value of paired t-tests between the empirical and predicted parameters.

Reflectance	Luminance [cd/m^2]				
r	plain	high transp. dark	high transp. light	low transp. dark	low transp. light
0.06	15	18	69	21	88
0.11	25	22	73	23	90
0.19	40	28	79	25	94
0.31	60	36	87	29	98
0.46	89	48	99	35	104
0.63	120	60	111	41	110
0.82	155	74	125	49	116
1.05	199	92	144	57	126
1.29	242	108	159	65	134
1.50	281	125	176	73	142
1.67	312	137	188	79	148
1.95	365	157	209	90	160
2.22	415	177	229	100	169

Table A.4: Target luminance values in different viewing conditions. The luminance values (in cd/m^2) corresponding to the 13 reflectance values (in *povray* units, first column) are shown for each viewing condition. Reflectance values from 0.11 to 1.67 were used as targets in the both experiments (MLDS and matching).

A.2 APPENDIX TO PART III: USING MLDS TO MEASURE SENSITIVITY

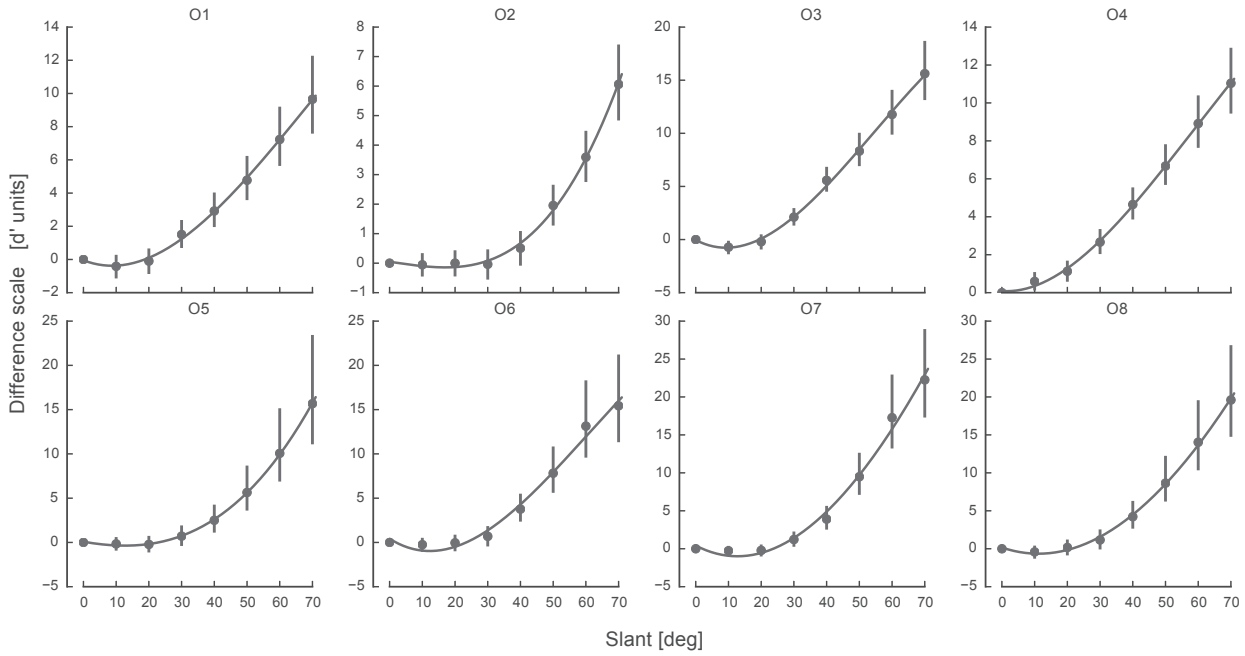


Figure A.6: Difference scales estimated by MLDS for all observers. Markers indicate discrete scale values obtained from MLDS, error bars indicate confidence intervals of scale estimates obtained by bootstrap. The continuous line indicate the cubic spline fitted to the discrete scale values.

A.2.1 *Goodness of fit of difference scales*

The goodness of fit was acceptable only for three observers (O2, O3, O4) when the data was fitted with the MLDS default parameters (probit link function with zero asymptotes). Consequently, we applied a workflow consisting on a series of refitting attempts until a satisfactory goodness of fit was achieved (as suggested in Knoblauch & Maloney, 2012, pp. 219-222). The goodness of fit statistics at

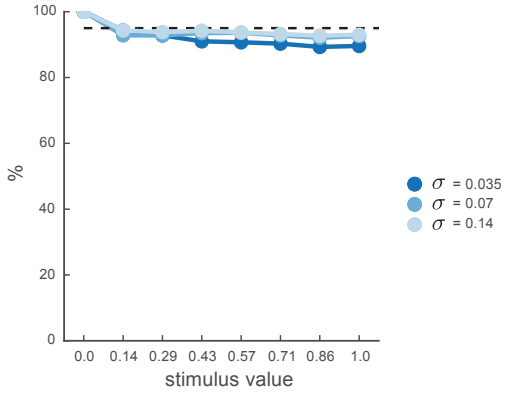


Figure A.7: Coverage analysis: % of simulations on which the 'true' value was included in the confidence interval of the estimated scale. At 95 % confidence, 95 % of the simulations (black dashed line) should include the true value. Tested at three different noise levels σ.

each step of the workflow for each observer is shown in Table A.5. First, we estimated the error rates from the data (analogous to guess and lapse rates in psychometric functions) and refitted the model with those error rates as non-zero asymptotes of the link function. We obtained non-significant p-values for the models of observers O5 and O7, and these scales were kept. For observers O1, O6 and O8, however, p-values were still significant. For these remaining observers, we then proceeded to manually divide the data into two halves, and we refitted the model for each half independently (420 trials each). When the goodness of fit of a scale from any of the two halves of data was found to be appropriate, it was kept for further analysis; this was the case for observer O1 and O6. For observer O8, both halves of the data were non-significant; in this case we choose the one that had a higher 'deviance accounted for'.

The data and goodness of fit statistics that were finally considered for the estimation of near-threshold performance are marked with a gray background in Table A.5, and the corresponding scales are shown in Figure A.6. Figure A.8 shows the goodness of fit plots as in Figure 3.3.

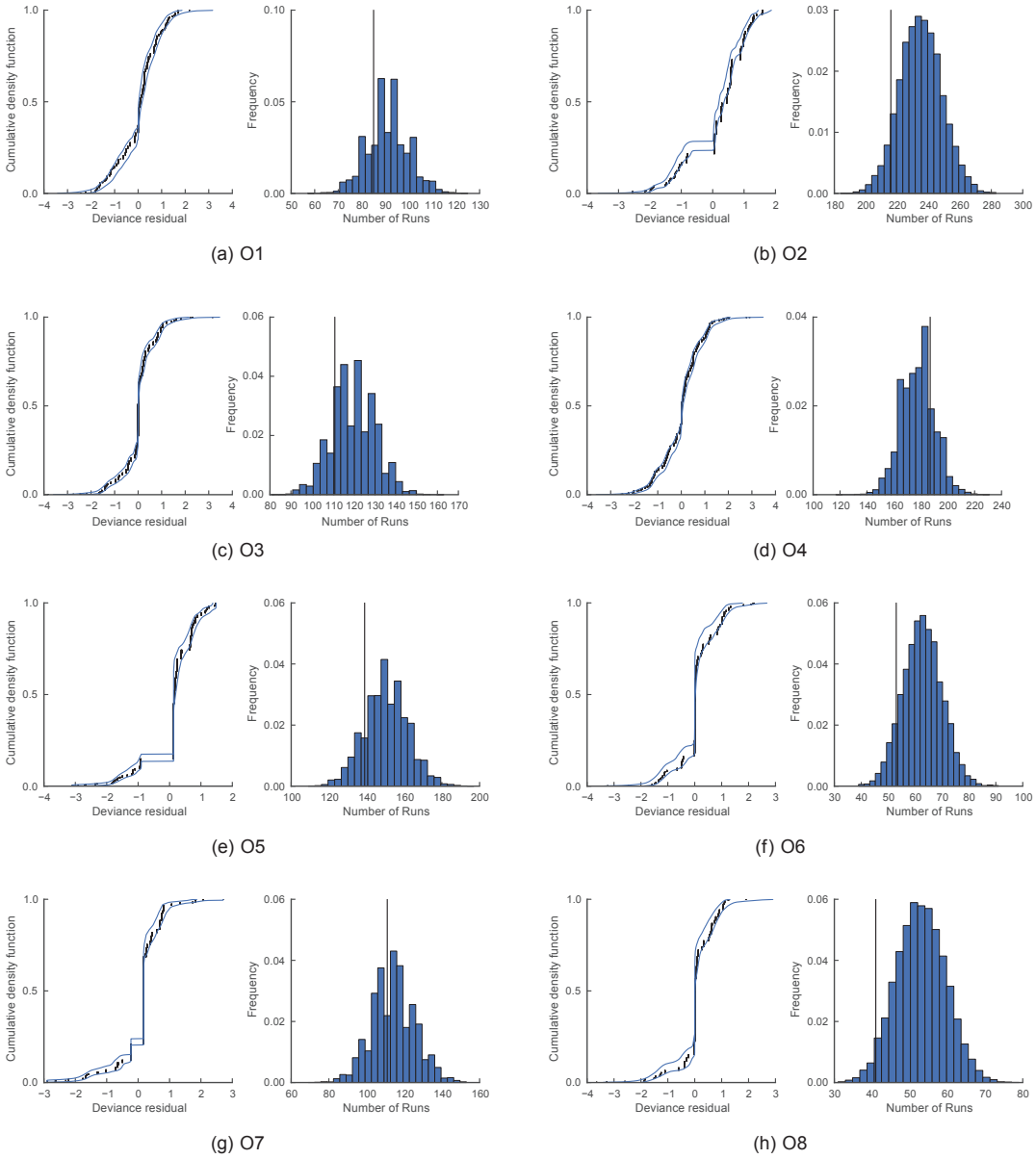


Figure A.8: Goodness of fit of difference scales obtained by Monte-Carlo simulation, as described in Section 3.5 and in Knoblauch and Maloney (2012). Each panel shows the two diagnostic graphs produced by the *MLDS* package for each observer.

		Asymptotes		GoF measure		
		lower	upper	AIC	DAF	p-value
O1	all	-	-	687	42	$< 10^{-3}$
	all	0.01	0.03	632	47	$< 10^{-2}$
	1/2	-	-	271	56	0.01
	2/2	-	-	299	50	0.28
O2	all	-	-	714	40	0.08
O3	all	-	-	383	68	0.23
O4	all	-	-	560	53	0.80
O5	all	-	-	520	57	0.01
	all	0.33	0.01	469	47	0.18
O6	all	-	-	508	58	$< 10^{-3}$
	all	0.06	0.01	492	57	$< 10^{-2}$
	1/2	-	-	201	68	0.10
	2/2	-	-	261	58	$< 10^{-3}$
O7	all	-	-	435	64	$< 10^{-3}$
	all	0.03	0.02	400	66	0.43
O8	all	-	-	390	68	$< 10^{-3}$
	all	$< 10^{-3}$	0.01	389	68	0.02
	1/2	-	-	222	64	0.05
	2/2	-	-	171	73	0.05

Table A.5: Goodness of fit measures and values of link function asymptotes used for the fitting of the difference scale model, for each observer and part of data considered. See text for a description of the workflow employed to achieve appropriate goodness of fit. AIC: Akaike information criterion; DAF: 'deviance accounted for'; p-value: p-value statistic based on Monte-Carlo simulation that accessed goodness of fit (see Section 3.5). Gray background indicate data that were considered for the analysis and the estimation of thresholds.

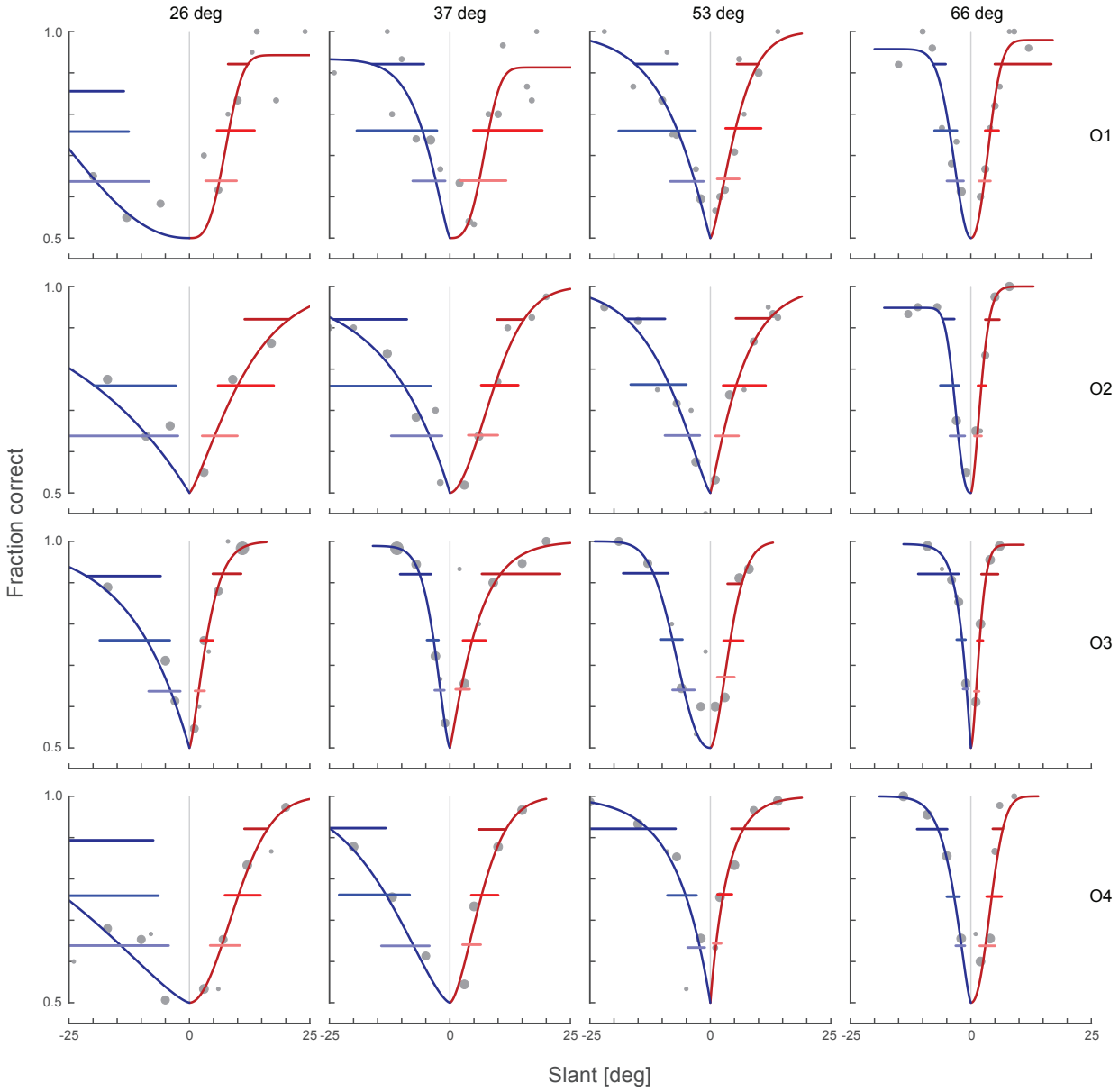
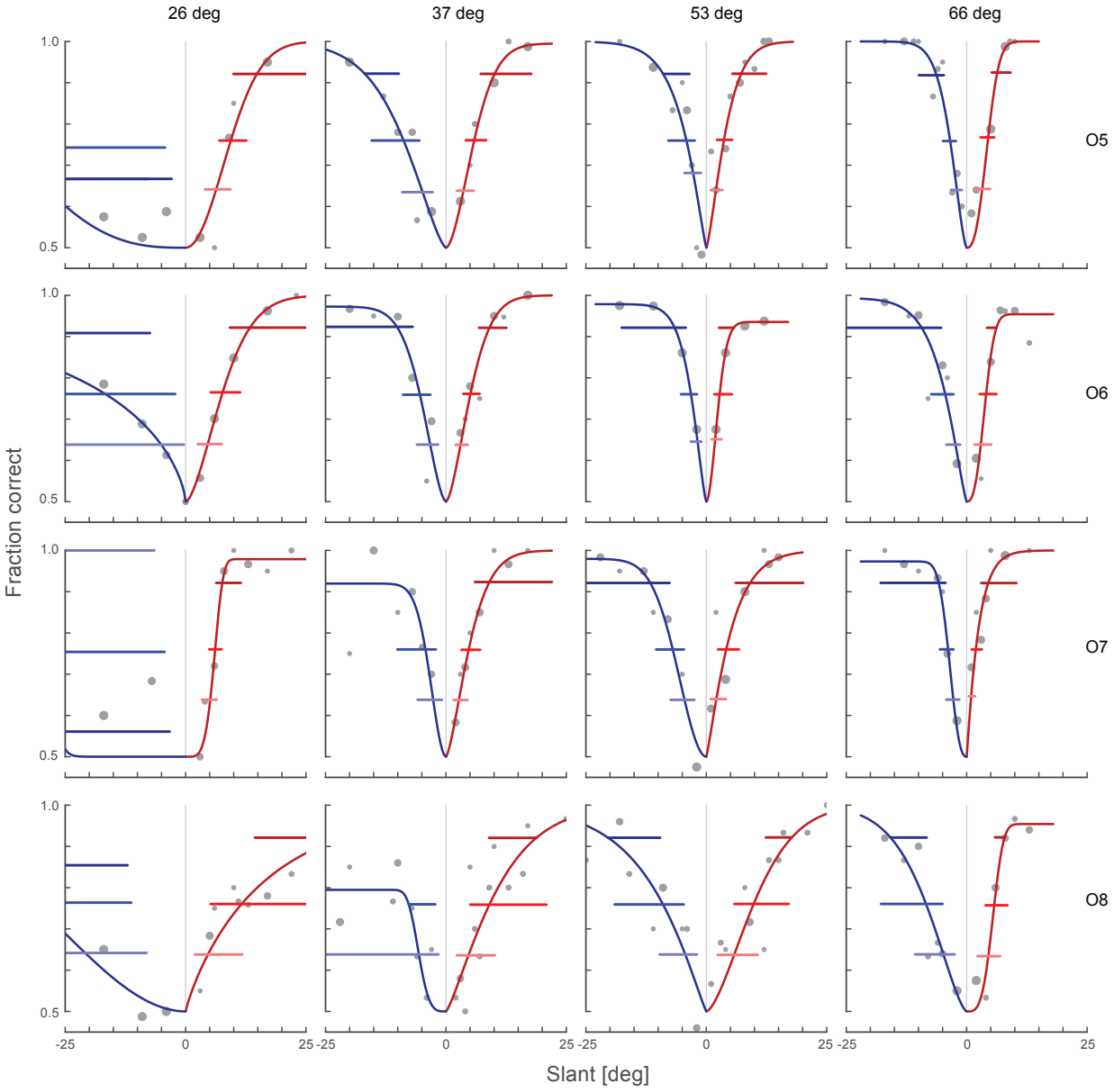


Figure A.9: Psychometric functions obtained in Experiment 2 for all evaluated standards values (columns) and observers (rows). Blue (red) lines indicate functions for comparisons below (above) the standard. Straight horizontal lines indicate 95 % confidence intervals for the thresholds calculated at performance of d′ 0.5, 1 and 2.


26 deg
37 deg
53 deg
66 deg
O5
O6
O7
O8
Fraction correct
1.0
0.5
-25
0
25
Slant [deg]

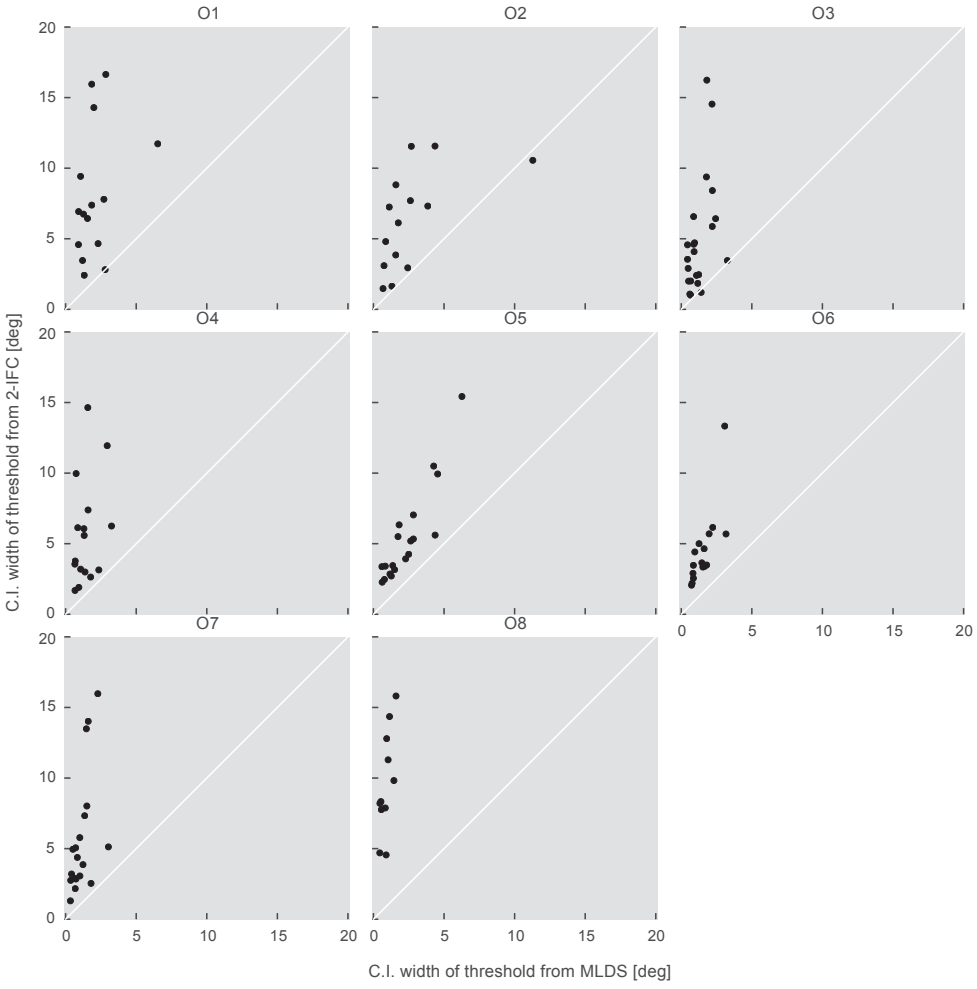


Figure A.10: Comparison of the variability in the threshold estimation. The width of the confidence intervals from Figure 11.2 and Figure 11.3 are plotted against each other for each observer individually. All standard values are plotted together.

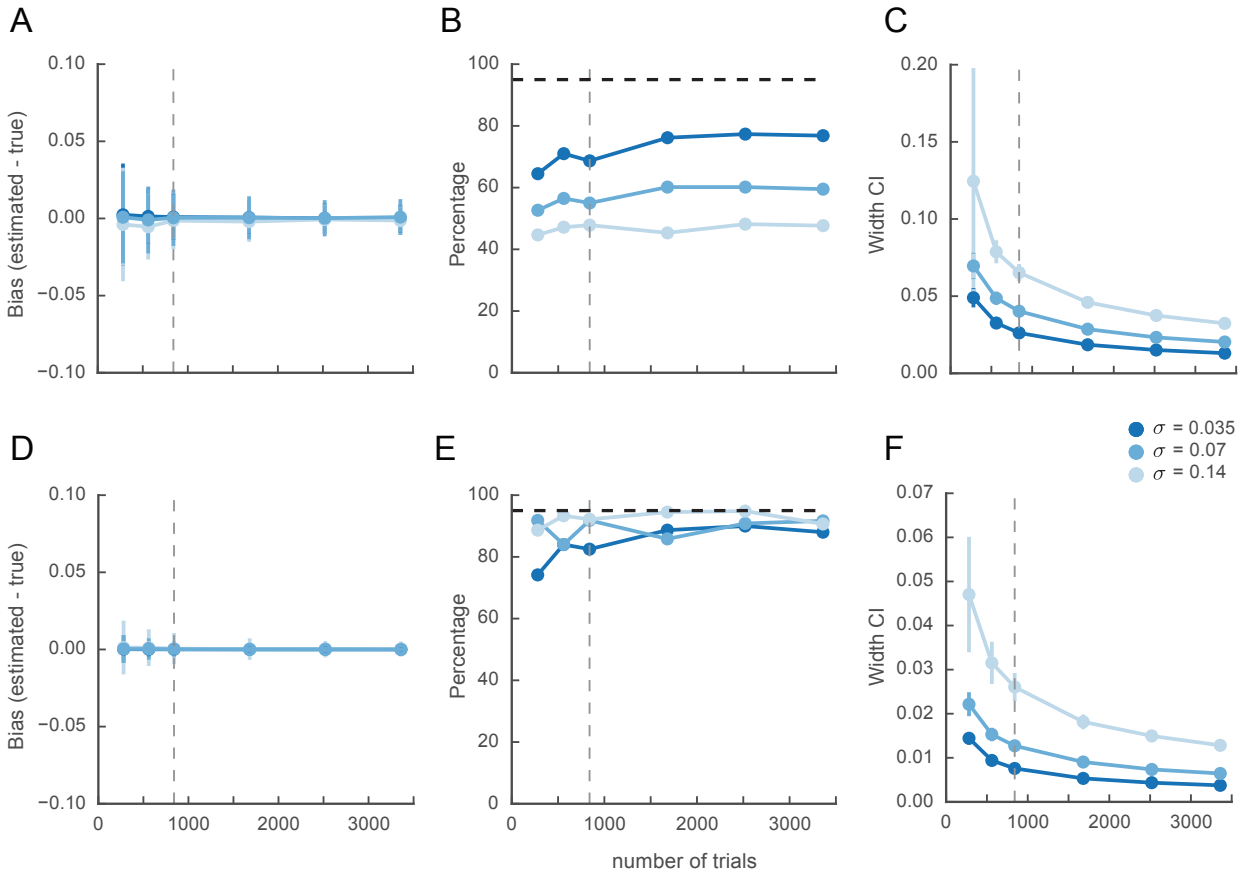


Figure A.11: Simulation results for threshold estimation using MLDS with variable number of trials. Bias (A, D), confidence intervals' coverage (B, E) and width (C, F) for the estimation of thresholds using MLDS as a function of the number of simulated trials, for three different noise values (σ). Panels A-C correspond to results for a standard stimulus value $st = 0.2$; panels D-F for $st = 0.8$. The vertical dashed line indicate the number of trials used in the actual experiments. The horizontal dashed lines in panel B and E indicate the expected coverage percentage. Errorbars indicate mean ± S.D. across n=100 simulations.

REFERENCES

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., p. 339-351). Cambridge, MA: MIT Press.

Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision*, *17*(1), 37. doi:10.1167/17.1.37

Anderson, B. L. (1999). Stereoscopic Surface Perception. *Neuron*, *24*(4), 919–928. doi:10.1016/S0896-6273(00)81039-9

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, *21*(24), R978–R983. doi:10.1016/j.cub.2011.11.022

Arend, L. E., & Goldstein, R. (1987). Simultaneous constancy, lightness, and brightness. *Journal of the Optical Society of America A*, *4*(12), 2281-2285. doi:10.1364/JOSAA.4.002281

Aytekin, M., & Rucci, M. (2012). Motion parallax from microscopic head movements during visual fixation. *Vision research*, *70*, 7–17. doi:10.1016/j.visres.2012.07.017

Baddeley, A., & Turner, R. (2005). spatstat: An r package for analyzing spatial point patterns. *Journal of Statistical Software*, *12*(6), 1–42.

Baird, J. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

Baird, J. (1989). The fickle measuring instrument. *Behavioral and Brain Sciences*, *12*(02), 269-270. doi:10.1017/S0140525X00048585

Barrow, H. G., & Tenenbaum, J. M. (1978). Recovering intrinsic scene characteristics from images. In A. Hanson & E. Riseman (Eds.), *Computer vision systems* (pp. 3–26). New York: Academic Press.

Brainard, D. H., Brunt, W. A., & Speigle, J. M. (1997). Color constancy in the nearly natural image. 1. Asymmetric matches. *Journal of the Optical Society of America A*, *14*(9), 2091–2110. doi:10.1364/JOSAA.14.002091

Brainard, D. H., & Radonjic, A. (2016). The use of graphics simulations

in the study of object color appearance. *Journal of Vision*, *16*(12), 2. doi:10.1167/16.12.2

Charrier, C., Maloney, L. T., Cherifi, H., & Knoblauch, K. (2007). Maximum likelihood difference scaling of image quality in compression-degraded images. *Journal of the Optical Society of America A*, *24*(11), 3418. doi:10.1364/JOSAA.24.003418

Chubb, C., Landy, M. S., & Econopouly, J. (2004). A visual mechanism tuned to black. *Vision Research*, *44*(27), 3223–3232. doi:10.1016/j.visres.2004.07.019

Cutting, J. E., & Millard, R. T. (1984). Three gradients and the perception of flat and curved surfaces. *Journal of experimental psychology: General*, *113*(2), 198–216.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychol. Methods*, *3*(2), 186–205. doi:10.1037/1082-989X.3.2.186

Devinck, F., & Knoblauch, K. (2012). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of vision*, *12*(3), 1–14. doi:10.1167/12.3.19

D'Zmura, M., & Iverson, G. (1993). Color constancy. II. Results for two-stage linear recovery of spectral descriptions for lights and surfaces. *Journal of the Optical Society of America A*, *10*(10), 2166–2180.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Ekroll, V., & Faul, F. (2013). Transparency perception: the key to understanding simultaneous color contrast. *Journal of the Optical Society of America A*, *30*(3), 342. doi:10.1364/JOSAA.30.000342

Emrith, K., Chantler, M. J., Green, P. R., Maloney, L. T., & Clarke, a. D. F. (2010). Measuring perceived differences in surface texture due to changes in higher order statistics. *Journal of the Optical Society of America A*, *27*(5), 1232. doi:10.1364/JOSAA.27.001232

Fechner, G. (1860). *Elemente der psychophysik*. Leipzig: Breitkopf und Hartel.

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research*, *94*, 62–75. doi:10.1016/j.visres.2013.11.004

Fleming, R. W. (2016). Confessions of a reluctant photorealist. *Journal of Vision*, *16*(12), 3. doi:10.1167/16.12.3

Fleming, R. W., Jäkel, F., & Maloney, L. T. (2011). Visual perception of thick transparent materials. *Psychological science*, *22*(6), 812–20. doi:10.1177/0956797611408734

Foster, D. H. (2003). Does colour constancy exist? *Trends in Cognitive Sciences*, *7*(10), 439–443. doi:10.1016/j.tics.2003.08.002

Foster, D. H. (2011). Color constancy. *Vision Research*, *51*(7), 674–700. doi:10.1016/j.visres.2010.09.006

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research*, *51*(7), 771–81. doi:10.1016/j.visres.2010.09.027

Gescheider, G. A. (1988). Psychophysical Scaling. *Annual Review of Psychology*, *39*(1), 169–200. doi:10.1146/annurev.ps.39.020188.001125

Gescheider, G. A. (1997). *Psychophysics: The Fundamentals* (3rd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Gilchrist, A. L. (1977). Perceived lightness depends on perceived spatial arrangement. *Science*, *195*(4274), 185–7.

Gilchrist, A. L. (1980). When does perceived lightness depend on perceived spatial arrangement? *Perception & psychophysics*, *28*(6), 527–38.

Gilchrist, A. L., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., … Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review*, *106*(4), 795–834. doi:10.1037/0033-295X.106.4.795

Goris, R. L. T., Putzeys, T., Wagemans, J., & Wichmann, F. a. (2013). A neural population model for visual pattern detection. *Psychological review*, *120*(3), 472–96. doi:10.1037/a0033136

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Heasly, B. S., Cottaris, N. P., Lichtman, D. P., Xiao, B., & Brainard, D. H. (2014). RenderToolbox3: MATLAB tools that facilitate physically based stimulus rendering for vision research. *Journal of Vision*, *14*(2), 6–6. doi:10.1167/14.2.6

Hillis, J. M., & Brainard, D. H. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *Journal of the Optical Society of America A*, *22*(10), 2090. doi:10.1364/JOSAA.22.002090

Hillis, J. M., & Brainard, D. H. (2007a). Distinct Mechanisms Mediate Visual Detection and Identification. *Current Biology*, *17*(19), 1714–1719. doi:10.1016/j.cub.2007.09.012

Hillis, J. M., & Brainard, D. H. (2007b). Do common mechanisms of adaptation mediate color discrimination and appearance? Contrast adaptation. *Journal of the Optical Society of America A*, *24*(8), 2122–2133. doi:10.1364/JOSAA.24.002122

Hoffman, D. M., Girshick, A. R., Akeley, K., & Banks, M. S. (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, *8*(3), 33.1–30. doi:10.1167/8.3.33

Kingdom, F. A. (2016). Fixed versus variable internal noise in contrast transduction: The significance of Whittle's data. *Vision research*, *128*, 1–5. doi:10.1016/j.visres.2016.09.004

Kingdom, F. A., & Prins, N. (2010). *Psychophysics : A practical introduction*. London: Academic Press.

Knill, D. C. (1998). Discrimination of planar surface slant from texture: human and ideal observers compared. *Vision research*, *38*(11), 1683–711. doi:10.1016/S0042-6989(97)00415-X

Knoblauch, K., & Maloney, L. T. (2008). MLDS : Maximum likelihood difference scaling in R. *Journal of Statistical Software*, *25*(2), 1–26. doi:10.18637/jss.v025.i02

Knoblauch, K., & Maloney, L. T. (2012). *Modeling Psychophysical Data in R*. New York: Springer.

Koenderink, J. J. (1999). Virtual Psychophysics. *Perception*, *28*(6), 669–674. doi:10.1068/p2806ed

Koenderink, J. J. (2013). Methodological background: experimental phenomenology. In J. Wagemans (Ed.), *Handbook of perceptual organization* (pp. 41–54). Oxford: Oxford University Press.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. volume i : Additive and polynomial representations*. Mineola, New York: Dover Publications.

Krueger, L. E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, *12*, 251–320.

doi:10.1017/S0140525X0004855X

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of vision*, *5*(5), 478–92. doi:10.1167/5.5.8

Logvinenko, A. D., & Maloney, L. T. (2006). The proximity structure of achromatic surface colors and the impossibility of asymmetric lightness matching. *Perception & Psychophysics*, *68*(1), 76–83. doi:10.3758/BF03193657

Logvinenko, A. D., Petrini, K., & Maloney, L. T. (2008). A scaling analysis of the snake lightness illusion. *Perception & Psychophysics*, *70*(5), 828–840. doi:10.3758/PP.70.5.828

Lu, Z.-L., & Sperling, G. (2012). Black-white asymmetry in visual perception. *Journal of Vision*, *12*(10), 8–8. doi:10.1167/12.10.8

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophyiscs. In R. Atkinson, R. Herrnstein, G. Lindzey, & R. Luce (Eds.), *Stevens' handbook of experimental psychology* (p. 3-74). Oxford, England: John Wiley & Sons.

Maertens, M., & Wichmann, F. A. (2013). When luminance increment thresholds depend on apparent lightness. *Journal of Vision*, *13*(6), 21. doi:10.1167/13.6.21

Maertens, M., Wichmann, F. A., & Shapley, R. (2015). Context affects lightness at the level of surfaces. *Journal of Vision*, *15*(1), 15–15. doi:10.1167/15.1.15

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of vision*, *3*(8), 573–85. doi:10.1167/3.8.5

Marks, L. E., & Algom, D. (1998). Psychophysical scaling. In M. H. Birnbaum (Ed.), *Measurement, judgment and decision making* (p. 81-178). San Diego: Academic Press.

Marks, L. E., & Gescheider, G. A. (2002). Psychophysical scaling. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology. vol. 4: Methodology in experimental psychology* (p. 91-138). New York: John Wiley & Sons.

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), 1909–1913. doi:10.1016/j.cub.2012.08.009

McNicol, D. (1972). *A primer of signal detection theory*. London: George Allen &

Unwin.

Miles, W. R. (1930). Ocular dominance in human adults. *The Journal of General Psychology*, *3*(3), 412–430. doi:10.1080/00221309.1930.9918218

Munsell, A. E. O., Sloan, L. L., & Godlove, I. H. (1933). Neutral Value Scales. I Munsell Neutral Value1 Scale. *Journal of the Optical Society of America*, *23*(11), 394-411.

Obein, G., Knoblauch, K., & Viénot, F. (2004). Difference scaling of gloss: nonlinearity, binocularity, and constancy. *Journal of vision*, *4*(9), 711–20. doi:10.1167/4.9.4

Pauli, H. (1976). Proposed extension of the CIE recommendation on "Uniform color spaces, color difference equations, and metric color terms". *Journal of the Optical Society of America*, *66*(8), 866-867.

Paulun, V. C., Kawabe, T., Nishida, S., & Fleming, R. W. (2015). Seeing liquids from static snapshots. *Vision Research*, 1–12. doi:10.1016/j.visres.2015.01.023

Radonjić, A., & Brainard, D. H. (2016). The nature of instructional effects in color constancy. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(6), 847–865. doi:10.1037/xhp0000184

Radonjic, A., Cottaris, N. P., & Brainard, D. H. (2015a). Color constancy in a naturalistic, goal-directed task. *Journal of Vision*, *15*(13), 3. doi:10.1167/15.13.3

Radonjic, A., Cottaris, N. P., & Brainard, D. H. (2015b). Color constancy supports cross-illumination color selection. *Journal of Vision*, *15*(6), 13. doi:10.1167/15.6.13

Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with r*. New York: Springer.

Rogers, B., & Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception*, *8*(2), 125–34.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2004). Some observations on the effects of slant and texture type on slant-from-texture. *Vision research*, *44*(13), 1511–35. doi:10.1016/j.visres.2004.01.013

Ross, H. E. (1997). On the possible relations between discriminability and apparent magnitude. *British Journal of Mathematical and Statistical Psychology*, *50*(2), 187–203. doi:10.1111/j.2044-8317.1997.tb01140.x

Rudd, M. E., & Zemach, I. K. (2007). Contrast polarity and edge integration in

achromatic color perception. *Journal of the Optical Society of America A*, *24*(8), 2134–56. doi:10.1364/JOSAA.24.002134

Runeson, S. (1977). On the possibility of smart perceptual mechanisms. *Scandinavian Journal of Psychology*, *18*, 172–179.

Saunders, J. a. (2003). The effect of texture relief on perception of slant from texture. *Perception*, *32*(2), 211–233. doi:10.1068/p5012

Saunders, J. a., & Backus, B. T. (2006). Perception of surface slant from oriented textures. *Journal of vision*, *6*(9), 882–97. doi:10.1167/6.9.3

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123. doi:10.1016/j.visres.2016.02.002

Shreiner, D., Woo, M., Neider, J., & Davis, T. (2005). *Opengl programming guide: The official guide to learning opengl, version 2, 5th edition*. Upper Sadle River, NJ: Addison-Wesley.

Singh, M. (2004). Lightness constancy through transparency: internal consistency in layered surface representations. *Vision Research*, *44*(15), 1827–1842. doi:10.1016/j.visres.2004.02.010

Singh, M., & Anderson, B. L. (2002). Toward a perceptual theory of transparency. *Psychological Review*, *109*(3), 492–519. doi:10.1037/0033-295X.109.3.492

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153–81.

Stevens, S. S. (1975). *Psychophysics : introduction to its perceptual, neural, and social prospects*. New York: John Wiley & Sons.

Thurstone, L. L. (1927a). Equally often noticed differences. *Journal of Educational Psychology*, *18*, 289–293.

Thurstone, L. L. (1927b). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Todd, J. T., Christensen, J. C., & Guckes, K. M. (2010). Are discrimination thresholds a valid measure of variance for judgments of slant from texture? *Journal of vision*, *10*(2), 1–18. doi:10.1167/10.3.22

Todd, J. T., Thaler, L., & Dijkstra, T. M. H. (2005). The effects of field of view on the perception of 3D slant from texture. *Vision research*, *45*(12), 1501–17.

doi:10.1016/j.visres.2005.01.003

Todd, J. T., Thaler, L., Dijkstra, T. M. H., Koenderink, J. J., & Kappers, A. M. L. (2007). The effects of viewing angle, camera angle, and sign of surface curvature on the perception of three-dimensional shape from texture. *Journal of vision*, *7*(12), 9.1–16. doi:10.1167/7.12.9

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: John Wiley & Sons.

Treisman, M. (1964a). Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology*, *16*(1), 11–22. doi:10.1080/17470216408416341

Treisman, M. (1964b). What do sensory scales measure? *Quarterly Journal of Experimental Psychology*, *16*(4), 387–391. doi:10.1080/17470216408416400

Umbach, N. (2013). *Dimensionality of the perceptual space of achromatic surface colors*. Verlag Dr. Hut. (Doctoral Dissertation, Eberhard-Karls-Universität Tübingen, presented March, 2014; published May 2014.)

Velisavljević, L., & Elder, J. H. (2006). Texture properties affecting the accuracy of surface attitude judgements. *Vision research*, *46*(14), 2166–91. doi:10.1016/j.visres.2006.01.010

Wallach, H. (1948). Brightness constancy and the nature of achromatic colors. *Journal of Experimental Psychology*, *38*(3), 310–324. doi:10.1037/h0053804

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of vision*, *5*(10), 834–62. doi:10.1167/5.10.7

Whittle, P. (1994). The psychophysics of contrast brightness. In A. L. Gilchrist (Ed.), *Lightness, brightness, and trasnparency* (p. 35-110). New York: Psychology Press.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*(8), 1314–1329. doi:10.3758/BF03194545

Wichmann, F. A., Janssen, D. H. J., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging*, *2017*(14), 36–45. doi:10.2352/ISSN.2470-1173.2017.14.HVEI-113

Wiebel, C. B., Aguilar, G., & Maertens, M. (2017). Maximum likelihood difference

scales represent perceptual magnitudes and predict appearance matches. *Journal of Vision*, *17*(4), 1. doi:10.1167/17.4.1

Wiebel, C. B., Singh, M., & Maertens, M. (2016). Testing the role of Michelson contrast for the perception of surface lightness. *Journal of Vision*, *16*(11), 17. doi:10.1167/16.11.17

Wood, S. (2006). *Generalized additive models : an introduction with r*. Boca Raton, FL: Chapman & Hall/CRC.

Zeileis, A., Koenker, R., & Doebler, P. (2015). glmx: Generalized linear models extended [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=glmx` (R package version 0.1-1)

Zeiner, K., & Maertens, M. (2014). Linking luminance and lightness by global contrast normalization. *Journal of Vision*, *14*(7), 3–3. doi:10.1167/14.7.3