Technische Universität Berlin
Fakultät IV - Elektrotechnik und Informatik
Dept. Computational Psychology

Abschlussarbeit

Optimizing data acquisition for Maximum Likelihood Conjoint Measurement

---

vorgelegt von
Jan Simon Zabel
Matrikelnummer: 410791
zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)
im Studiengang Informatik

Erstgutachter: Dr. Guillermo Aguilar
Zweitgutachter: Prof. Dr. Kenneth Knoblauch

16.12.2023

AFFIDAVIT

I hereby declare that the thesis submitted is my own, unaided work, completed without any unpermitted external help. Only the sources and resources listed were used.

Berlin, December 16, 2023

Jan Simon Zabel

ABSTRACT

Understanding brightness perception relies on perceptual scales, which is a function mapping the physical unit of luminance to perceived brightness. Maximum Likelihood Conjoined Measurement (MLCM) is a difference scaling method that can estimate these perceptual scales by having a participant compare two stimuli and recording the answers. This thesis introduces an optimization approach of the data acquisition for MLCM, adopting two sampling strategies. A static and a dynamic sampling strategy. Using a simulated observer, data in the shape of an actual experiment was generated to estimate perceptual scales. This was done with both sampling strategies and multiple noise levels and then evaluated against the ground truth functions, which represent the perceptual encoding functions in the simulation. The static sampling strategies omits trials with high agreement. The dynamic sampling strategy reduces the amount of initial trial shown and then only repeat those with low agreement. Both approaches are able to reduce the amount of trials and consequently the experimental duration by $45 - 49\%$ without impacting the accuracy or precision of the perceptual scales noticeably. The improved data collection process increases the efficiency of the experiment while upholding the integrity of the result.

Keywords: mlcm, perceptual scales, scaling methods, brightness

ZUSAMMENFASSUNG

Sogenannte Wahrnehmungsskalen helfen dem Verständnis von Helligkeitswahrnehmung und dienen als Funktionen, um von der physischen Einheit Luminanz auf die wahrgenommene Helligkeit abzubilden. Maximum Likelihood Conjoined Measurement (MLCM) ist eine Differenzskalierungsmethode, welche diese Wahrnehmungsskalen schätzen kann, indem einem Beobachter zwei Stimuli zum Vergleich gezeigt werden und die Antworten hinsichtlich der Intensität aufzeichnet werden, um die Datensammlung für MLCM effizienter zu gestalten. Dies wird mit zwei Musterziehungsstrategien erzielt. Eine statische und eine dynamische Strategie. Mithilfe eines simulierten Beobachters werden Daten in gleicher Form zu der aus einem echten Experiment generiert. Die geschätzen Wahrnehmungsskalen werden mit eingestellten Wahrnehmungsfunktionen verglichen, welche die Wahrnehmung eines Menschen simulieren sollen. Dies geschieht für mehrere Rauschwerte. Die statische Musterziehungsstrategie zeigt die Vergleiche mit hoher Übereinstimmung nicht. Die dynamische Musterziehungsstrategie verringern die Anzahl der Vergleiche, welche

anfangs durchgeführt werden und wiederholt dann nur die Vergleiche, welche eine geringe Übereinstimmung haben. Beide Ansätze schaffen es, die Anzahl der Vergleiche und dadurch die Zeit für das Experiment um $45 - 49\%$ zu senken, ohne einen Einfluss auf die Genauigkeit oder Präzision zu haben. Der verbesserte Datensammelprozess erhöht die Effizienz des Experiments und behält dabei die Integrität der Ergebnisse bei.

CONTENTS

1 Introduction 1
1.1 Introduction into brightness perception 1
1.2 Measuring brightness perception 1
1.3 Related work 2
1.4 Experimental design for MLCM 4
1.5 Timecost of the experimental procedure 4
1.6 Optimizing the experiment 6
2 Methodology 9
2.1 Simulation 9
2.2 How can we reduce the amount of trials 12
2.2.1 The static sampling strategy 12
2.2.2 The dynamic sampling strategy 13
2.3 Varying parameters 15
2.4 Evaluation 16
3 Results 17
3.0.1 Different noise levels 19
3.0.2 Parameters of the dynamic sampling strategy . 22
3.0.3 Alternative ground truth functions 24
4 Discussion 27
4.0.1 Estimating the perceptual scales 27
4.0.2 Interpretation 27
4.0.3 Limitations of the sampling strategies 28
4.0.4 Sampling the stimulus domain 29
4.0.5 Other sampling strategies 29
4.0.6 Recommendation 30
4.0.7 Open questions 30
4.1 Conclusion 31
References 33
A Appendix 35
A.1 Other noise levels 35

1 INTRODUCTION

1.1 INTRODUCTION INTO BRIGHTNESS PERCEPTION

Brightness perception is a part of visual sciences that refers to the measuring, explaining and quantifying how we perceives brightness. Brightness perception can be influenced by the *luminance* of an object, its reflectivity and its surrounding context. These aspects are also referred to as physical dimensions. While *luminance* is a physical unit that can be measured, perceived brightness is a subjective experience that does not always directly correlate with the *luminance* of an object and is more difficult to measure. This difference between *luminance* and perceived brightness can be seen in White's Illusion (1979) (see Figure 1.1), where two objects with identical *luminance* may be perceived differently from each other. White's Illusion is a bar-like structure, alternating between white and black bars. On those bars two gray targets are superimposed. The targets have the same width as the bar in the background and a fraction of the height, such that a part of the bar in the background is both visible above and below the target. The targets can have multiple *luminance levels* between perceived black (zero) and perceived white (one) and can be either on a white bar or on a black bar, this is referred to as a *context*. Both targets can be on bars of the same *luminance* or a different *luminance*, but not on the same bar. It is not fully understood how White's Illusion can influence our perception of brightness. This is why related work such as Vincent, Maertens, and Aguilar (in preparation) have researched the relation between stimulus variations and perceptual magnitudes and is where this thesis aims to build upon by improving the data collection method.

1.2 MEASURING BRIGHTNESS PERCEPTION

The perceived brightness can differ between objects with identical *luminances* and also between people who each perceive brightness different from each other. In order to understand this difference, the effects of variations in *luminance* on perceived brightness have to be measured. First, we need to find a way how we can express the mapping from *luminance* to perceived brightness. Perceptual encoding functions are one possible way to do that. They give us a base on which we can build our experiment on to measure the perceived brightness of an object.

Figure 1.1: Example of White's Illusion. The left stimuli appears more intense than the right one, even though both stimuli have an identical *luminance*. Both stimuli have a different *context*, the left one "on black" and the right one "on white".

Scaling Methods allow us to measure these perceptual encoding functions by estimating them using statistics. Scaling methods do not produce the functions, but estimate perceptual scales, which resemble the encoding functions through multiple data-points. One scaling method, Maximum Likelihood Conjoint Measurement (MLCM), has been used to estimate perceptual scales that capture how the physical dimensions *luminance* and *context* contribute to the response (Knoblauch and Maloney, 2012, chapter 8), this is explained in more detail in Section 1.3.

These perceptual scales can be measured because *luminance* and the *context* exhibit a form of measurable regularity. That means that *luminance* and the *context* are relevant to the visual system, affecting the perceptual encoding function, and can be expressed empirically using a transfer function (Georgeson, 2014): Given a varying stimulus S and measuring the resulting response R, the relationship between the physical dimensions of *luminance* and *context* of the stimulus to the response can be modeled in a way $\Psi(S) = R$.

An example for a possible perceptual encoding function and the estimated perceptual scale can be seen in Figure 1.2A and B respectively. Figure 1.2A shows a potential mapping of *luminance* for both *contexts* into a perceived brightness for White's Illusion. Figure 1.2B shows the perceptual scales estimated by MLCM which resemble the underlying perceptual encoding function. Figure 1.2C is a perceptual scale estimated using MLCM from a participant in an experiment conducted by Vincent et al..

1.3 RELATED WORK

A study about measuring brightness perception was conducted by Vincent et al. to research the effect of stimulus variations on perceptual

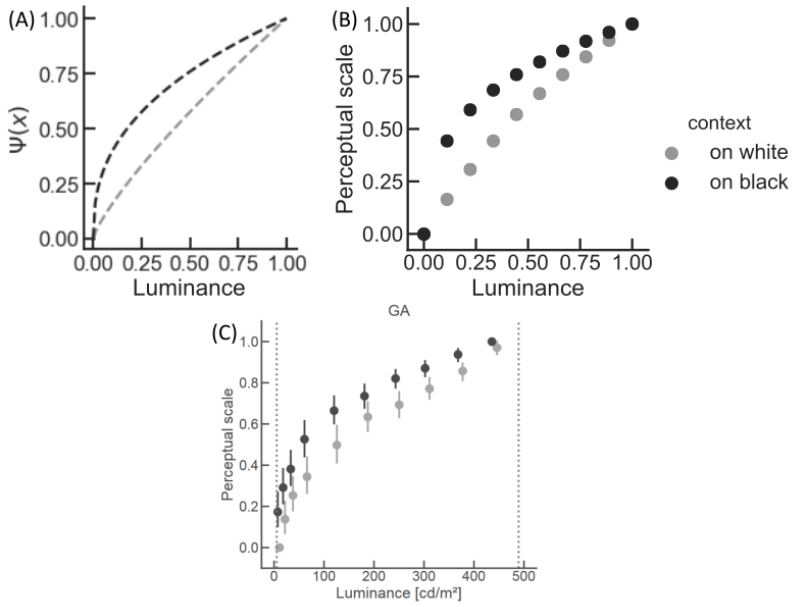


Figure 1.2: Perceptual encoding functions and perceptual scales generated by MLCM. (A) Perceptual encoding functions for White's Illusion with two *contexts* "on white" and "on black". The x-axis presents the *luminance*, a measurable physical unit of light emitted by an object. The y-axis presents perceived brightness $\Psi(X)$ as the result of the perceptual encoding functions for each *context*. This perceptual encoding function is not observable, so this is only an estimate. (B) estimated perceptual scales using MLCM for ten *luminance levels* for each *context*. The x-axis is again the *luminance*, the y-axis is the measured perceived brightness from multiple trials. This can also be seen in Figure 1.1 where both targets have the same *luminance*, but the target "on black" has a higher intensity than the target "on white". This perceptual scale resembles the perceptual encoding function which can be seen in Figure 1.2A. (C) Perceptual scale for White's stimulus obtained with MLCM from an observer "GA" in Vincent et al. (in preparation). The *luminance* is not normalized to one

magnitudes. In their study, perceptual scales for White's Illusion were created using MLCM to research the perceptual encoding process using a scaling method. The result is an estimate of the perceptual encoding function, which cannot be directly measured. Their model of measuring brightness perception uses a variety of Stimuli-Values S to trigger different experiences $\Psi(S)$, which then result in different responses R from observers. This relationship is modeled using functions $\Psi = f_1(S)$, $R = f_2(\Psi)$ and $R = f_3(S)$. In order to measure said response and transform the results into perceptual scales, an experiment as described in 1.4 was conducted to estimate the perceptual scales. The study has shown that MLCM is able to estimate non-linear perceptual scales for multiple participants for White's Illusion. The resulting scales show that the *luminance* for the "on black" target is al-

ways mapped to a higher perceived brightness than that of "on white". The difference in perceived brightness for the same *luminance* values is lowest at the edges (close to zero and one) of the scale and highest in the middle range (close to 0.5 and 0.6). The estimated perceptual scales are in agreement with the knowledge regarding the effect of White's Illusion. This leads to the conclusion that MLCM can provide a link between physical dimensions and perceptual experiences in the form of perceptual encoding.

1.4 EXPERIMENTAL DESIGN FOR MLCM

Perceived brightness is measured by showing a collection of trials consisting of two stimuli, similar to what can be seen in Figure 1.1. In front of the participant is an input-device with two buttons, each corresponding to either the left or the right target in the trial. The participants task is to determine which of the two targets they perceive as more intense and to press the corresponding button. There are 15 repeats for each unique trial. Each round of trials compares every *luminance level* and every *context* with each other, with the exception of the same *luminance* and *context* for both stimuli, as well as ignoring the order of stimuli. This is set up as an 2AFC experiment, so the participants must choose either left or right, they cannot choose to not give an answer or anything other than left or right. There is no time limit for the trials, participants can take as long as they want to give an answer.

1.5 TIMECOST OF THE EXPERIMENTAL PROCEDURE

While each trial of comparing two stimuli can be done in just a second to just a few seconds, comparing multiple *luminance levels* and *contexts* will add up to hundreds or even thousands of trials

Let N_L be the amount of different *luminance levels*, N_C the amount of *contexts*::

$$T = N_L \cdot N_C$$

as the amount of possible unique targets. We can use T to calculate the amount of unique trials:

$$\#Unique\ trials = \frac{T \cdot (T-1)}{2}$$

This gives us:

$$\frac{20 \cdot (20-1)}{2} = 190 \text{ unique trials}$$

The timecost increases if we decide to add more *luminance levels* and *contexts* as can be viewed in the table 1.1. The table shows the

exponential growth of unique comparisons (trials) for *luminance levels* and *contexts*, as well as the total amount of trials with 15 repeats of every unique comparison. Increasing the amount of *luminance levels* by 30% will increase the amount of unique trials by more than 60%. Consequently, the timecost increases with the amount of unique trials. Doubling the *luminance levels* and *contexts* will make the experiment unfeasible.

Luminance Levels	Contexts	Unique comparisons
10	2	$190(\cdot 15 = 2850)$
13	2	$325(\cdot 15 = 4875)$
10	3	$435(\cdot 15 = 6525)$
20	4	$3160(\cdot 15 = 47400)$

Table 1.1: Exponential growth of trials when adding more luminance levels or contexts

(A) (B)

Figure 1.3: Two MLCM trials of White's Illusion. (A) An Easy trial with two stimuli. The difference in *luminance* is high. There is probably high agreement on which target a participant will choose. The right target is perceived as brighter. (B) A difficult trial with two stimuli. The difference in *luminance* is low. There is probably low agreement on which target a participant will choose. The left target has a higher luminance.

Not all trials carry the same informative value. There are "easy trials" with predictable results as can be seen in Figure 1.3A. This trial has a high agreement among observers. A high agreement means, that an observer is very likely to pick the same target again during most, if not all trial repeats. The resulting relative frequency of answers is zero or one, referring to 0% or 100% respectively. The opposite can be said for a difficult trial in Figure 1.3B where an observer may give different answers in multiple showings of the unique trial. The resulting relative frequency is between zero and one. All unique trials are made up of two stimuli, and the resulting relative frequencies for all unique trials and for the 15 trial repeats is presented using a heatmap. Figure 1.4 shows the relative frequency from an observer in Vincent et al. (in preparation) as a heatmap. The x- and y-axis show the *contexts* and *luminances* being compared. The cells show the relative frequency of answers. The large amount of zeroes and ones indicates that a lot of

the unique trials are easy and thus predictable. The easy trials are very common among comparisons within the same context and less common in trials with differing contexts. These easy trials still create a time cost and make up more than 50% of all unique trials, so it can lead to an important question: Do all unique trials have to be shown to participants if we can predict some of the result?

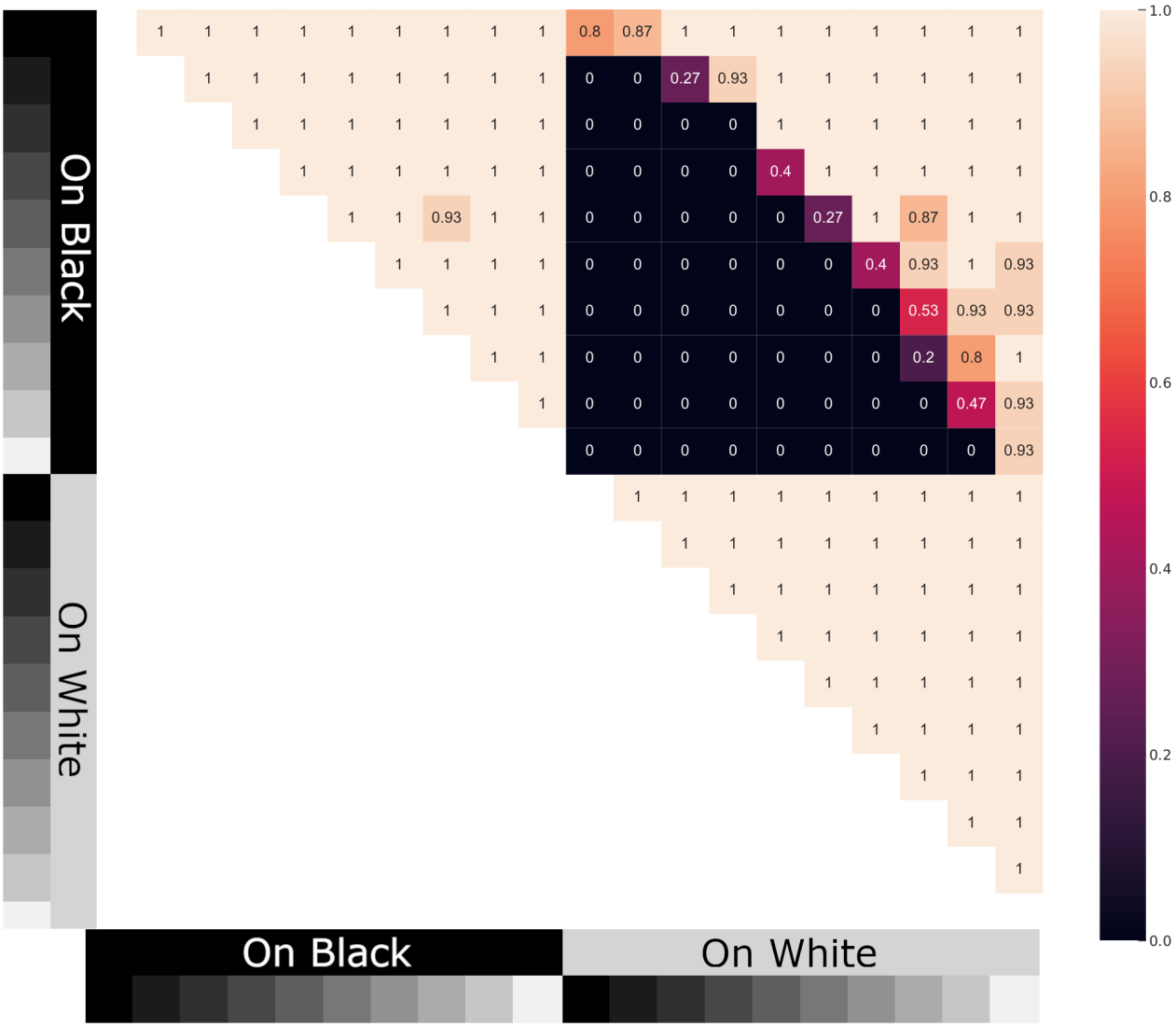


Figure 1.4: Heatmap of relative frequency of White's Illusion trial answers for one participant. Each value of a cell represents the relative frequency of a participant choosing one stimulus over another in 15 trial repeats, 2850 Trials in total. x- and y-axis represent both *contexts* and *luminance levels*. The stimuli on the x- and y-axis are compared against each other.

1.6 OPTIMIZING THE EXPERIMENT

Being able to reduce the number of trials could reduce the overall time it takes to do an experiment. It could also leave time for more of the informative trials, thus increasing efficiency and reduce tiredness in observers. Knowing which trials carry more informational value provides the opportunity to show more informative trials to participants, leading to potentially better results using the same amount of time or similar results using less time. This raises the following research question: Can we reduce the amount of less informative trials and consequently the experiments duration, for a fixed set of unique stimuli, without impacting the quality of the encoding function estimated using MLCM?

In this thesis I explore this question by simulating the experiment using a simulated observer. I have developed two novel sampling strategies strategies that could reduce the amount of trials with high agreement and improve the data collection for the experiment. I evaluated the estimated perceptual scales against the perceptual encoding functions available in a simulation using the Root-Mean-Squared-Error (RMSE).

2 METHODOLOGY

The goal of the thesis is to find strategies to optimize the data acquisition for the experiment. I have developed two strategies that can reduce the number of uninformative trials and prioritize the more informative trials.

The experiment conducted by Vincent et al. (in preparation) provided me with some prior knowledge of which trials are more informative. This knowledge was used as a basis to develop the Static sampling strategy, omitting a part of the less informative trials from the experiment. I also have developed a strategy that improves the data acquisition of the experiment without requiring prior knowledge like the Static sampling strategy does. The Dynamic sampling strategy can prioritize more informative trials and show them more often than less informative trials. I evaluated these sampling strategies through a simulated experiment without actual participants.

2.1 SIMULATION

I set up the simulation-environment for the experiment, creating a simulated observer to act as a replacement for a participant. A visualization for this process can be seen in Figure 2.1.

Step one is setting up the simulation with two ground truth functions $\Psi_w(s)$ and $\Psi_b(s)$ to have the computer simulate brightness perception for both *contexts* using the *luminance* s, a noise level σ to simulate randomness, the ten *luminance levels* and the two *contexts*. The ground truth functions are modified power functions:

$\Psi_b(s) = m(s-y)^\alpha + n$ and $\Psi_w(s) = o(s-z)^\beta + p$

where $\alpha, \beta \geq 0$ are the exponents and $m, n, o, p, y, z \in \mathbb{R}$ can modify and shift the power functions in more ways.

Step two is generating all unique trials containing the 10 *luminance levels* and two *contexts*.

Step three is using the simulated observer, calculating the decision variable for each trial using the ground truth functions and noise. The decision variable is calculated as

$\delta = \Psi_{c_2}(s_2) - \Psi_{c_1}(s_1) + \epsilon$

where c_1 and c_2 refers to the ground truth function based on the two *contexts*. The values s_1 and s_2 denote the *luminance value*. The Gaussian noise $\epsilon \sim N(0, \sigma^2)$ also affects the decision variable, adding randomness. The simulated observer chooses the first stimulus if $\delta < 0$, otherwise it chooses the second one.

Step four is creating data in the same format as actual data from a human participant, such that the data can be fed into the MLCM algorithm.

Step five is MLCM producing the perceptual scales based on the simulated data. The perceptual scales are estimates that approach the perceptual encoding function. The perceptual encoding functions are referred to as "ground truth functions" in the simulations. The developed static and dynamic sampling strategies alter the generated trials by omitting some of them. This happens after step two and before step three before the computer acts as the simulated observer. The simulation then continues until the end as normal for the Static sampling strategy. At the end of step three, the Dynamic sampling strategy will do a small analysis of results and repeat step three with a selection of trials again before continuing as normal until the end of the simulation.

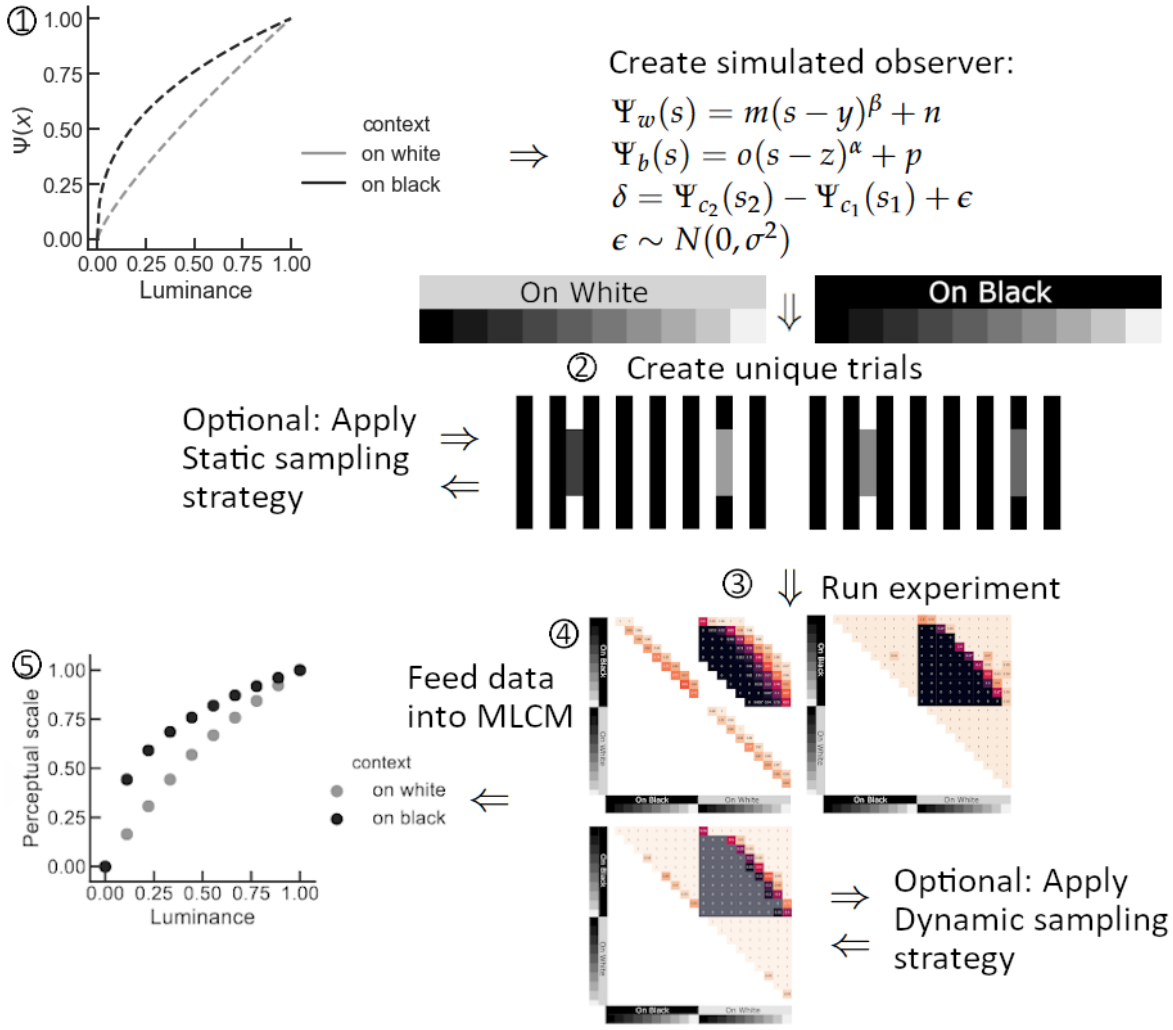


Figure 2.1: Process of the simulation from setup to estimating the perceptual scales

After simulation data is obtained, the estimated perceptual scales are compared to the ground truth functions to see how accurate they are. The perceptual scales can also be compared against the scales for the other sampling strategies. It is expected that the data gathered by the sampling strategies is less than that of the full experiment since less trials will be shown. The simulation for this thesis is an extension of the simulation used in Vincent et al. (in preparation). (The simulation is available at: `https://git.tu-berlin.de/janzabel/white_scaling_bachelor`)

The ground truth function $\Psi(X)$ changes how *luminance* and *context* are translated into perceived brightness. The function, as well as the *luminance levels* are normalised between 0 and 1, 0 being the lowest possible luminance and 1 being the highest possible luminance. The functions are anchored at 0 for the lowest *luminance* and the "on white" position. There is one ground truth function for every of the two *contexts* since every *context* alters how brightness is perceived. The ground truth function for "on white" and "on black" will be different from each other throughout the experiments. The "on black" function will translate a *luminance level* to a higher perceived brightness than the function for "on white" based on the reports in Vincent et al. (in preparation). An example for ground truth functions can be seen in Figure 2.2.

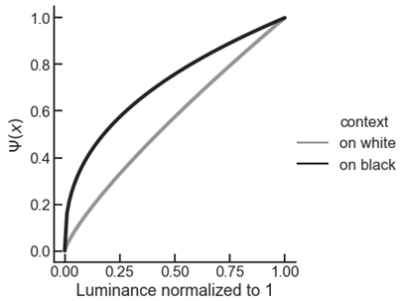


Figure 2.2: Two ground truth functions. One each for "on white" and "on black" setup as arguments for the simulated observer. The ground truth functions can have multiple shapes and serve as perceptual encoding functions in the simulation. The displayed ground truth functions are: $\Psi_{White}(X) = X^{0.4}$ and $\Psi_{Black}(X) = X^{0.8}$

It is also possible to change the amount of noise at the decision stage. The noise simulates the uncertainty of participants, allowing the trials to give more realistic results for stimuli with similarly perceived brightness, as well as introducing a small amount of lapses. The noise will vary between 0 and 0.1 at the decision stage, 0 being a deterministic simulation, meaning that a trial with two stimuli will always give the same result, no matter how often it is shown. Increasing the noise will increase the chance the simulation picks a different answer for the same trial in another round of trials. A realistic estimate for noise

at the decision stage is $\sim 5\%$ and a realistic range $3.21\% \leq \sigma \leq 6.72\%$ as reported by Vincent et al. (in preparation).

2.2 HOW CAN WE REDUCE THE AMOUNT OF TRIALS

One way to improve the efficiency of data acquisition is to use a sampling strategy. Sampling means choosing a selection of trials to show to participants. The goal is to show a reduced total amount of trials for every experiment. There are multiple ways to sample the trials: sampling the trials before showing them and sampling the trials while they are being shown. I have developed two strategies that could reduce the amount of trials.

2.2.1 *The static sampling strategy*

The Static sampling strategy aims to sample before any trials are shown to a participant or simulated. The way this works is by comparing the *luminance levels* and not showing trials with high differences in *luminance*. I expect participants to have a high agreement on big differences in *luminance* for the same *context*, so when both targets are "on white" or both targets are "on black". I use a percentile difference in *luminance* of $> 20\%$ for the same *context*. For different *contexts*, so when one target is "on white" and one target is "on black", a percentile difference of $> 50\%$ is deciding whether a trial is being shown or not. The effects of this Static sampling strategy can be viewed in Figure 2.3 where the heatmap shows less trials than before.

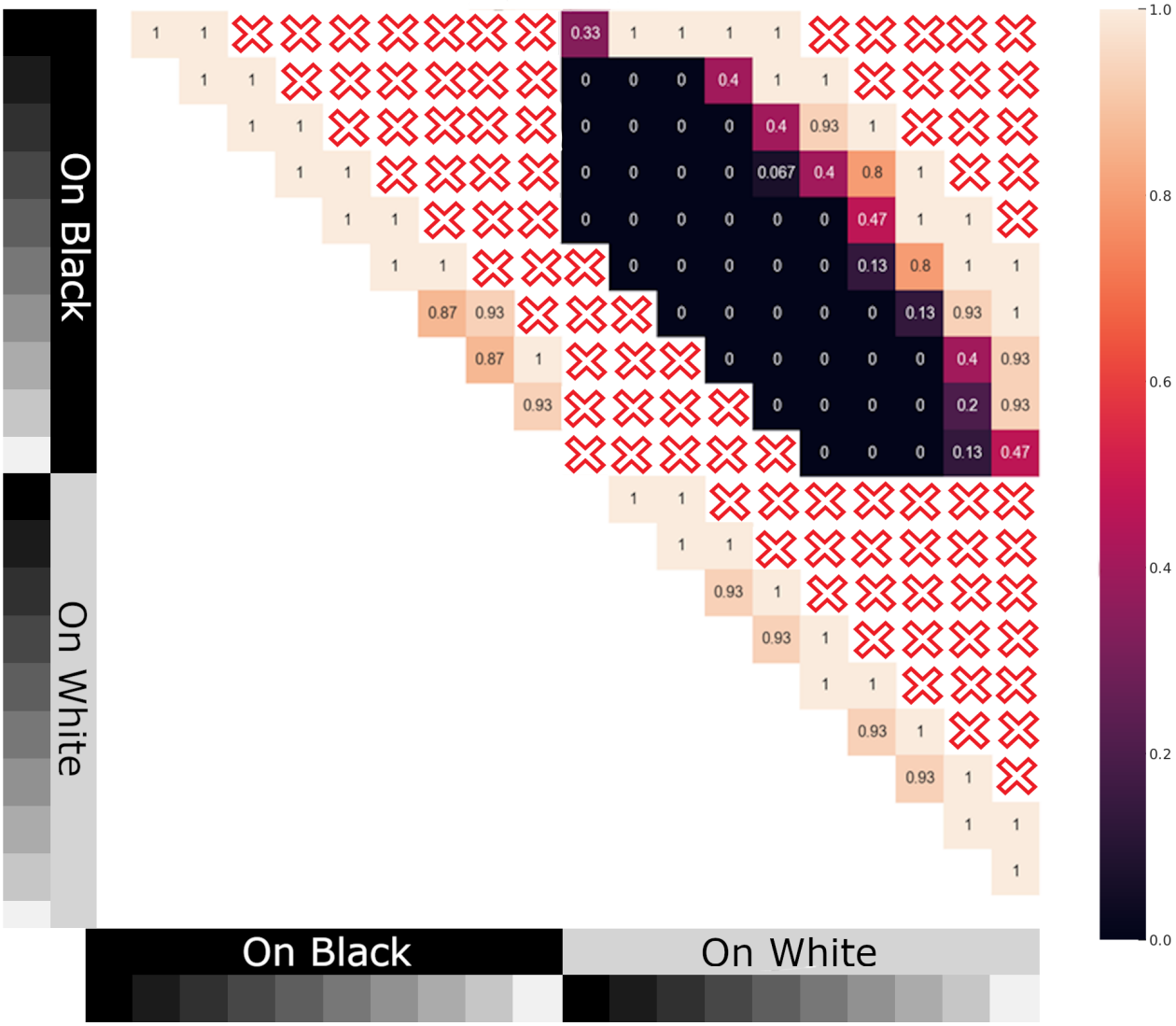
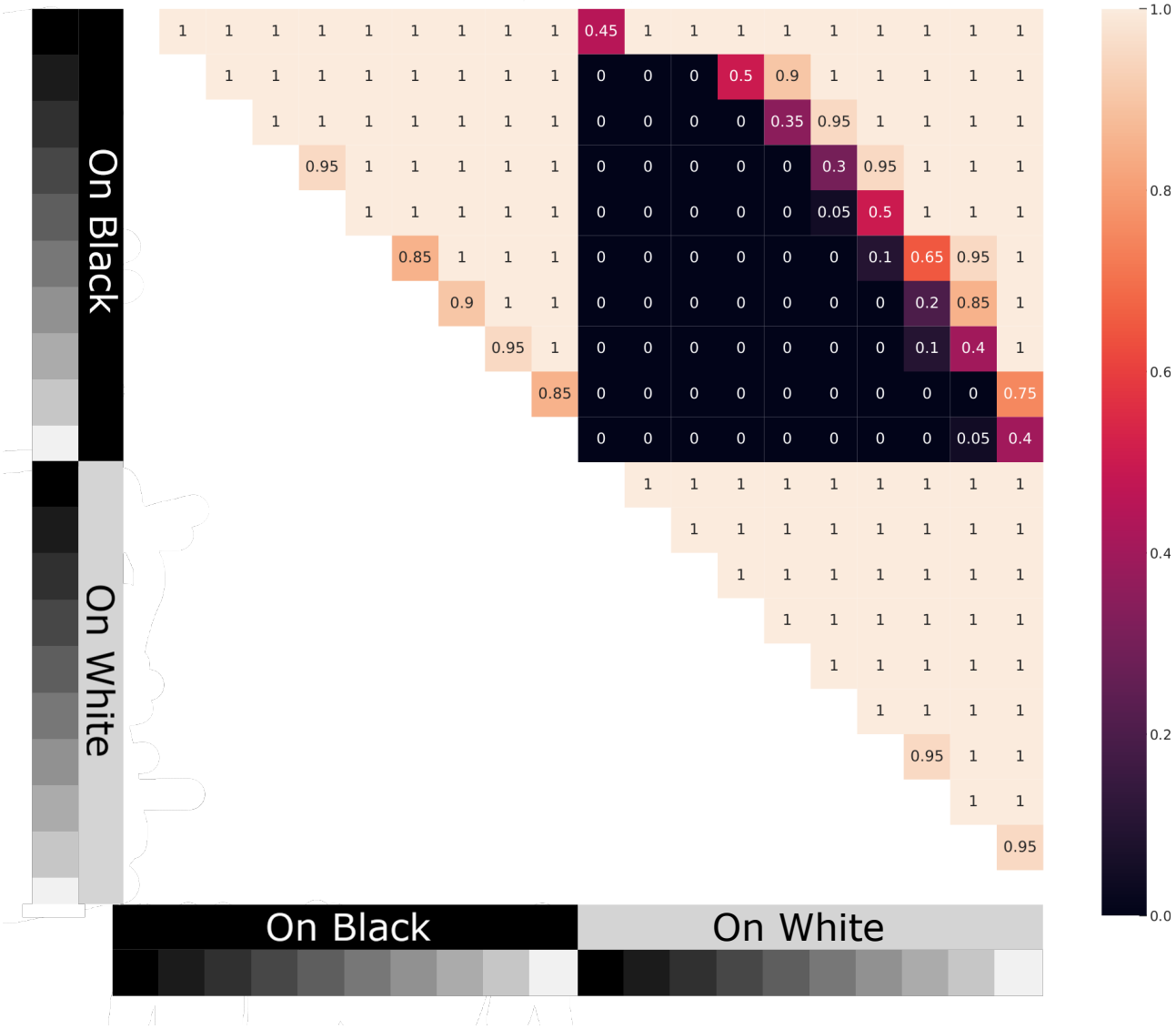


Figure 2.3: Heatmap of relative frequency of an experiment with White's illusion (1979). There are 15 trial repeats for the Static sampling strategy from a simulated observer. Trials with a *luminance* difference of $> 20\%$ for an identical *context* are removed from the collection of shown trials. Trials with a *luminance* difference of $> 50\%$ for a different *context* are also removed from the collection of trials. The removed trials are denoted by a red X.

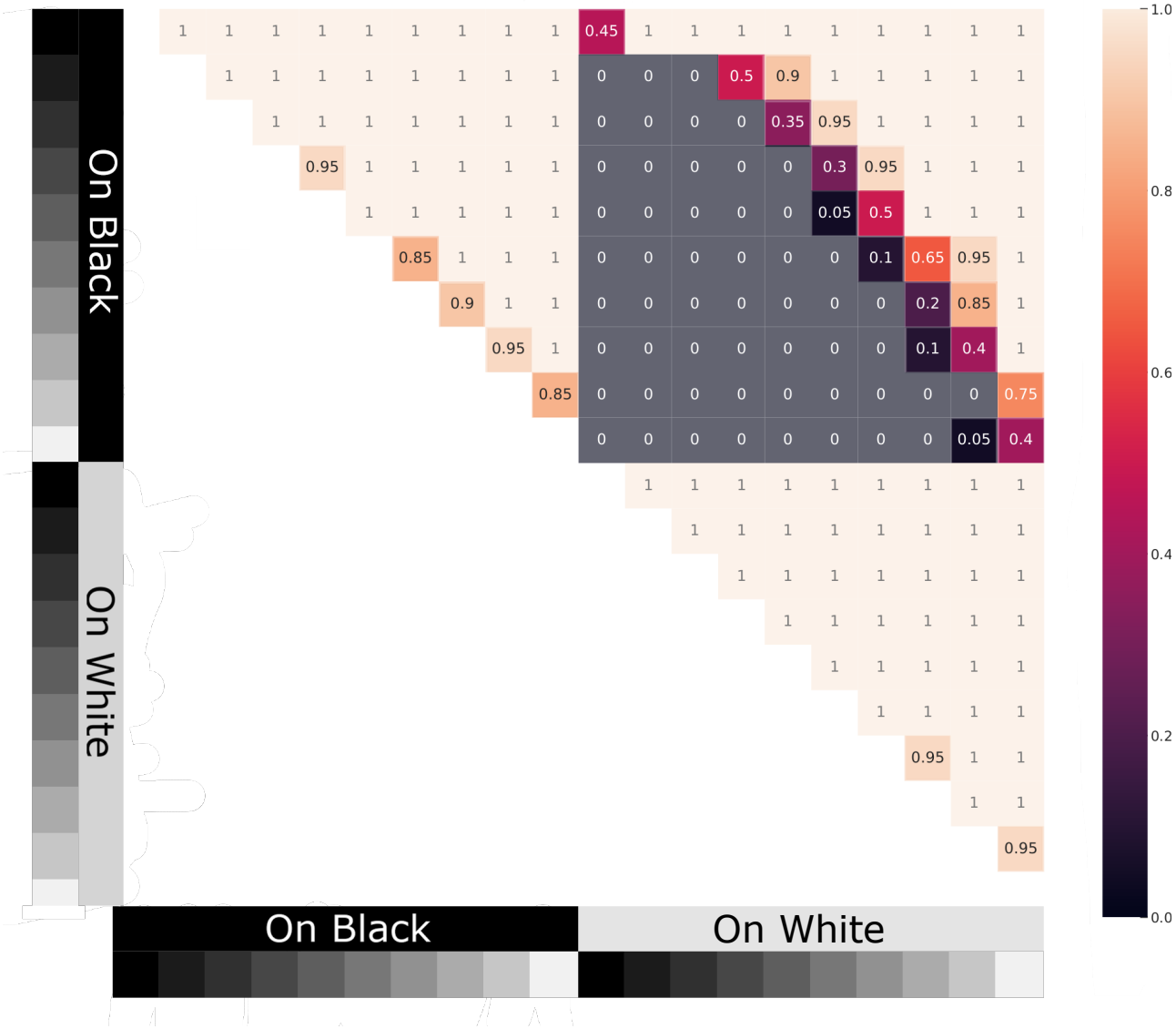
2.2.2 *The dynamic sampling strategy*

The Dynamic sampling strategy aims to sample while the trials are being shown to a participant or simulated. The way this works is by splitting the experiment into two parts, the initial stage and the sample stage. The initial stage is similar to the original experiment in that all trials are being shown, no matter the difference in *luminance*, but reducing the amount of repeats for each trial from 15 times down to

seven times. This gives a first impression and can be used to sample the trials in a way where trials with high agreement results, so 0% or 100%, are not shown again. Trials with low agreement results, so anything $>$ 0% and $<$ 100%, are shown again in the sample stage another 8 times, aiming to get the same accuracy as the original experiment for those trials. This process can be seen in Figures 2.4A and 2.4B.



(A)



(B)

Figure 2.4: Heatmaps of relative frequency showcasing the Dynamic sampling strategy for an experiment with White's Illusion from a simulated observer. (A) Initial run creates results for sample run. (B) Isolate low agreement results (The trials which are not grayed out). Repeat only those trials in sample run.

2.3 VARYING PARAMETERS

The ground truth functions shown in Figure 2.2 are just two of multiple ground truth functions that can be used in the simulation. Changing ground truth functions can be used to test for which range of power functions the sampling strategies work and at which point their accuracy and precision is reduced. Another example for a different ground truth function is shown in Figure 2.5. Other ground truth can also show if the data collected by the sampling strategies is enough to esti-

mate functions of different shapes, such as a cubic function. The same can be said for the noise which will vary between $0\% - 10\%$ (0.0-0.1) at the decision stage to test the limits of the sampling strategies.

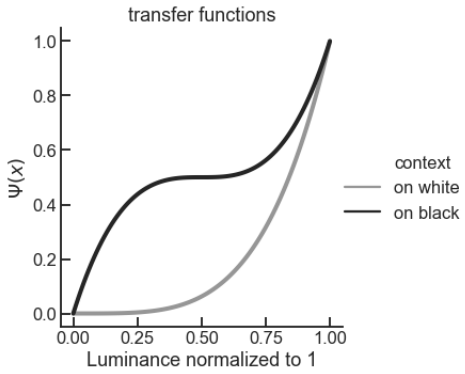


Figure 2.5: Alternative Ground Truth Functions.
$\Psi_{Black}(X) = 4 \cdot (X - \frac{1}{2})^3 + 0.5$ and $\Psi_{White}(X) = X^4$. These can be used to test if both sampling strategies are still able to estimate perceptual scales for these types of ground truth functions.

2.4 EVALUATION

As the experiment is simulated, we have full knowledge of the ground truth functions which would not be available in an actual experiment. This allows for a direct comparison between the estimated perceptual scales $\hat{\Psi}_w(s), \hat{\Psi}_b(s)$ and the ground truth functions $\Psi_w(s)$, $\Psi_b(s)$ for each of the N=10 *luminance values* s used in the simulation. For this comparison, the Root-Mean-Squared-Error (RMSE) is used to calculate the accuracy of the perceptual scales.

The RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((\Psi_w(s_i) - \hat{\Psi}_w(s_i))^2)} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}((\Psi_b(s_i) - \hat{\Psi}_b(s_i))^2)}$$

The simulation was run 1000 times and the average RMSE was calculated for accuracy. This is how much the perceptual scales are different from the ground truth functions. A small RMSE corresponds to high accuracy. Then a percentile confidence interval of 95% is calculated to get the precision of the estimated perceptual scales. The precision is a range where 95% of the estimated perceptual scales will be. A high precision corresponds to a small confidence interval.

3

RESULTS

Figure 3.1 shows an average of the perceptual scales for all sampling strategies for a realistic average noise level at the decision stage of 0.05. A comparison between the three sampling strategies depicted in the perceptual scales shows that the sampling strategies do not seem to impact the estimated scales in a way which would alter the accuracy or precision. The amount of trials is reduced by $\sim 45-50\%$. Both sampling strategies collect enough data to successfully estimate accurate and precise perceptual scales with MLCM. This gets even clearer when looking at the RMSE in Figure 3.2 for each of the sampling strategies as both the static and Dynamic sampling strategy have a less than 0.5% decrease in accuracy when compared to using no (full) sampling strategy.

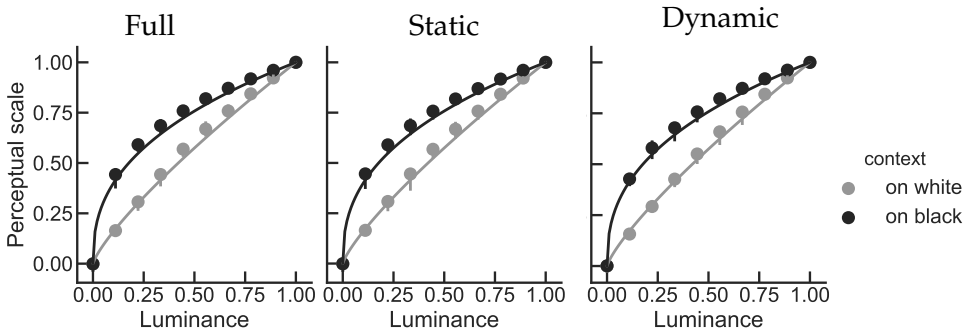


Figure 3.1: Perceptual scales estimated by MLCM. Corresponding ground truth functions are $\Psi_{White}(X) = X^{0.4}$ and $\Psi_{Black}(X) = X^{0.8}$. The noise level at the decision stage is 0.05. The labels Full, Static and Dynamic refer to the sampling strategies.

Sampling Strategies	**RMSE**
Full	0.0696
Static	0.0701
Dynamic	0.0737

Table 3.1: Table of RMSE for the sampling strategies with a noise of 0.05. A lower RMSE corresponds to higher accuracy.

Figures 3.3A-3.3E show the perceptual scales estimated by MLCM for the functions $\Psi_{White}(X) = X^{0.4}$ and $\Psi_{Black}(X) = X^{0.8}$ and varying levels of noise between 0.0 and 0.1 at the decision stage in the simulation. Each row of scales represent a simulation with one level of noise. This is a selection of noise levels I deemed relevant. More scales of

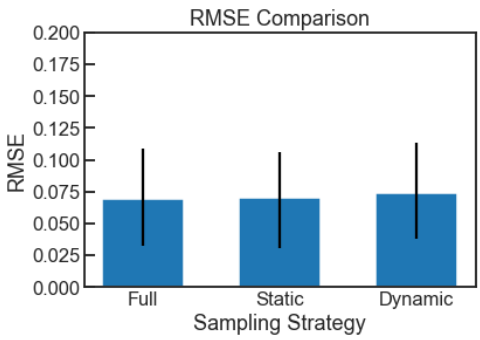


Figure 3.2: RMSE averaged over 1000 simulations from a simulated observer for all sampling strategies. Lower RMSE means a closer approximation of the ground truth function. The RMSE is calculated against the ground truth functions. The exact numbers can be viewed in table 3.1.

simulations with other noise levels (0.01, 0.03, 0.07, 0.08 and 0.09) can be found in the Appendix.

Figure 3.3A depicts the scales of a simulation without noise, making it deterministic. The results are inaccurate. MLCM is not able to approach the underlying ground truth functions. Figure 3.3B shows an unrealistically low level of noise and the scales are inaccurate. MLCM is unable to approach the ground truth functions using perceptual scales accurately, regardless of the sampling strategy. Comparing Figure 3.3A, B and C for any of the sampling strategies visualizes the trend that the accuracy of the scales improves with a rising noise level. This is until the noise in the simulation gets unrealistically high, as can be seen in Figure 3.3D. The scales have a high accuracy and represent the ground truth functions from a noise level 0.04 as shown in Figure 3.3C, which is well within the noise range recorded in Vincent et al. (in preparation) for human participants.

Using the Full sampling strategy and using the Static sampling strategy results in almost identical scales across all noise levels. The increase in efficiency when using the Static sampling strategy is 45.2%, reducing the amount of trials from 2850 down to 1560 per experiment.

The Dynamic sampling strategy has a slightly lower accuracy and precision. The accuracy refers to the position of the dots. The precision refers to the size of the confidence intervals. The increase in efficiency is divided into two parts. An increase of 53.3% with the initial stage, but this is lowered because of the sample stage which has a different amount of trials every time. The average is around 200. The overall increase in efficiency is around 46.3%. Both sampling strategies were able to estimate the scales for realistic noise level estimates.

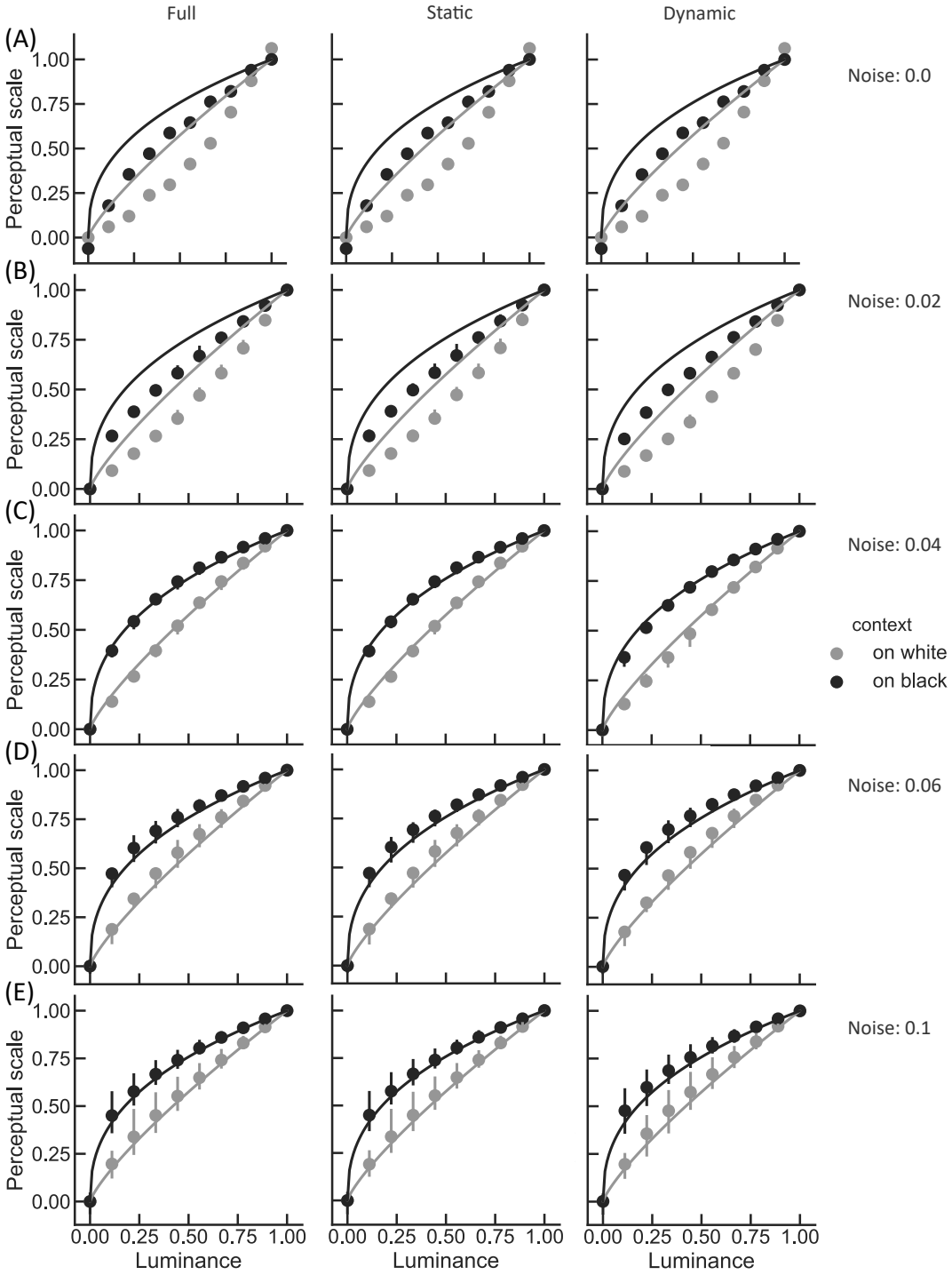


Figure 3.3: Perceptual Scales estimated by MLCM using a simulated observer with all sampling strategies for the noise levels (A) 0.0 (B) 0.02 (C) 0.04 (D) 0.06 (E) 0.1

3.0.1 *Different noise levels*

Every visual system is unique and Vincent et al. (in preparation) has shown that participants have varying levels of noise when comparing

the intensity of stimuli. Because there is no fixed noise for everyone, it is relevant to test the sampling strategies for multiple levels of noise. The Figures 3.4A-C depict the changes of RMSE from the estimated perceptual scales for multiple levels of noise between 0.0 and 0.1.

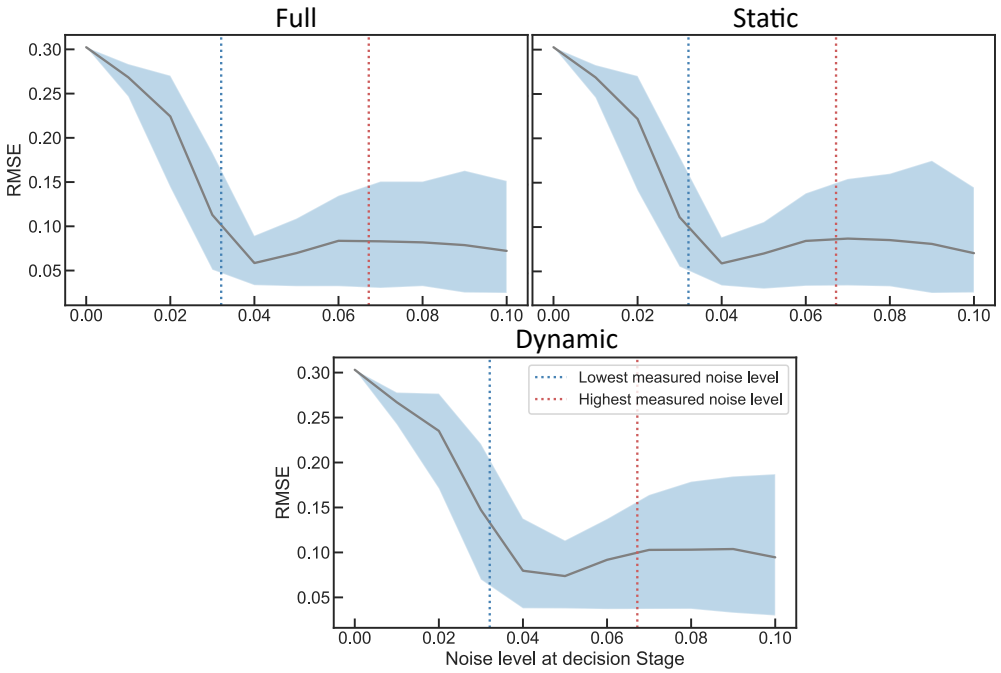
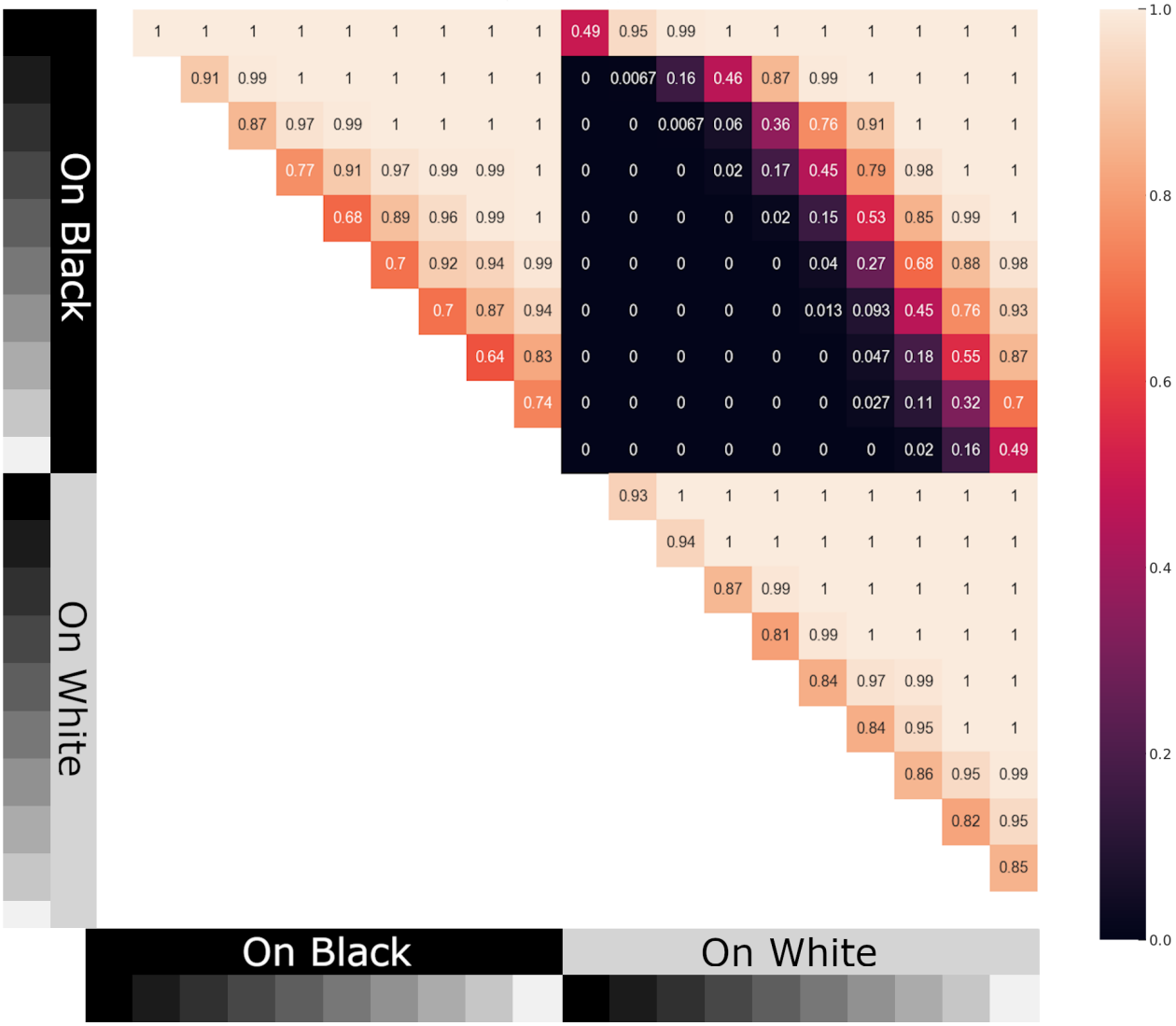
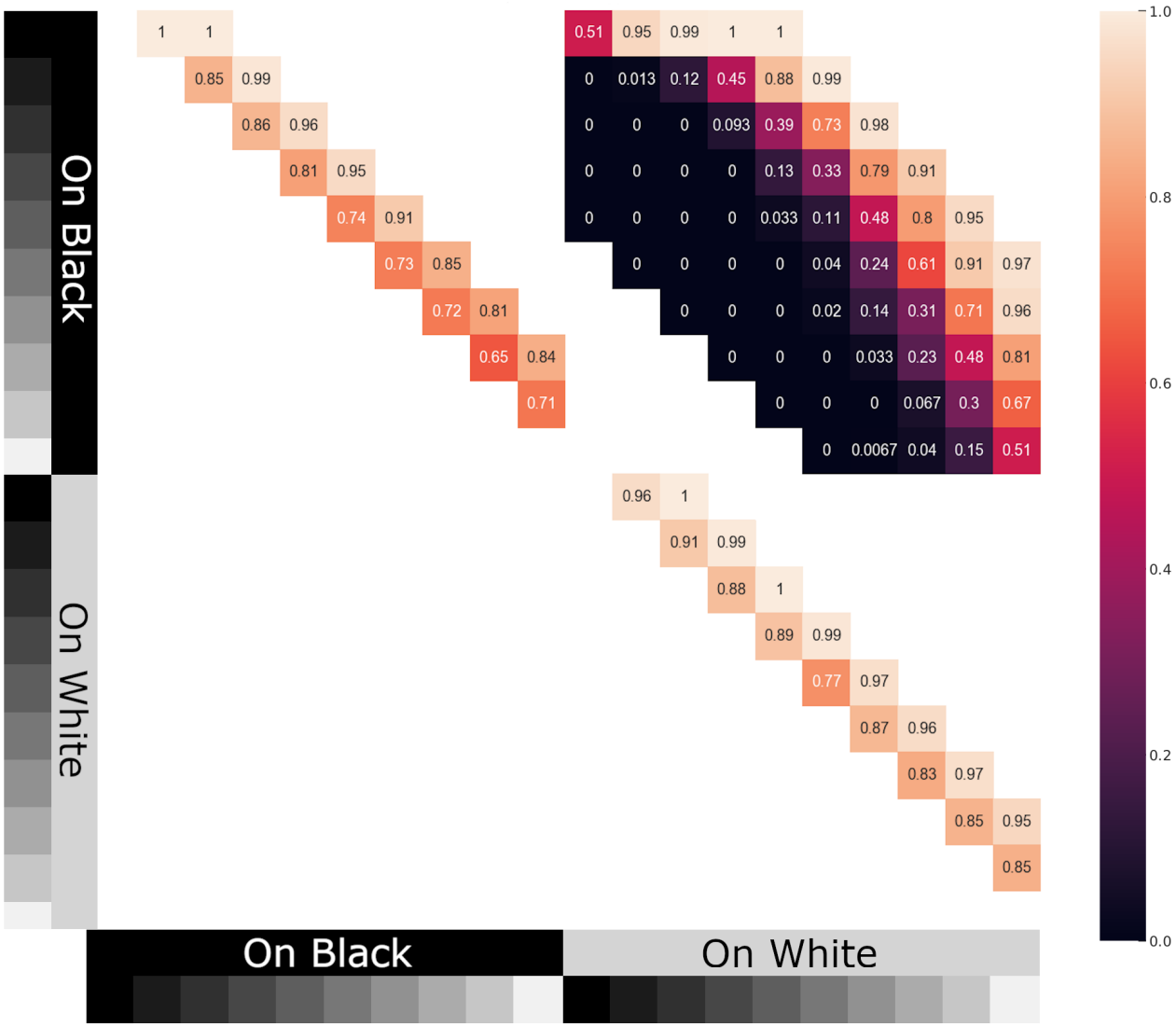


Figure 3.4: RMSE Plots for noise levels between 0.0 and 0.1 in increments of 0.01. Averaged over 1000 simulations from the simulated observer. The RMSE is the mean deviation for all ten *luminance levels* and two *contexts*. The x-axis shows the noise level at the decision stage. The y-axis shows the RMSE. The blue area is the 95% confidence interval. The blue and red markings show the lowest and highest noise levels recorded by Vincent et al. (in preparation).

The overall deviation of the perceptual scales from the ground truth functions is less than 10% for a realistic noise level range for all sampling strategies and lowest at around $6 - 8\%$ for the noise range at the decision stage of 0.04-0.06. The accuracy is good from a noise of 0.04 and upwards. The precision decreases with rising noise levels. The RMSE for the full sampling strategy and the Static sampling strategy is almost identical, deviating only at very high noise levels when the amount of outliers increases to the point of affecting the estimated scales. These outliers are removed from the collection of trials by the Static sampling strategy. An example of this can be seen in Figures 3.5A and B. The Dynamic sampling strategy has the lowest accuracy and precision when compared to the other sampling strategies, but both accuracy and precision are within 10% difference when compared to the other sampling strategies for a realistic noise range.



(A)



(B)

Figure 3.5: Relative Frequency of an experiment with White's Illusion (1979) from a simulated observer. A drawback of the Static sampling strategy occurs for very high noise levels. Some relevant trials at the top left area are removed from the collection of trials that is simulated. This simulation was run with an unrealistically high level of noise and exists only to present a problem which could occur. (A) 150 trials and a high noise of 0.1, full sampling strategy (B) 150 trials and a high noise of 0.1, Static sampling strategy

3.0.2 *Parameters of the dynamic sampling strategy*

The accuracy and precision of the Dynamic sampling strategy can be improved when changing the amount of trials in the initial stage and in the sample stage, at the cost of being less efficient. Figure 3.6 shows the RMSE for varying arguments of the Dynamic sampling strategy. I

vary the amount of trials in the stages, always reaching 15 repeats in total for trials with low agreement results. If the initial stage has seven trial repeats then the sample stage has 15-7=8 trial repeats. The more trial repeats there are in the initial stage, the less efficient but more accurate and precise the estimated perceptual scales are. The far right entry at 15 Initial repeats and zero sample trial repeats represents the full experiment without any sampling strategies applied. The accuracy is high between seven and 15 Initial trial repeats. For all simulations, seven initial trial repeats and eight sample trial repeats were used.

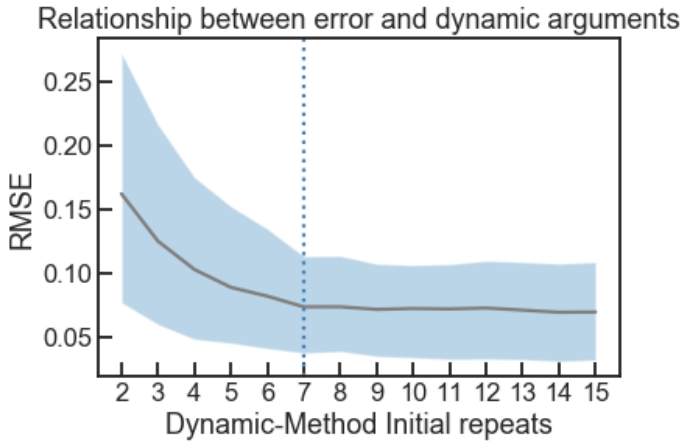


Figure 3.6: RMSE for multiple arguments of the Dynamic sampling strategy. Experiment with White's Illusion from the simulated observer, averaged over 1000 simulations for a noise of 0.05 at the decision stage. The x-axis shows the amount of trials in the initial stage and indirectly the amount of trials in the sample stage. The y-axis is the RMSE. The ground truth functions are those used before in Figures 3.3A-3.3E. The light-blue zone is the 95% confidence interval of RMSEs.

3.0.3 *Alternative ground truth functions*

As said in Section 2.3, there are multiple ground truth functions to use in the simulation. Figure 3.7A-3.7C presents multiple sets of two alternative ground truth functions:

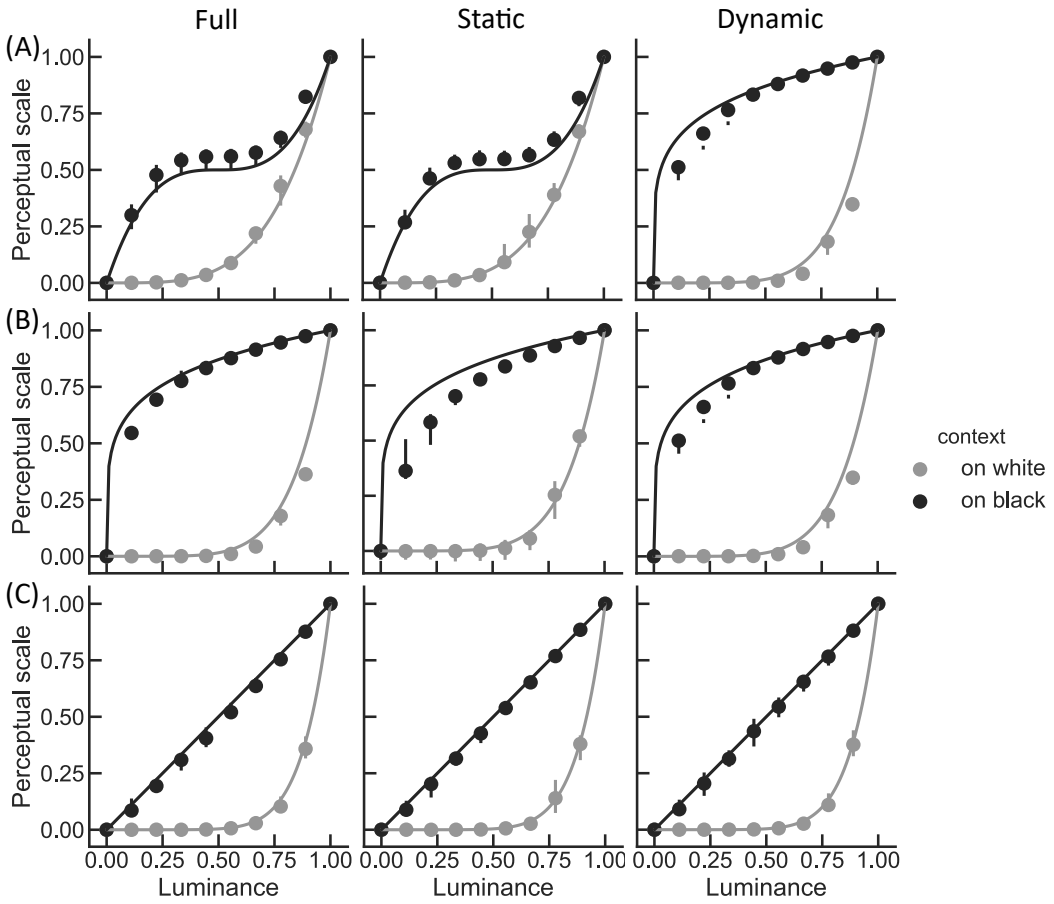


Figure 3.7: Alternative ground truth functions for the simulated observer in the simulations. The noise at the decision stage is 0.05. The sampling strategies are full, static and dynamic from left to right. (A) Set of alternative ground truth functions for the simulated observer. $\Psi_{White}(X) = X^4$ and $\Psi_{Black}(X) = 4 \cdot (X - \frac{1}{2})^3 + 0.5$ (B) $\Psi_{White}(X) = X^6$ and $\Psi_{Black}(X) = X^{0.2}$ (C) $\Psi_{White}(X) = X^8$ and $\Psi_{Black}(X) = X$

Not all sampling strategies can estimate all functions equally well and the limits for the given arguments of each sampling strategies are reached. For the functions presented in Figure 3.7A, both the static and the Dynamic sampling strategy are able to estimate both functions, but with lower precision than using the full sampling strategy. The decline in precision for the Static sampling strategy is noteworthy as this shows that extreme functions like $f(X) = X^4$ appear to be the limit for the arguments used. The accuracy for the Static sampling strategy is higher than when using the full sampling strategy.

An even more extreme example can be seen in Figure 3.7B, where the Static sampling strategy is not able to generate enough data for MLCM to estimate accurate perceptual scales. Both scales for $f(x) = X^6$ and $f(X) = X^{0.2}$ have significantly worse accuracy and precision at the edges where the Static sampling strategy reduces the amount of trials. The Dynamic sampling strategy does better here and is able to both more accurately and more precisely estimate both functions. The last set of functions also contains a linear function X where both sampling strategies are able to estimate the function with high accuracy and precision.

The RMSE plots for all sets of functions can be seen in Figure 3.8A-C

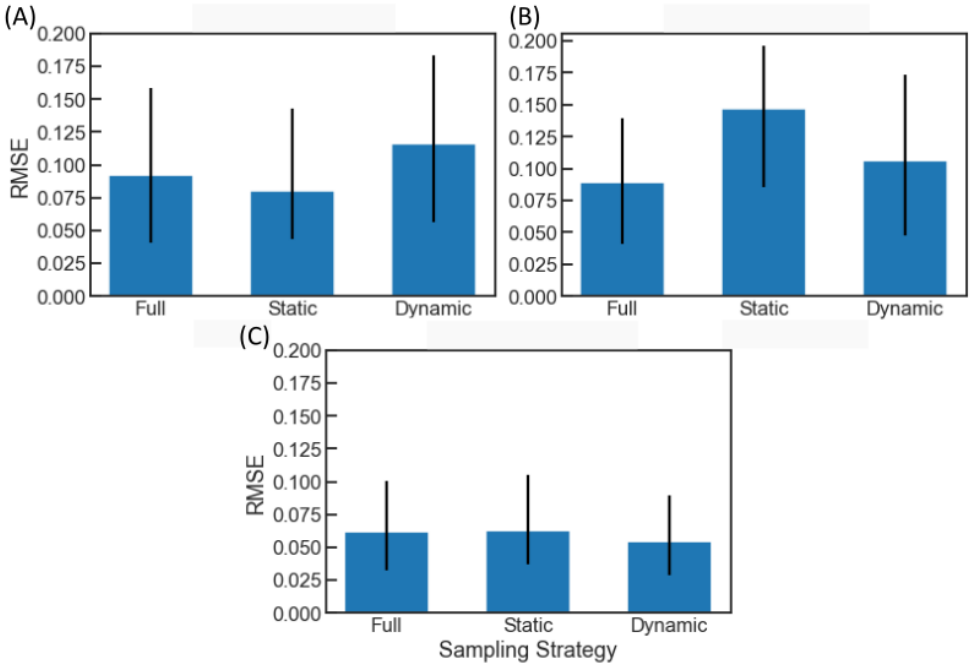


Figure 3.8: RMSE for the different sets of alternative ground truth functions shown in Figure 3.7A-C (A) RMSE for the functions shown in Figure 3.7A (B) RMSE for the functions shown in Figure 3.7B (C) RMSE for the functions shown in Figure 3.7C

These alternative and somewhat extreme ground truth functions which do not represent realistic perceptual encoding functions show that the sampling strategies do work to some degree at these extremes, but not everywhere. The highly decreased accuracy for the Static sampling strategy for Figure 3.7B is obvious and can be seen clearly in Figure B.

4

DISCUSSION

In this thesis I studied whether it is possible to improve the data acquisition for MLCM. This was achieved by reducing the amount of trials and consequently the experiments duration, for a fixed set of unique stimuli, without impacting the quality of the encoding function estimated using MLCM. After setting up the simulation using the ground truth functions and noise, the sampling strategies were applied to the trials. The simulated observer creates data which is then fed into the MLCM algorithm. The resulting perceptual scales are then compared to the ground truth functions of the simulation to calculate the RMSE. Multiple parameters for the Dynamic sampling strategy are also compared.

4.0.1 *Estimating the perceptual scales*

For a realistic noise level, MLCM is able to estimate perceptual scales that resemble the ground truth functions accurately using both sampling strategies. The sampling strategy reduce the amount of trials by $45,3\%$ for the static strategy and between $45-50\%$ for the dynamic strategy. The deviation is between 5% and 9% for the ten data points of the perceptual scales within the realistic noise range. The Static sampling strategy is able to estimate almost identical scales for most of the parameters used in the simulations, with minor deviations in unrealistically high noise level environments. The perceptual scales are not completely identical because of noise. Even after the average of 1000 simulations, there is a difference of less than 0.5% between the perceptual scales of the Static sampling strategy and the full (no) sampling strategy. The Dynamic sampling strategy is able to estimate scales which are similar but in almost all cases slightly worse than the perceptual scales estimated by using the full sampling strategy or the Static sampling strategy, being close to 10% in deviation. Both sampling strategies are able to estimate scales which are similar to the ones estimated by using the full sampling strategy for a wide range of perceptual encoding functions, but break at extreme functions like X^4 and $X^{0.2}$ where the accuracy of the perceptual scales drops to undesirable levels.

4.0.2 *Interpretation*

Both sampling strategies are able to generate enough data for MLCM to successfully estimate perceptual scales which resemble a multitude

of encoding functions while increasing the efficiency of the simulated experiment by around 45%-49%. The amount of trials can be reduced without impacting the quality of the encoding function estimated using MLCM, at least when using White's Illusion (1979), but likely also for other stimuli when adapting the sampling strategies. The improvement is expected to be similar for the actual experiment. As can be seen in Figures 3.3A and 3.3B, MLCM is unable to estimate perceptual scales which resemble the perceptual encoding function when the noise level at the decision stage is low ($< 2\%$). The accuracy is low, such that the scales do not approach the ground truth functions. The data used to estimate the perceptual scales can be interpreted as deterministic or close to deterministic. Deterministic results can be seen in every heatmap presented and consist of only two kinds of values in these heatmaps, zeroes and oness. These results provide the insight that the perceptual scales estimated by MLCM are largely unaffected by values that are zeroes and ones. This is not to say that the zeroes and ones are not informative, another algorithm may be able to estimate perceptual scales or something similar with these values, they just are not informative for MLCM. The same is indicated in every simulated experiment using the Static sampling strategy, as the estimated scales are almost identical to the ones estimated without using any sampling strategy. The Static sampling strategy aims to exclusively remove trials that have a very high agreement, which are exactly the trials that result in zeroes and ones.

The alternative ground truth functions shown in Section 3.3 showcase the limits of the sampling strategies as the Static sampling strategy has a lower accuracy for the scales shown in Figures 3.7B since the relevant trials to estimate the scales are part of the removed trials for these extreme functions. Unexpected is that Figure 3.7A shows a better result for the Static sampling strategy, where doing all trials for a cubic function might have been detrimental for the estimated scales. Also unexpected is that in Figure 3.7C, the Dynamic sampling strategy has a better accuracy than both no- and the Static sampling strategy. This is caused by the linear scale which means that there are only very few informative trials and that the results are not representative as the experiment has not been adjusted for these extreme functions. Figures 3.7A and B show the expected results where the Dynamic sampling strategy is less accurate than using the full sampling strategy, and in case of Figure 3.7B more accurate than the Static sampling strategy as the Static sampling strategy is not tuned for this set of ground truth functions.

4.0.3 *Limitations of the sampling strategies*

The Static sampling strategy has almost identical results when comparing it to using all trials. There is little to no loss in accuracy or

precision for a wide range of ground truth functions. The time for each experiment has been reduced by almost 50%, but this can vary between functions used and the set *luminance* thresholds at which to omit trials. The Static sampling strategy does however require some prior knowledge of the results to be set up efficiently. It is not agnostic. Minor improvements can be made using common sense, for example using a low *luminance* threshold for the same *context*, but this is difficult for different *contexts*. There are more optical Illusions than White's Illusion and the Static sampling strategy can not be applied to those and guarantee improved efficiency. The Dynamic sampling strategy, unlike the Static sampling strategy, is agnostic. Using multiple initial trial repeats, it can be applied to almost any form of Stimulus without the prior knowledge of the results. The gain in efficiency is comparable to that of the Static sampling strategy, averaging at 50%, but this is dependent on the ground truth functions and noise level used. It also comes at the cost of having lower precision and in some cases also a lower accuracy compared to a well set up Static sampling strategy.

The sampling strategies can work together by eliminating each others weaknesses. The Dynamic sampling strategy has the advantage of being agnostic, so it could be used to get a first estimate of the data. This estimate can then be used to set up the arguments for the Static sampling strategy and improve the accuracy and precision of all future results while maintaining a more efficient experimental procedure. Or using both sampling strategies at the same time for a very high increase in efficiency at the cost of a lower accuracy and precision of the Dynamic sampling strategy.

4.0.4 *Sampling the stimulus domain*

I use a linear scale normalized from zero to one in steps of 0.1. While I do have prior knowledge of the data and can use an improved non-linear scale based on the already created perceptual encoding functions, I test the sampling strategies on multiple types of ground truth functions with different shapes for which I do not have any prior knowledge. To achieve a uniformity of results, I use a linear scale for all simulated experiments. This also allows for the sampling strategies to be adapted to similar experiments without having to change the scales. For the actual experiment with participants a non-linear scale was used as reported by Vincent et al. (in preparation). The sampling strategies may need to be adjusted to improve the efficiency of data acquisition for a non-linear scale.

4.0.5 *Other sampling strategies*

Another sampling strategy which I have not used in any of the experiments is the random sampling strategy. It works by deciding to show

only 90% randomly sampled trials instead of every trial. Researchers can always decide to use a different percentage of trials to show. An example for this procedure can be seen in Shooner and Mullen (2022). I expect this to have an impact on the accuracy and precision similar to the Dynamic sampling strategy, which I try to mitigate using the Static sampling strategy.

4.0.6 *Recommendation*

The Dynamic sampling strategy proved to work even for extreme ground truth functions and is expected to work for other Stimuli as well. Without prior knowledge of the results, the Dynamic sampling strategy should be applied first with a linear scale to get a first impression of the perceptual scales and their shape. The data can then be analyzed to find parameters for the Static sampling strategy for improved accuracy, as well as a potential non-linear scale where more trials are conducted near "relevant" *luminance levels* for improved scales in the future. If the decrease in accuracy is a non-issue, both the static- and Dynamic sampling strategy can be applied together.

4.0.7 *Open questions*

The best approach for a realistic function would be a logit function $ln\frac{p}{1-p}$, but that function can not be normalised between zero and one. Both edges are $-\infty$ and ∞ respectively. Instead, I chose a multitude of power functions to use in the simulation.

One suggested way of finding lapses is that when looking at the results, any low agreement result that is surrounded by high agreement results, could be a lapse. This means that if any results does not have high agreement, but every "neighbour", so trials with similar *luminance levels*, has high agreement, the trial is most likely a lapse and the result could be corrected without falsifying the data.

Another aspect that could be considered in the future is the time it takes a participant to answer a trial. A trial with high agreement takes very little time since a participant does not have to think much before making a decision. A trial with very low agreement can take longer since participants might be unsure about which stimuli they deem more intense and will hesitate to give an answer within a short time frame. This does not take into account the time a participant might rest his eyes between trials or if they think about something else while doing the experiment, thus needing longer to answer disregarding the trial. There might be more benefits or limitations to this approach that I am not considering right now.

MLCM is able to estimate accurate perceptual scales when the data is non-deterministic. There could be a way to apply a reversed strategy. If we can determine for which trials a participant might be

able to answer with a high agreement, we could use that data to get similar scales as those estimated using MLCM. How this option can be approached can be a topic for another thesis, but the trials with high agreement are arguably easier to determine than those with low agreement.

Similar to finding lapses, the Dynamic sampling strategy could be improved by adding trials with high agreement of neighbours with low agreement to the selection of trials shown, reducing the efficiency but aiming to improve the accuracy to full sampling strategy levels.

4.1 CONCLUSION

This bachelor thesis researched if the data acquisition for MLCM can be optimized. The results of the simulations showed that the amount of unique trials in an experiment can be reduced without impacting the quality of the perceptual scales that MLCM estimates, or only impacting the quality by reducing the accuracy for less than 10%. Both developed sampling strategies can achieve an improved data acquisition through different means. The Static sampling strategy requires prior knowledge of the resulting data, but does not affect the accuracy or precision for a wide range of functions and noise levels. The Dynamic sampling strategy is agnostic, but affects the accuracy and precision of the estimated perceptual scales. While other sampling strategies like random sampling do offer a more efficient data acquisition, they do not focus on finding more informative trials. The dynamic approach also adds the opportunity to look at consistency and inconsistency for future trials while the experiment in being conducted. I did not expect the Static sampling strategy to work the way it does, simply removing trials and not replacing them with fake but predictable results instead. MLCM being unaffected by the missing trials was an unexpected discovery. The dynamic strategy, similar to the random sampling strategy, was expected to have an effect on the accuracy and precision of the estimated scales and works as intended.

REFERENCES

Georgeson, M. (2014). Spatial fourier analysis and human vision. *Tutorial Essays in Psychology: Guide to Recent Advances*, 39–88.

Knoblauch, K., and Maloney, L. T. (2012). *Modeling psychophysical data in r* (Vol. 32). Springer Science and Business Media.

Murray, R. F. (2021). Lightness perception in complex scenes. *Annual Review of Vision Science*, *7*, 417–436.

Shooner, C., and Mullen, K. T. (2022). Linking perceived to physical contrast: Comparing results from discrimination and difference-scaling experiments. *Journal of Vision*, *22*(1), 13–13.

Vincent, J., Maertens, M., and Aguilar, G. (in preparation). What matching can't do: Estimating perceptual encoding functions with perceptual scaling. *Journal of Vision*.

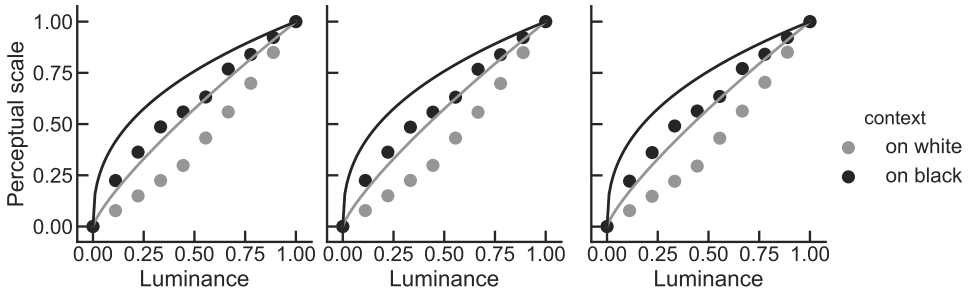
White, M. (1979). A new effect of pattern on perceived lightness. *Perception*, *8*(4), 413–416.

A

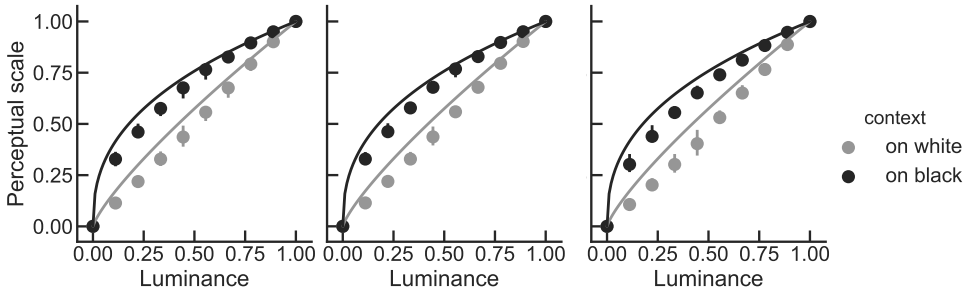
APPENDIX

A.1 OTHER NOISE LEVELS

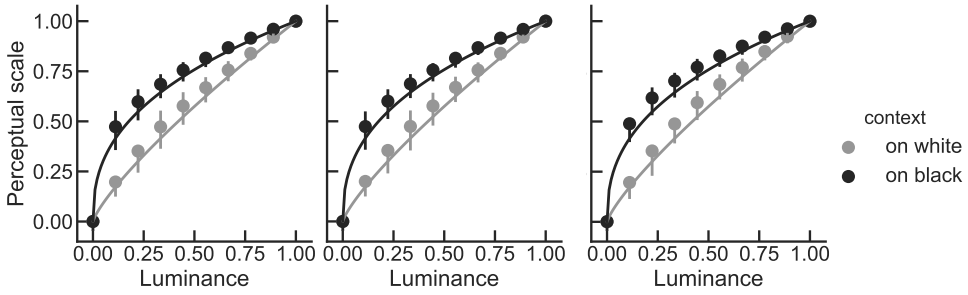
Here are more perceptual scales from simulations with noise levels that are not shown in the result section.



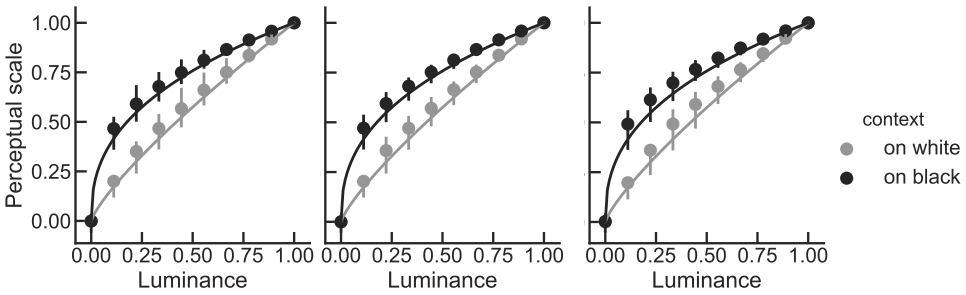
(F) Noise = 0.01



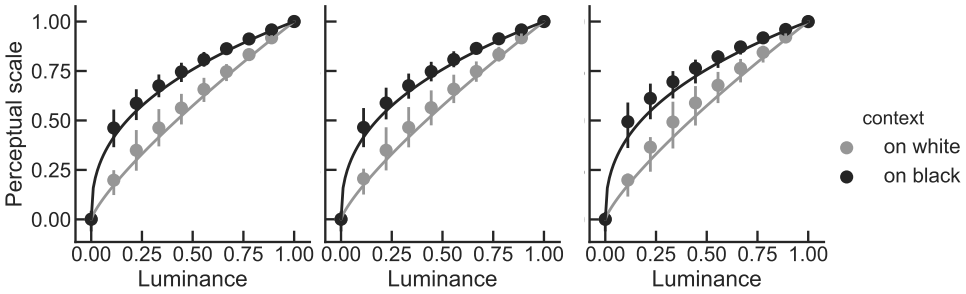
(G) Noise = 0.03



(H) Noise = 0.07



(I) Noise = 0.08



(J) Noise = 0.09