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

Abschlussarbeit

Investigating Inter-Individual Differences in Human Brightness Perception

---

vorgelegt von
Mohamad Anas Allaham
zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)
im Studiengang Informatik

Erstgutachterin: Prof. Dr. Marianne Maertens
Zweitgutachter: Prof. Dr. Felix Wichmann

21. Oktober 2022

SELBSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 21. Oktober 2022

Mohamad Anas Allaham

ABSTRACT

Brightness effects are often probed in psychophysical experiments since they can provide insight into the underlying mechanisms of visual perception. Such experiments are usually conducted using only a small number of observers and stimuli. A concomitant consequence of such an approach is the difficulty in comparing results from different studies and reliably determining the direction of some brightness effects.

The present work fills this need by conducting a psychophysical experiment that examined the variability among observers in perceiving the direction of brightness effects using a large set of stimuli. Observers judged the direction of brightness effects using a five-point rating scale. With few exceptions, the results show that the observed directions are consistent with those previously reported in literature. Nevertheless, inter-individual differences were evident with respect to the direction of effect and the confidence with which it was perceived. Further, Krippendorff's alpha yielded a value of 0.644, indicating low inter-observer reliability for the entire dataset.

ZUSAMMENFASSUNG

Helligkeitseffekte werden oft in psychophysischen Experimenten untersucht, da sie einen Einblick in die zugrundeliegenden Mechanismen der visuellen Wahrnehmung geben können. Solche Experimente werden in der Regel mit einer geringen Anzahl von Probanden und Stimuli durchgeführt. Ein Nachteil eines solchen Ansatzes ist die resultierende Schwierigkeit, die erhobenen Ergebnisse verschiedener Studien zu vergleichen und die Richtung einiger Helligkeitseffekte zuverlässig zu bestimmen.

Die vorliegende Arbeit erfüllt diesen Bedarf durch die Durchführung eines psychophysischen Experiments, das die Variabilität der wahrgenommenen Richtung von Helligkeitseffekten unter Probanden bei einer großen Menge von Stimuli untersuchte. Die Bewertung der Richtung von Helligkeitseffekten erfolgte anhand einer fünfstufigen Rating-Skala. Die Ergebnisse zeigen, dass die beobachteten Richtungen bis auf wenige Ausnahmen mit den zuvor in der Literatur berichteten Richtungen übereinstimmen. Nichtsdestotrotz zeigten sich interindividuelle Unterschiede hinsichtlich der Richtung des Effekts sowie der Sicherheit, mit der er wahrgenommen wurde. Darüber hinaus ergab Krippendorffs Alpha einen Wert von 0.644, was eine geringe Interbeobachter-Reliabilität des gesamten Datensatzes andeutet.

ACKNOWLEDGMENTS

I wish to thank my supervisors, Lynn Schmittwilken and Joris Vincent, for accompanying me along the way to my first academic milestone.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Sensations evoked by sensory stimuli, such as a light or a sound, result in meaningful psychological experiences of perception. These perceptions are context-dependent, variable and subjective. For instance, two people may not necessarily agree on how loud the same voice sounds, how soft the same touch feels or how bright the same light appears. The present work investigates such inter-individual differences in brightness perception. In the following, a brief introductory overview is given of what it means for an object to be bright.

1.1 LUMINANCE AND BRIGHTNESS

In a general sense, a light source, such as a lit candle, emits light in every direction. If shone onto a surface, the incident amount of light per unit area is referred to as *illuminance*. The proportion of illuminance the surface is able to reflect is called its *reflectance*. It is worthwhile to note that unlike reflectance, illuminance is independent of the nature of the surface itself (Hurvich and Jameson, 1966, p. 4). For instance, when viewed under identical lighting conditions, a white surface would have the same illuminance as a black one.

The mathematical product of illuminance and reflectance is in turn known as *luminance*. Measured in candelas per square meter (cd/m^2), luminance can be understood as the intensity of directional light that reaches the eye by being reflected off, emitted by or transmitted through a surface. *Brightness* is defined as apparent luminance; the psychological impression of light intensity. These definitions are based on those outlined in Adelson (2000).

It evidently follows that brightness is dependent on the two components of its physical equivalent; reflectance and illuminance. If we were to distinguish between two objects viewed against a uniform background, the object with a higher luminance would appear brighter. Likewise, if the objects were equally luminant they would be perceived as equally bright. Nevertheless, luminance is ambiguous in the sense that multiple combinations of illuminance and reflectance can result in the same non-unique luminance signal. In order to differentiate the reflectance of a surface from its illuminance, our visual system appears to attempt to disentangle this luminance signal, for which it usually does not have enough information (Blakeslee and McCourt, 2015a; Blakeslee et al., 2008). This is demonstrated by way of examples in the following section.

1.2 SURROUND-CONTEXT BRIGHTNESS EFFECTS

A *surround-context brightness effect*, or simply "*brightness effect*", is an instance where the perceived brightness of an object is changed by its surround. Such an effect occurs when an object with a constant luminance appears to vary in brightness depending on the context in which it is viewed. A popular example of this is the "checker-shadow stimulus" illustrated in Figure 1.1. The figure depicts a checkerboard consisting of dark-gray and light-gray squares that alternate in a regular pattern. The brightness of a uniformly-lit square[1] can be understood as the shade of gray the square appears to be painted in. Luminance would in turn refer to the actual gray-level value of the square's pixels; the color that an "eyedropper" tool would detect. Two squares of the checkerboard, although equal in luminance, appear to differ in brightness due to being viewed in different contexts.

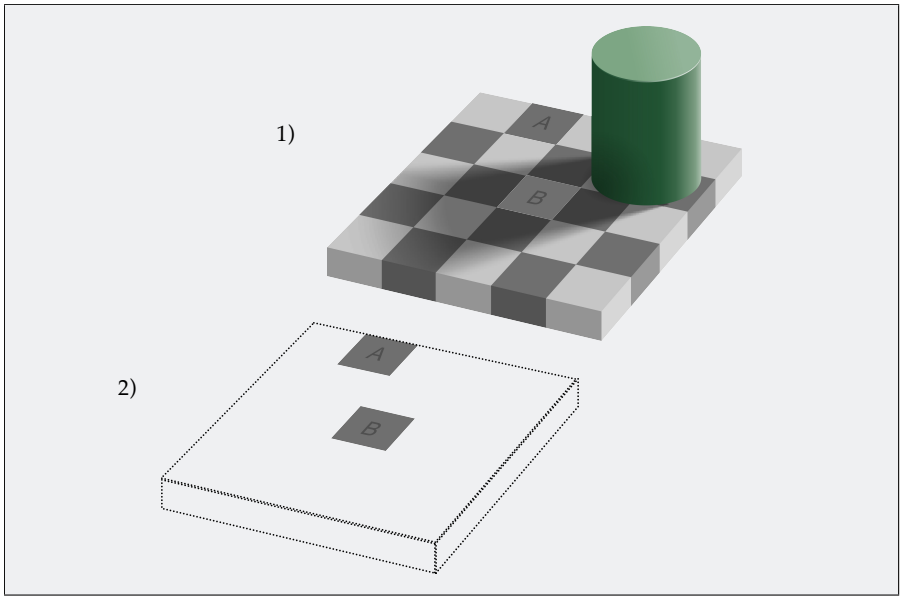

Figure 1.1: 1). The checker-shadow stimulus (Adelson, 1995). The equiluminant squares "A" and "B" appear to differ in brightness. Square "A" is observed in plain view while square "B" lies in the shadow cast by the green cylindrical object. This contributes to square "B" appearing brighter than square "A". 2). The squares are shown against a uniform background with the checkerboard and the shadow-casting object omitted, confirming the surround-context effect.

Figure 1.2 further demonstrates how different surround-contexts can have different effects on the same surface.

1 A uniformly-lit square is a square that is either completely in shadow or completely out of shadow. A square that is only partially shadowed would have at least two different brightness profiles.

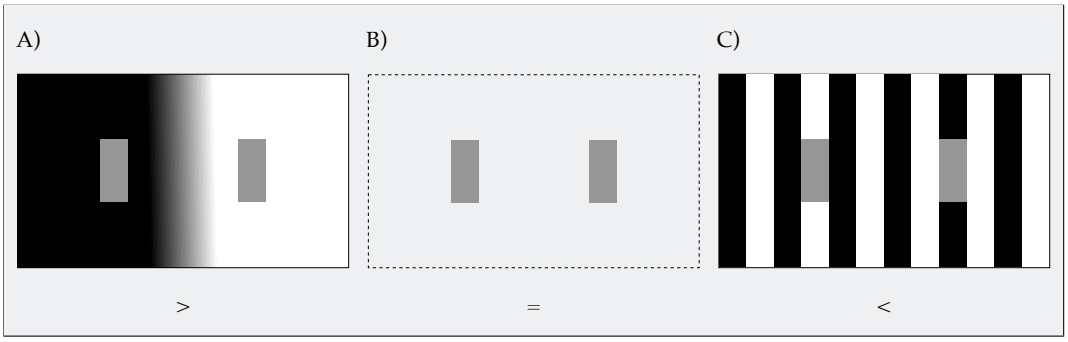

Figure 1.2: The importance of context on perceived brightness. Three physically identical pairs of equiluminant gray patches are viewed in three contexts. The symbols >, = and < refer to the direction of the respective brightness effect; which of the two patches is brighter. A). The simultaneous brightness contrast effect, where the left gray patch appears brighter than the right patch. Adapted from Maertens et al. (2015). B). The patches appear equally bright when viewed against a uniform background. C). White's effect (White, 1979), where the right patch appears brighter than the left patch.

White's stimulus and the simultaneous brightness contrast effect are extensively studied in literature and often used as illustrative examples of evident brightness effects with consistent directions. That is, when examining the figure above, people would—in all likelihood—agree on the indicated direction of effect for each stimulus. Nevertheless, some brightness stimuli tend to elicit disagreements among observers. Figure 1.3 presents an example of such a stimulus.

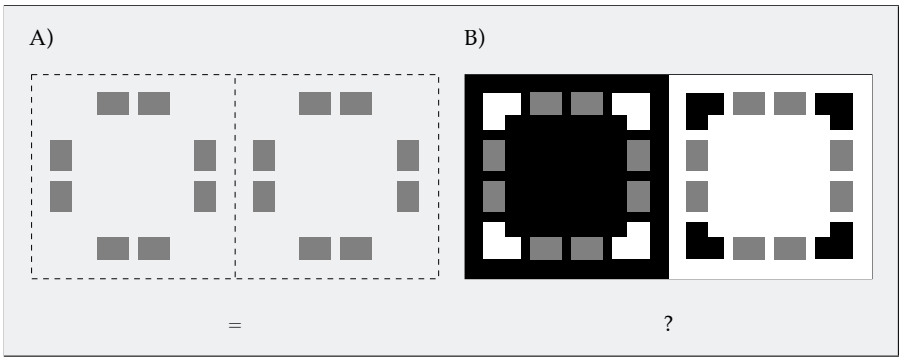

Figure 1.3: The "cube" stimulus (Agostini and Galmonte, 2002; Domijan, 2015) where observers are likely[2] to disagree on the direction of effect. A). Two groups of equiluminant rectangular gray patches appear equally bright against a uniform background. B). The "cube" stimulus; observers disagree on which or whether a group of patches (left or right) appears brighter than the other.

These are only a few examples of many known brightness effects. It is of particular interest to study such phenomena since they reveal information about the underlying mechanisms and assumptions used

2 This is corroborated by the findings of this work, which are reported in Chapter 3.

by our visual system when attempting to resolve the aforementioned luminance ambiguity (Blakeslee and McCourt, 2015b). Inter-individual differences in perceiving these stimuli would imply inter-individual differences in the processes that underlie visual perception, which raises a compelling need for investigating how people differ or concur in their perceptions.

1.3 QUANTITATIVE AND QUALITATIVE MEASUREMENTS OF BRIGHTNESS

Being a perceptual quantity, brightness does not have an objective measure. That is, it cannot be physically quantified using a photometer. Instead, brightness is estimated in psychophysical experiments. Typically, such experiments are conducted with a focus on measuring the strength (or magnitude) of brightness effects; which is to say, these experiments are usually *quantitative* in nature.

A popular technique widely used in such experimental designs is known as *matching*; where human observers adjust the luminance of a test stimulus until it perceptually matches that of a reference stimulus (du Buf, 2001). An alternative, more recent method is *maximum likelihood difference scaling*, which is capable of estimating interval perceptual scales[3] (Aguilar and Maertens, 2020). One of the downsides to these types of experiments is that many stimulus repetitions would be necessary for studying a sufficiently wide luminance range (Abebe et al., 2017). As a result, the length of the experiment increases, effectively confining the number of stimuli that can be used, as well as the number of observers willing to participate. To ensure that both the duration and the difficulty of the task are manageable, quantitative experiments are usually conducted using a small number of stimuli and observers. A resulting problem is the inability to compare the directions of brightness effects reported in existing studies, as different experiments use different stimuli and different participants. Consequently, it is difficult to reliably determine the direction of some brightness effects, particularly if they are understudied in literature.

Rather than measuring the magnitude of brightness effects, the present work is oriented towards measuring their direction. Such a *qualitative* approach allows for conducting a more extensive experiment where a significant number of brightness stimuli can be assessed by the same set of observers, effectively overcoming the previously mentioned limitations of quantitative studies. To this end, a simple rating scale is used to enable a quick and orderly recording of perceptual judgements.

3 Another method known as *maximum likelihood conjoint measurement* is also considered when estimating these scales (Aguilar and Maertens, 2020).

1.4 STUDY OBJECTIVES

The present work conducts a qualitative experimental study as described earlier investigating how the perception of the direction of brightness effects differs or agrees inter-individually. In this context, the first goal is producing comparable data on a large set of brightness stimuli. The data will then be analyzed to firstly assess the variability in the induced brightness effects across observers, and secondly to determine the average direction of effect of each stimulus. The study's final aim is measuring the consistent agreement among observers and determining the reproducibility of the dataset should the experiment be repeated using a different set of participants.

RESEARCH QUESTION: How do human observers differ or concur in their judgments of the perceived direction of selected brightness effects?

It is worthwhile to note that this work, although explorative in nature, does provide a dataset that is useful beyond its scope. For instance, the data can be used to test the validity and compare the performance of computational models that attempt to predict brightness as perceived by humans. So far, this has not been possible due to the aforementioned issue of the inability to reliably determine the direction of effect from existing studies. Further, the dataset can be useful in studying the stimuli themselves and investigating the extent to which their effects evoke and depend on similar underlying perceptual mechanisms. I do hope this work would be of value for researchers in the visual perception community and aid later efforts of better understanding the subtleties of brightness perception.

2 METHOD

To accurately assess brightness as a dependent experimental variable across different human observers, a certain consistency in the manner in which luminance is presented and viewed has to be ensured. This is of particular importance for ascertaining that the inter-individual differences the participants may exhibit are not due to differing experimental conditions. The experiment was therefore performed in a laboratory setting where a high level of control over pertinent parameters, such as luminance, visual angle and visual distance, can be achieved. In contrast to a naturalistic, real-world setting, this controlled environment effectively allows us to record subject responses under equivalent conditions and account for any extraneous variables that could influence visual behavior, such as ambient illumination. As a result, we obtain a dataset that is coherent and reflective of inter-individually comparable perceptual judgments. This approach also allows for the experiment to be replicated by other members of the visual perception community and expanded upon in terms of adding additional data from further observers. To this end, a complete documentation of this study is presented.

2.1 STIMULI

Rendered images representing 45 brightness effects drawn from Murray (2020), Domijan (2015) and Robinson et al. (2007) comprise the stimuli set. Each stimulus consists of two equiluminant gray regions ("targets"), one on the left and one on the right, embedded within a black-and-white surround. Being the only gray areas of an otherwise black-and-white image, the targets are meant to be unambiguously identifiable by naïve observers. Collectively, the stimuli depict different surround-contexts and instantiate a wide variety of brightness phenomena. Figure 2.1, Figure 2.2 and Figure 2.3 illustrate the stimulus set used.

The stimuli were presented on a neutral gray background with a luminance of 50.16 cd/m^2. In every case, black regions had a luminance of 0.31 cd/m^2, white regions a luminance of 100 cd/m^2, while the luminance values of the gray targets varied. Table 2.1 gives an overview of the stimulus parameters used.

In an effort to increase measurement accuracy and avert the effect of misleading outliers, six repetitions of each stimulus were presented. Three of these repetitions showed mirror-flipped versions of the stimuli, so as to diminish memory and carry-over effects. To further miti-

gate these, the stimuli were displayed in a fixed order that guarantees at least five intervening trials between "similar" stimuli; that is, stimuli that depict variations of the same effect or are differently sized or mirror-flipped versions of one another.

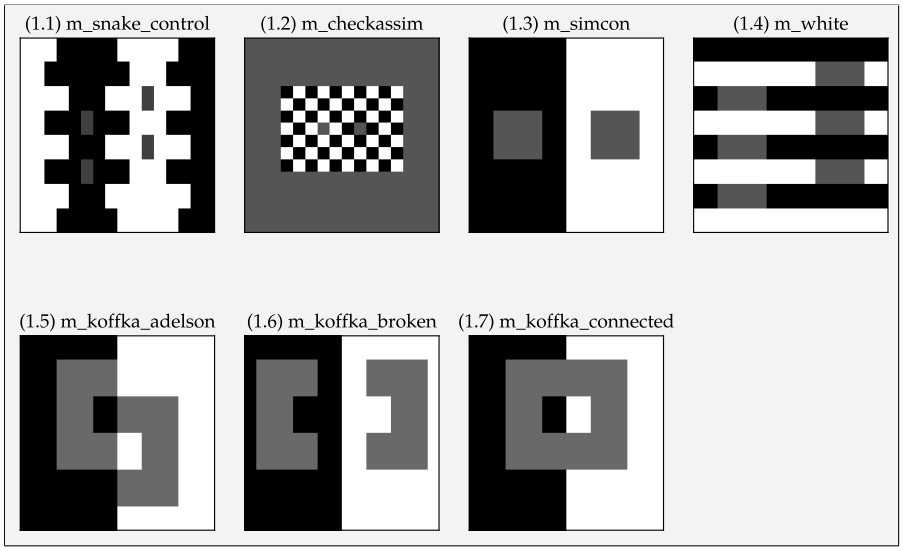

Figure 2.1: Stimuli from Murray (2020). Stimulus 1.1 was rotated by 90° counterclockwise to enable a left-right target comparison. The initial "m" refers to Murray and is used as an identifier for the publication.

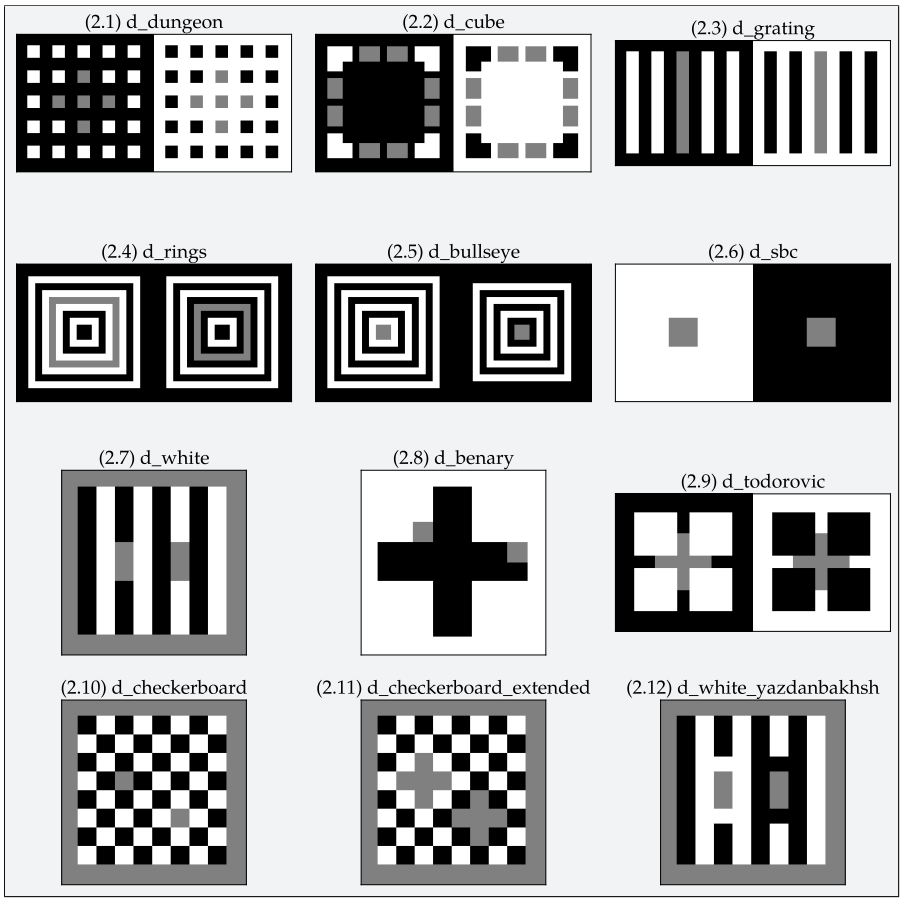

Figure 2.2: Stimuli from Domijan (2015). The initial "d" refers to Domijan and is used as an identifier for the publication.

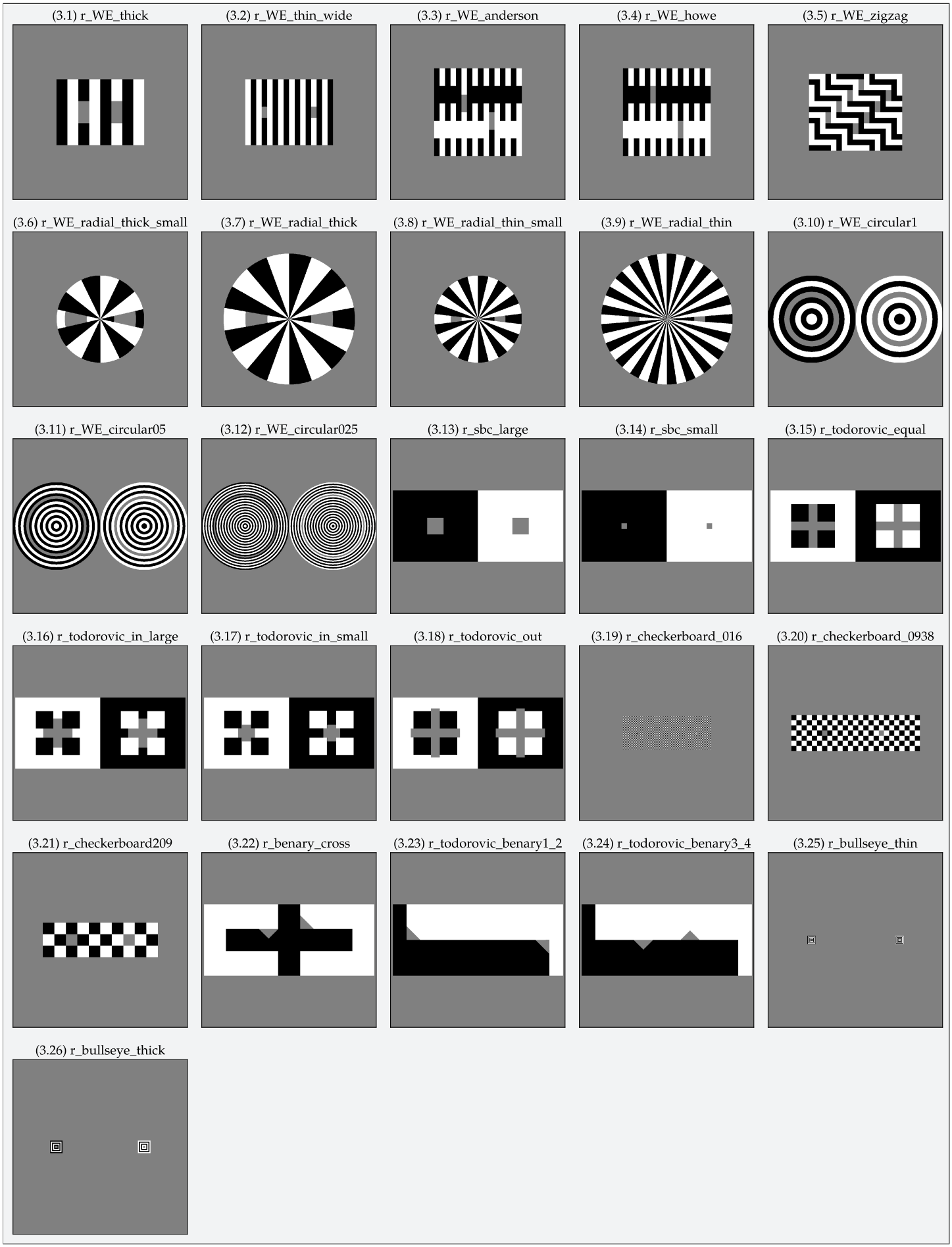

Figure 2.3: Stimuli from Robinson et al. (2007). Stimuli 3.6, 3.7, 3.8 and 3.9 were rotated by 90° counterclockwise to enable a left-right target comparison. The initial "r" stands for Robinson et al. and is used as an identifier for the publication.

Stimulus	Id	Complete name	Size (w x h)	Target luminance (cd/m^2)
m_snake_control	1.1	Snake control figure	8 x 8°	25.23
m_checkassim	1.2	Checkerboard assimilation	8 x 8°	33.54
m_simcon	1.3	Classic simultaneous contrast figure	8 x 8°	33.54
m_white	1.4	White's illusion	8 x 8°	33.54
m_koffka_adelson	1.5	Koffka-Adelson figure	8 x 8°	41.60
m_koffka_broken	1.6	Koffka ring, broken	8 x 8°	41.60
m_koffka_connected	1.7	Koffka ring, connected	8 x 8°	41.60
d_dungeon	2.1	Dungeon illusion	23.9 x 12°	50.16
d_cube	2.2	Cube illusion	24 x 12°	50.16
d_grating	2.3	Grating illusion	26.5 x 12°	50.16
d_rings	2.4	Ring patterns	24 x 12°	50.16
d_bullseye	2.5	Bullseye display	24 x 12°	50.16
d_sbc	2.6	Contrast-contrast effect	24 x 12°	50.16
d_white	2.7	White's effect	12 x 12°	50.16
d_benary	2.8	Benary's cross	12 x 12°	50.16
d_todorovic	2.9	Todorović's illusion	24 x 12°	50.16
d_checkerboard	2.10	Checkerboard contrast	12 x 12°	50.16
d_checkerboard_extended	2.11	Checkerboard contrast extended	12 x 12°	50.16
d_white_yazdanbakhsh	2.12	White's effect-Yazdanbakhsh	12 x 12°	50.16
r_WE_thick	3.1	White's effect-thick	32 x 32°	50.16
r_WE_thin_wide	3.2	White's effect-thick-wide	32 x 32°	50.16
r_WE_anderson	3.3	White's effect-Anderson	32 x 32°	50.16
r_WE_howe	3.4	White's effect-Howe	32 x 32°	50.16
r_WE_zigzag	3.5	White's effect-zigzag	32 x 32°	50.16
r_WE_radial_thick_small	3.6	White's effect-radial-thick-small	32 x 32°	50.16
r_WE_radial_thick	3.7	White's effect-radial-thick	32 x 32°	50.16
r_WE_radial_thin_small	3.8	White's effect-radial-thin-small	32 x 32°	50.16
r_WE_radial_thin	3.9	White's effect-radial-thin	32 x 32°	50.16
r_WE_circular1	3.10	White's effect-circular-1	32 x 32°	50.16
r_WE_circular05	3.11	White's effect-circularo.5	32 x 32°	50.16
r_WE_circular025	3.12	White's effect-circularo.25	32 x 32°	50.16
r_sbc_large	3.13	Simultaneous brightness contrast-large	32 x 32°	50.16
r_sbc_small	3.14	Simultaneous brightness contrast-small	32 x 32°	50.16
r_todorovic_equal	3.15	Todorović-equal	32 x 32°	50.16
r_todorovic_in_large	3.16	Todorović-in-large	32 x 32°	50.16
r_todorovic_in_small	3.17	Todorović-in-small	32 x 32°	50.16
r_todorovic_out	3.18	Todorović-out	32 x 32°	50.16
r_checkerboard_016	3.19	Checkerboard-0.16	32 x 32°	50.16
r_checkerboard209	3.20	Checkerboard-209	32 x 32°	50.16
r_checkerboard_0938	3.21	Checkerboard-0.94	32 x 32°	50.16
r_benary_cross	3.22	Benary cross	32 x 32°	50.16
r_todorovic_benary1_2	3.23	Todorović-Benary 1–2	32 x 32°	50.16
r_todorovic_benary3_4	3.24	Todorović-Benary 3–4	32 x 32°	50.16
r_bullseye_thin	3.25	Bullseye-thin	32 x 32°	50.16
r_bullseye_thick	3.26	Bullseye-thick	32 x 32°	50.16

Table 2.1: Stimulus parameters and naming scheme.

2.2 TASK

The observers were requested to discriminate between the targets of each stimulus by indicating which of the two (left or right) they perceived as brighter. An ordinal five-point Likert-type scale was used as a psychometric tool to measure subject responses. The rating scale consisted of the categories "equally bright", "maybe brighter" and "definitely brighter" as demonstrated in Figure 2.4. In wording the response options, the adverbs "maybe" and "definitely" were chosen over other common terms, such as "somewhat" and "much", as the intention behind using this five-point design was not to quantify the magnitude of the effects, but to capture the confidence with which observers answer.

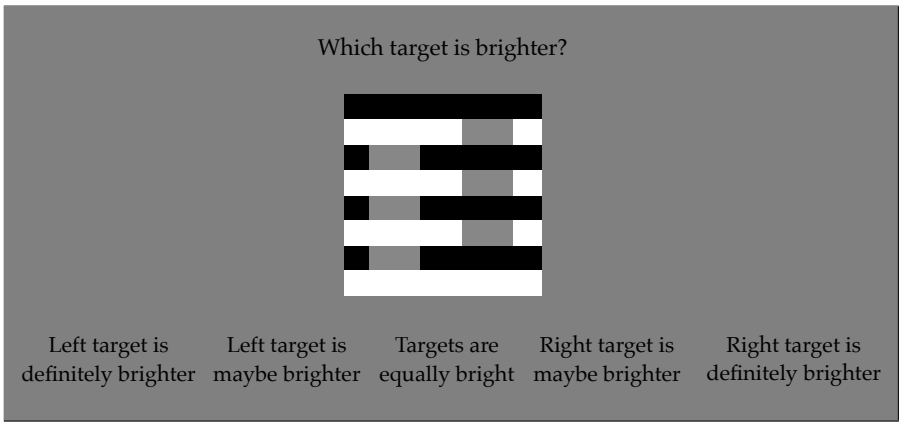

Figure 2.4: Representation of the experimental task and rating scale. The participants were required to select one of the five answers to advance to the next trial. As such, the task can be designated as a forced-choice discrimination task (Kingdom and Prins, 2016, pp. 24, 29).

Prior to beginning the experiment, five practice trials were conducted to ensure the familiarity of the subjects with the task and the given rating scale. In these trials, dummy stimuli depicting exemplary scenarios for each of the five response options were presented. Figure 2.5 gives an overview of these stimuli. The first two practice trials showed examples of the endpoints of the scale (stimuli 4.1 and 4.2) and were completed under the experimenter's supervision. Before these two answers were recorded, the observers were verbally asked which response option they would select. This was done to ensure that the subjects would not avoid choosing answers at the extreme ends of the scale, as is often the case in rating experiments (Cunningham and Wallraven, 2011, pp. 73–74). Incidentally, every participant had "correctly" chosen the leftmost and rightmost answers without having been explicitly instructed to do so. The dummy stimuli were also used as catch trials to ensure that the observers were attending to the experimental stimuli and paying attention throughout the experiment.

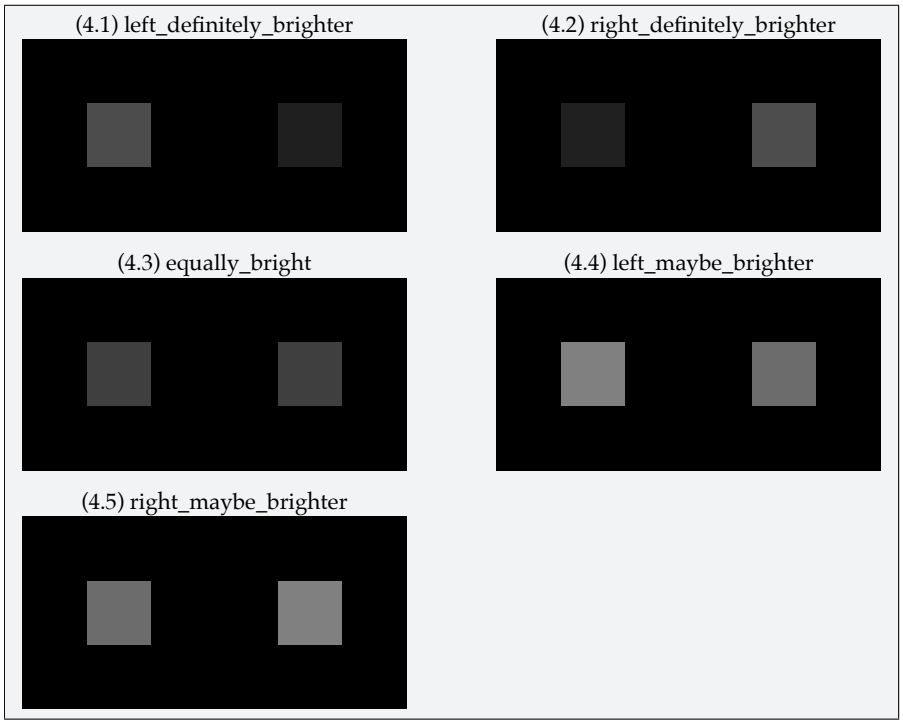

Figure 2.5: Practice and catch stimuli, each named after the corresponding response option in the rating scale. The targets of stimuli 4.1 and 4.2 had a luminance of 10 and 30 cd/m^2. Comparably, the targets of stimuli 4.4 and 4.5 were 40 and 50 cd/m^2 in luminance. Stimulus 4.3 had targets with a luminance of 20 cd/m^2. As catch trials, these stimuli appeared in order upon completion of 20%, 40%, 60%, 80% and 99% of the experimental trials.

2.3 APPARATUS

The experiment was displayed on a 24-inch VIEWPixx/3D LCD monitor (VPixx Technologies Inc., Saint-Bruno, QC, Canada) with a resolution of 1920 x 1080 pixels and a refresh rate of 120 Hz. Luminance output was within a range of 0-100 cd/m^2 at a 16-bit resolution. Python's High-Resolution Luminance (HRL)[1] library was used to implement the experiment and present the stimuli. A handheld five-button ResponsePixx response box (VPixx Technologies Inc., Saint Bruno, QC, Canada) acted as the response input device. Observers were able to navigate between the five response options by pressing the left and right buttons. Once the desired response option had been highlighted, they confirmed their answer by pressing the middle button.

1 Source code for HRL can be found on GitHub: https://github.com/computational-psychology/hrl

2.4 PARTICIPANTS

16 observers, of whom four are designated as "expert" and 12 as "naïve", participated in the study. Two additional observers were excluded from the analysis as post-experiment debriefings showed that they failed to perform the task appropriately. The first of the two reported disregarding the brightness effects they perceived and basing their answers on the immediate first impression they gained from the stimuli, which is that the target patches are equally bright. This led to a counterfactual response style that is inconsistent with their actual perception, as they noted brightness effects becoming more and more apparent the longer they viewed the stimuli. The second excluded participant had conjectured the physical identicalness of all target patches and indicated purposely trying to overlook the effects, rendering the data unsuitable for inclusion and use.

Following standard ethical research practices, informed consent was obtained from all observers prior to beginning the experiment. The expert participants consisted of the author (p_{05}) and three visual perception researchers (p_{01}, p_{02} and p_{03}). Participation was financially incentivized and naïve subjects were reimbursed for their time. Both expert and naïve observers had normal or corrected-to-normal visual acuity.

2.5 PROCEDURE

After having all their questions answered by the experimenter, the subjects were requested to seat themselves in a dark experimental chamber where they were able to perform the task under controlled conditions. The experimental setup was shielded from extraneous light sources by opaque blackout curtains, so as to ensure that the stimuli were perceived in the desired luminance levels. A chin-rest, on which the observers placed their chins, was positioned at a distance of 80 cm from the display monitor, effectively standardizing both the visual distance and the visual angle. During the experiment, the observers were not given feedback on their answers, nonetheless they were encouraged to ask questions and report issues; should they for instance not be able to locate the targets or mistakenly choose an answer that does not correspond to their perception. Every observer completed a total of 280 trials (five practice trials, five catch trials and 270 experimental trials)[2]. Three optional breaks built into each experimental session were offered once 25%, 50% and 75% of the trials were completed. On average, the participants took 17 minutes to complete the task, albeit no time limit was imposed.

2 Expert observer p_{03} took part in the experiment in its early stages and only completed 225 experimental trials (five repetitions of each stimulus).

2.6 DATA QUESTIONS

The objective of the research question is to investigate how the perception of the direction of brightness effects varies inter-individually. Based on the previously described experimental design, the research question will be answered from three aspects within the scope of the present work. The first of which is measuring the variability in the perceived direction of effect for each stimulus across observers. High variability would indicate low agreement on the direction of effect and/or the confidence with which it is perceived. Likewise, a lack of variability would imply high inter-observer agreement. This will enable us to categorize the stimuli by the consistency of the effects they induce across observers, and in so doing, identify those stimuli that produce the least and most consistent effects.

1ST DATA QUESTION: How variable are brightness judgements across observers for each stimulus?

The second aspect to be addressed is determining the average direction of effect for each of the 45 brightness stimuli across observers. This will enable us to identify and group together those stimuli whose effects are on average perceived in the same direction by observers. Further to reiterate, the rating scale takes into account how confident an observer is when perceiving an effect in either direction. As a result, determining the average direction of effect will further give us a sense of the average confidence with which the effect was perceived. This effectively allows us to identify the stimuli that elicit the least and most confident responses.

2ND DATA QUESTION: What is the average direction of effect per stimulus across observers?

Lastly, there is the question of the degree of consistent agreement among participants and whether such an agreement can be reproduced when repeating the experiment. Answering this could provide us with an indication of how similar brightness perception is among human observers.

3RD DATA QUESTION: To what extent do human observers consistently agree on the perceived direction of brightness effects?

3

RESULTS

Due to the ordinal nature of the data, the median of the six judgments made by each observer for each stimulus was adopted as the most appropriate measure of central tendency. Nevertheless, it should be noted that some caution is warranted when interpreting the results in the sections to follow. The median, being the midpoint between six observations, can take values that are in between the five original response categories, which may lead to erroneous interpretations:

Figure 3.1: Internal coding of the rating scale. From left to right, the response categories were assigned numerical values between 1 and 5. As a result of taking the median of an even number of observations, four additional points with the values 1.5, 2.5, 3.5 and 4.5 have to be considered. These are indicated by the small, unlabeled circles.

In Figure 3.1, the four in-between points are not to be misinterpreted as distinct response categories representative of responses that are exactly halfway between two original (labeled) scale points. This would assume equidistant response categories and contradict the premise that the data is ordinal. For instance, a median of 1.5 should not be interpreted as an observer choosing a response that is halfway between "1" and "2", whatever semantic meaning this response may have. Instead, the value would merely indicate that on average, the observer responded with "1" or "2". With this in mind, we can begin addressing each of the data questions raised earlier in turn in the following three sections.

3.1 VARIABILITY OF BRIGHTNESS EFFECTS

The first question concerns the variability in brightness judgments for each stimulus across observers. The interquartile range was calculated[1] and used as a measure of this variability. Figure 3.3 illustrates a heatmap representation of the average brightness judgments made by each observer for each stimulus. The columns of the heatmap represent

1 The interquartile range was calculated using SciPy's stats.iqr function with linear interpolation.

individual observers, while the rows correspond to the 45 brightness stimuli. The rows have been sorted according to the interquartile range, whose values are displayed on the right axis. Following these results, the stimuli can be grouped together based on the consistency of the effects they induce. For instance, we can observe that the stimuli with an interquartile range of zero induced the most consistent (least variable) effects, while the stimuli with an interquartile range of two produced the least consistent (most variable) effects across observers.

A noteworthy observation is that each participant was unique in how they perceived the set of stimuli in its entirety; that is, no two columns of the heatmap are identical. This may come at no surprise, since it is not unreasonable to assume that for such a large stimulus set, observers will inevitably vary in at least how confident they are in some of their judgments. However, even if we were to disregard confidence and only consider the observed direction of effect, each observer would still exhibit a unique perception of brightness and disagree with at least one other observer on the direction of effect of at least one stimulus. The highest agreement on the direction of effect can be seen between participants p_{01} and p_{03}, who agree on all stimuli apart from "m_snake_control".

Without necessarily restricting ourselves to the direction of effect, other similar response patterns can be readily detected. For instance, participants p_{09} and p_{14} gave similar responses to nearly all stimuli, further suggesting that groups of people might share the same visual behavior. Likewise, expert observer p_{02} was unique in perceiving some brightness effects in the opposite direction to all other participants. While it is possible that their brightness perception is completely individualistic, it could also be shared by a distinctive minority of people not represented in this sample of 16 observers. Such response patterns are useful to detect because they give us insight about those stimuli that differentiate between observers' perceptions and give rise to inter-individual differences.

3.2 AVERAGE DIRECTION AND CERTAINTY OF BRIGHTNESS EFFECTS

The second question addresses the average direction of effect and the implicit average certainty with which it was perceived. Based on the median response, Figure 3.4 illustrates the distribution of the brightness stimuli in terms of these two characteristics. First, as presented in the figure, the stimuli can be initially grouped into three main categories by the direction of the effect they produce. The bluish spectrum represents stimuli where the left target was perceived as brighter, the reddish spectrum represents stimuli where the right target was perceived as brighter, and the gray dots represent stimuli whose targets were perceived as equally bright.

Second, the stimuli can be further grouped by the certainty of the brightness effects they induce in either direction. The more saturated the shade of blue or red is, the higher the confidence with which the effect was perceived. For instance, we can observe that the left targets of the stimuli "m_white" and "d_rings" were perceived as brighter with the most certainty. Likewise, the right targets of the stimuli "r_WE_circular05", "r_WE_circular025", "r_WE_circular1", "r_bullseye_thin", "r_bullseye_thick", "r_checkerboard_016" and "r_checkerboard_0938" were perceived as brighter with the most certainty. Incidentally, four of these stimuli also produced the least variable effects across observers. These are represented in in Figure 3.2.

Figure 3.5 further illustrates the underlying distribution of average responses for each stimulus (each distribution was previously represented by the median in Figure 3.4). A few interesting results can be directly extracted; for instance, "m_white" is the only stimulus whose effect was perceived in the same direction by every observer. On the other hand, "d_cube" appears to have caused the most division on the perceived direction of effect among observers.

It follows that both the direction of effect and the confidence with which it is perceived vary to different extents depending on the stimulus. An interesting observation worth noting is that the effects of many "classical" brightness stimuli, such as the simultaneous brightness contrast, were perceived with uncertainty. This was particularly the case with naïve observers, suggesting that these effects may not be as apparent if the observers are unfamiliar with them.

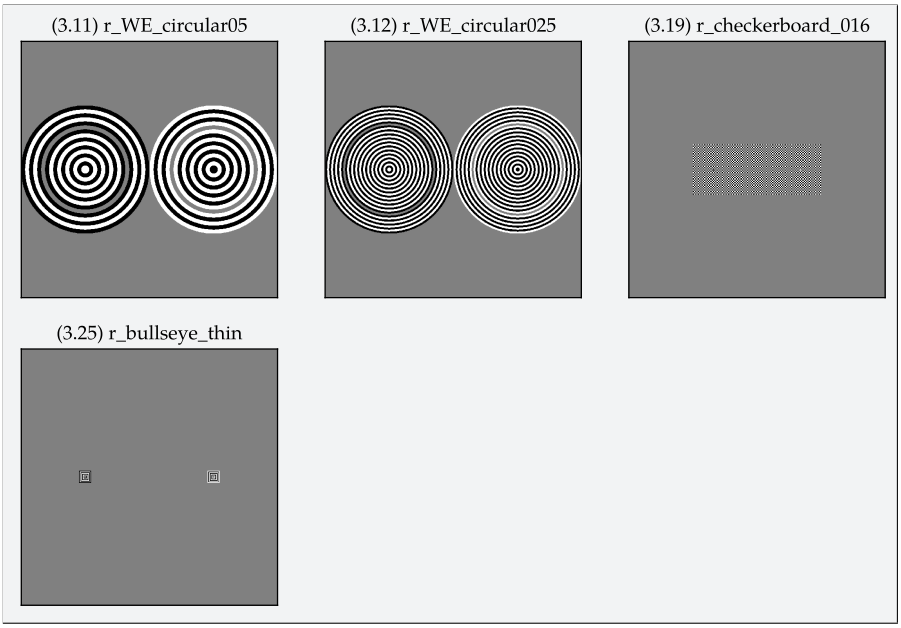

Figure 3.2: Stimuli eliciting of the most confident and least variable brightness judgments across observers.

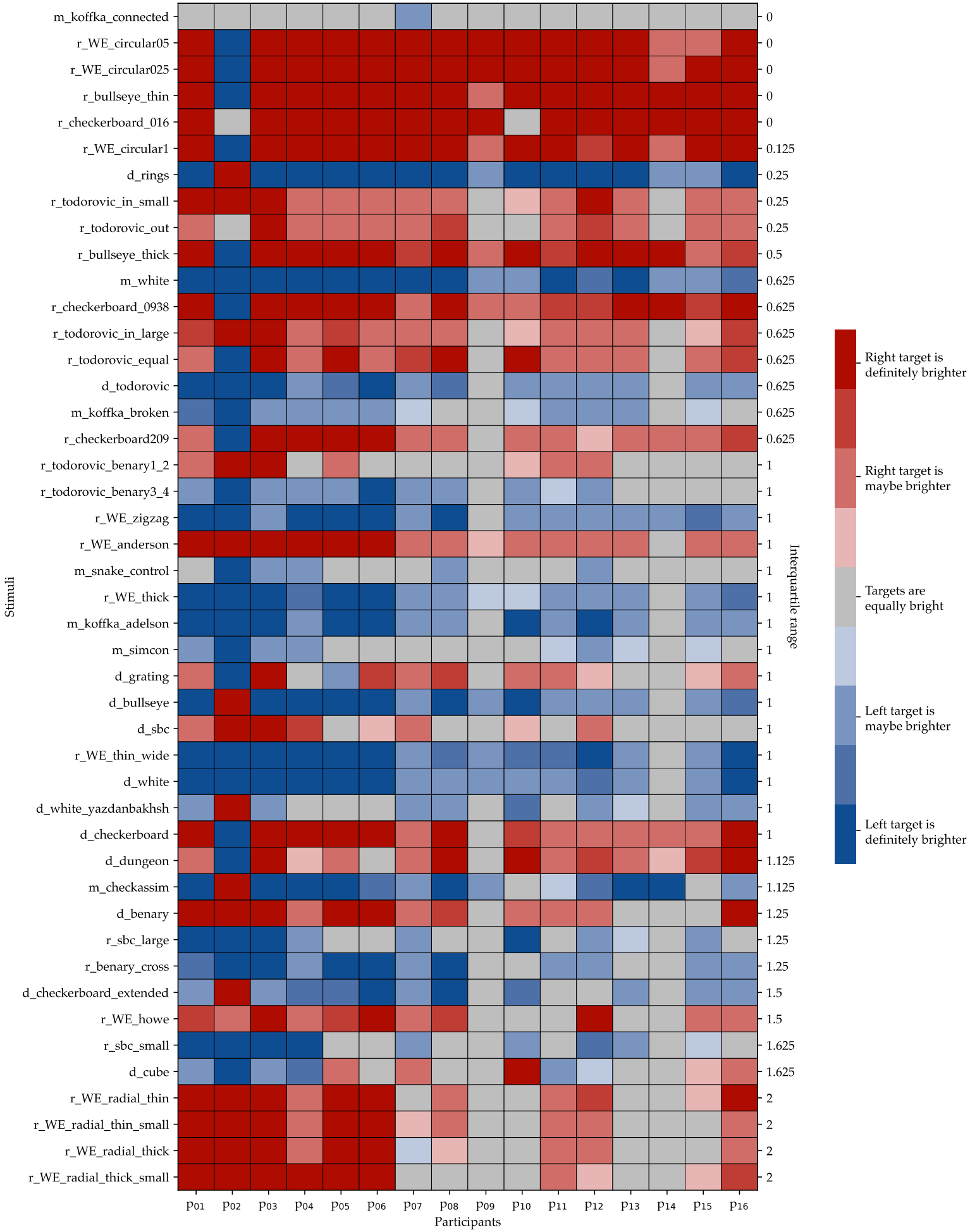

Figure 3.3: Heatmap representation of the dataset. Each cell represents the average direction of effect perceived by an observer. As a measure of the variability in each row, the respective interquartile range value is displayed on the right axis. Participants p_{01}, p_{02}, p_{03} and p_{05} are expert while the rest are naïve.

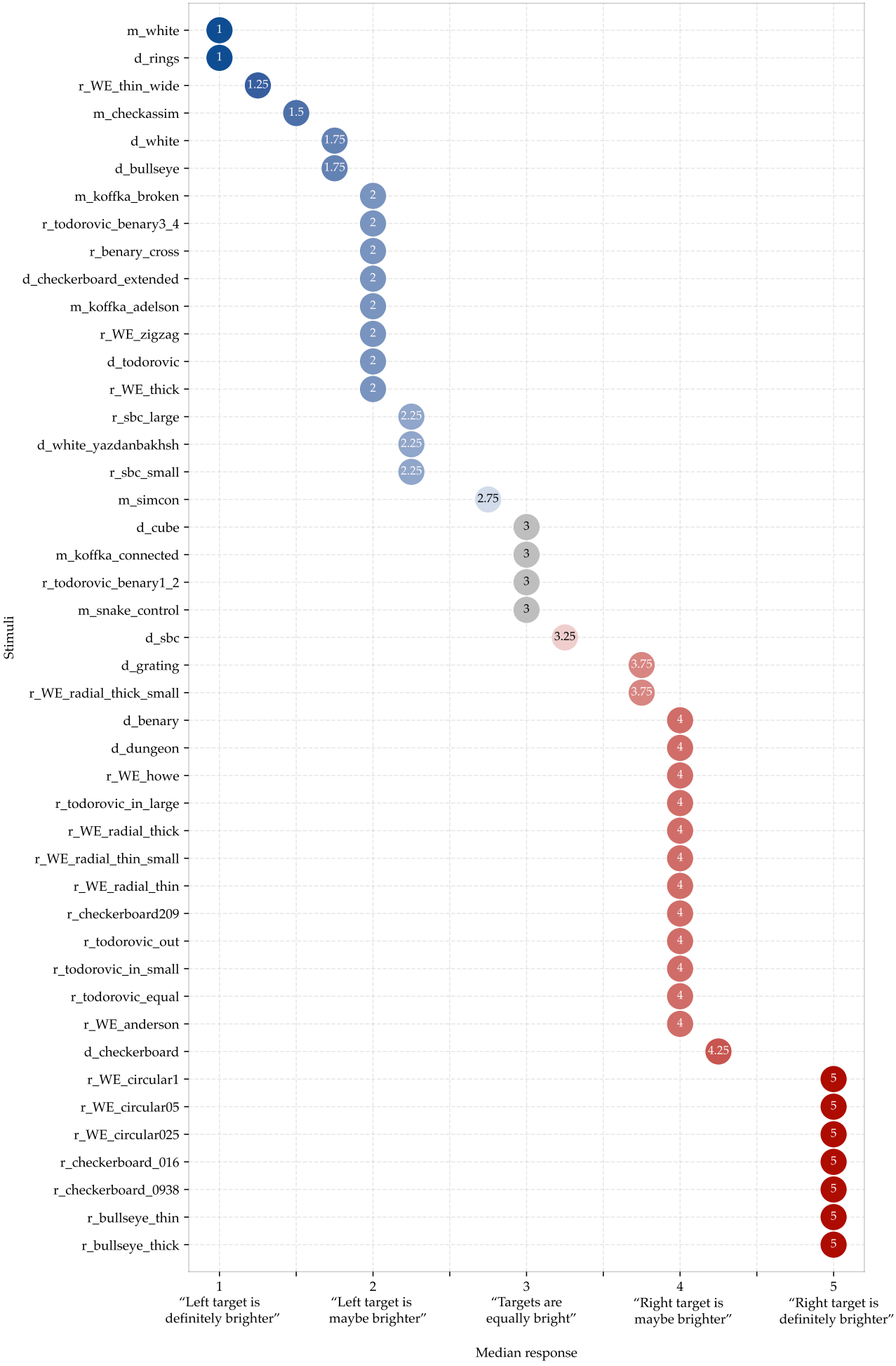

Figure 3.4: Stimulus groupings based on the median response. The stimuli can be grouped by the direction of effect (right or left/red or blue), its certainty (distance from 3/shade of either color) or both (stimuli that share the same median value). The targets of the stimuli with a median of 3 were perceived as equally bright.

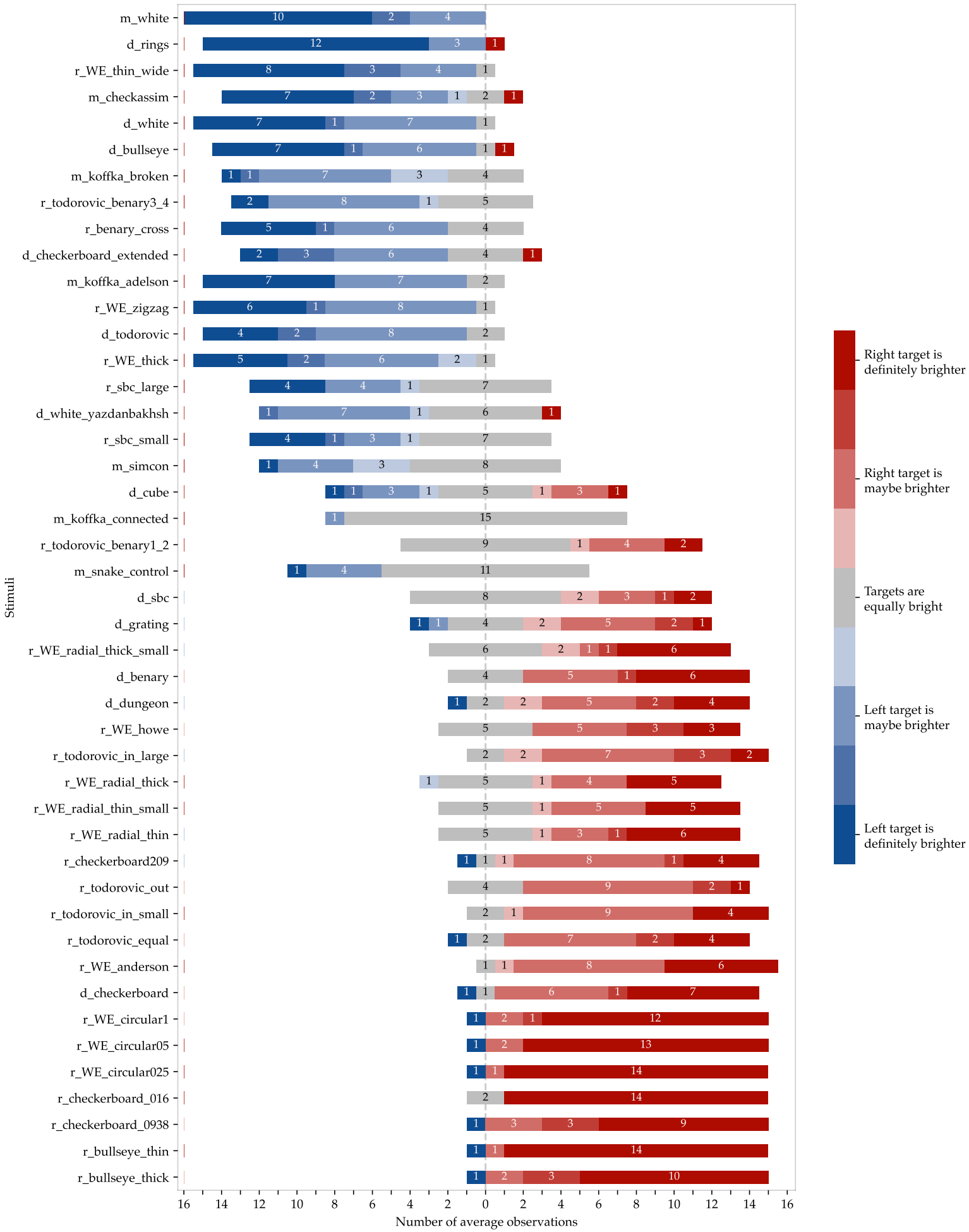

Figure 3.5: Distribution of average responses per stimulus. The numerical values on each bar represent the number of observers whose average response falls within the respective response category. For instance, 10 out of 16 observers perceived the right target of "r_bullseye_thick" as definitely brighter.

3.3 RELIABILITY OF BRIGHTNESS EFFECTS

The third question examines *inter-observer reliability*; the consistent, reproducible agreement among different observers (Gwet, 2021, p. 4). A subtle distinction between the terms "agreement" and "reliability" should be underlined. Agreement is what is measured, and reliability is an inference drawn from it (Krippendorff, 2004b). In the context of this work, inter-observer reliability is evaluated by measuring the agreement in perceptual judgments achieved among observers and refers to the extent to which this agreement can be reproduced when repeating the experiment under the same conditions.

In order to assess reliability, Krippendorff's alpha (α), a chance-corrected "*agreement measure with appropriate reliability interpretations*" (Krippendorff, 2004a, p. 221), was used as the most suitable metric. α applies to ordinal data, measures agreement between more than two observers, and considers partial agreement by assigning different weights to different response categories (Krippendorff, 2011). Its value ranges between 0 and 1 when evaluating reliability, and as a guideline, Krippendorff suggests considering data as reliable if $\alpha > 0.800$, unreliable when $\alpha < 0.667$ and tentatively reliable for $0.667 \leq \alpha \leq 0.800$ (Krippendorff, 2004a, pp. 222, 241). Although α can qualify as being mathematically complex and computationally intensive, the following subsection gives an overview of how it can be calculated for our specific dataset. This is useful for a better and more nuanced understanding of its result.

3.3.1 *Calculating Krippendorff's alpha*

Krippendorff's alpha is usually calculated using so-called coincidence matrices (Krippendorff, 2011). The following discussion deviates from this and provides a more simplistic approach. Unless otherwise stated, the use of definitions and equations follows Krippendorff (2011).

First, some terminology and notation should be introduced. Let us suppose that three observers ($o_{1,2,3}$) were requested to judge three different stimuli ($s_{1,2,3}$) once using our rating scale. Their responses are represented in the following matrix M, where each row corresponds to an observer and each column represents a stimulus:

$$M : \begin{array}{c} \begin{matrix} s_1 & s_2 & s_3 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 5 \\ 2 & 2 & 3 \\ 1 & 2 & 5 \end{pmatrix} \end{array} \begin{matrix} o_1 \\ o_2 \\ o_3 \end{matrix} .$$

(a) C is defined as the set of response categories; the possible values that an observation can take. In this case, $C = \{1, 2, 3, 4, 5\}$. Note

that the value "4", despite being in C, was never assigned to any stimulus.

(b) A *unit* u is a multiset containing the responses for a stimulus. In this example, M comprises three units corresponding to each of its columns: $u_1 = \{1,1,2\}, u_2 = \{1,2,2\}$ and $u_3 = \{3,5,5\}$.

(c) R denotes a multiset containing every recorded response. In other words, R is a multiset that contains every element of M. In this case, $R = \{1,1,1,2,2,2,3,5,5\}$.

(d) $\nu(S,x)$ denotes the number of occurrences of an element x in a multiset S. For instance, $\nu(R,1) = 3$ and $\nu(u_1,2) = 1$.

(e) n_x is a short-hand notation for $\nu(R,x)$. For example, $n_5 = \nu(R,5) = 2$.

(f) δ_{ck} is a *difference function* used to quantify the ordinal difference between two observations $c,k \in R$. This difference is in turn interpreted as the level of disagreement the pair of observations (c,k) exhibits.

Krippendorff's alpha is defined as follows:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\text{Average } \delta_{ck} \text{ within all units}}{\text{Average } \delta_{ck} \text{ within all data}}, \tag{3.1}$$

where D_o is the *observed disagreement*; the average difference between pairs of observations within every unit, and D_e the *disagreement expected by chance*; the average difference between all pairs of observations within R (Krippendorff, 2004a, p. 223). The difference function δ_{ck} is given by:

$$\delta_{ck} = (\sum_{g=c}^{g=k} n_g - \frac{n_c + n_k}{2})^2. \tag{3.2}$$

Note that δ_{ck}, being a difference function, is by definition symmetric: for $c \neq k$, $\delta_{ck} \overset{!}{=} \delta_{kc}$. Likewise, $\delta_{ck} \overset{!}{=} 0$ for $c = k$.

The observed disagreement D_o is defined as:

$$D_o = \frac{1}{|R|} \sum_c \sum_k (o_{ck} \cdot \delta_{ck}), \tag{3.3}$$

where

$$o_{ck} = \sum_u \frac{\overbrace{\nu(u,c) \cdot \nu(u,k)}^{\text{Number of } (c,k) \text{ pairs in unit } u}}{|u| - 1}; \tag{3.4}$$

whereas the disagreement expected by chance D_e is determined as follows:

$$D_e = \frac{1}{|R|(|R|-1)} \sum_c \sum_k (n_c \cdot n_k \cdot \delta_{ck}). \tag{3.5}$$

1ST WORKED EXAMPLE: Two observers were requested to judge three stimuli using our rating scale; that is, using the ordinal values in $C = \{1,2,3,4,5 \mid 1 < 2 < 3 < 4 < 5\}$. Their responses are represented in the following 2 x 3 matrix:

$$A : \begin{pmatrix} 1 & 1 & 5 \\ 1 & 1 & 5 \end{pmatrix}. \\ u_1 \quad u_2 \quad u_3$$

α can be calculated by taking the following steps:

1. Determine $R = \{1,1,1,1,5,5\}$ as the multiset of all observations.
2. Define P as the set of all unique (c,k) pairs with $c,k \in R$. In this case, $P = \{(1,5),(5,1),(1,1),(5,5)\}$.
3. Calculate the difference δ_{ck} for every $(c,k) \in P$:

$$\begin{aligned} \delta_{15} &= (\sum_{g=1}^{g=5} n_g - \frac{n_1 + n_5}{2})^2 \\ &= (n_1 + n_5 - \frac{n_1 + n_5}{2})^2 \\ &= (\nu(R,1) + \nu(R,5) - \frac{\nu(R,1) + \nu(R,5)}{2})^2 \\ &= (4 + 2 - \frac{4+2}{2})^2 \\ &= 9 \\ &= \delta_{51}. \end{aligned}$$

$$\begin{aligned} \delta_{11} &= (n_1 - \frac{n_1 + n_1}{2})^2 \\ &= (\nu(R,1) - \frac{\nu(R,1) + \nu(R,1)}{2})^2 \\ &= (4 - \frac{4+4}{2})^2 \\ &= 0 \\ &= \delta_{55}. \end{aligned}$$

4. Calculate the disagreement expected by chance D_e:

$$
\begin{aligned}
D_e &= \frac{1}{|R|(|R|-1)} \sum_c \sum_k (n_c \cdot n_k \cdot \delta_{ck}) \\
D_e &= \frac{1}{6(6-1)} (\underbrace{4 \cdot 4 \cdot 0}_{(1,1)} + \underbrace{4 \cdot 2 \cdot 9}_{(1,5)} + \underbrace{2 \cdot 4 \cdot 9}_{(5,1)} + \underbrace{2 \cdot 2 \cdot 0}_{(5,5)}) \\
D_e &= 4.8.
\end{aligned}
$$

5. Calculate o_{ck} for every $(c,k) \in P$:

$$
\begin{aligned}
o_{15} &= \sum_u \frac{\nu(u,c) \cdot \nu(u,k)}{|u|-1} \\
&= \frac{\nu(u_1,1) \cdot \nu(u_1,5)}{|u_1|-1} + \frac{\nu(u_2,1) \cdot \nu(u_2,5)}{|u_2|-1} + \frac{\nu(u_3,1) \cdot \nu(u_3,5)}{|u_3|-1} \\
&= \frac{2 \cdot 0}{2-1} + \frac{2 \cdot 0}{2-1} + \frac{0 \cdot 2}{2-1} \\
&= 0.
\end{aligned}
$$

$$
\begin{aligned}
o_{51} &= \frac{\nu(u_1,5) \cdot \nu(u_1,1)}{|u_1|-1} + \frac{\nu(u_2,5) \cdot \nu(u_2,1)}{|u_2|-1} + \frac{\nu(u_3,5) \cdot \nu(u_3,1)}{|u_3|-1} \\
&= \frac{0 \cdot 2}{2-1} + \frac{0 \cdot 2}{2-1} + \frac{2 \cdot 0}{2-1} \\
&= 0.
\end{aligned}
$$

$$
\begin{aligned}
o_{11} &= \frac{\nu(u_1,1) \cdot \nu(u_1,1)}{|u_1|-1} + \frac{\nu(u_2,1) \cdot \nu(u_2,1)}{|u_2|-1} + \frac{\nu(u_3,1) \cdot \nu(u_3,1)}{|u_3|-1} \\
&= \frac{2 \cdot 2}{2-1} + \frac{2 \cdot 2}{2-1} + \frac{0 \cdot 0}{2-1} \\
&= 8.
\end{aligned}
$$

$$
\begin{aligned}
o_{55} &= \frac{\nu(u_1,5) \cdot \nu(u_1,5)}{|u_1|-1} + \frac{\nu(u_2,5) \cdot \nu(u_2,5)}{|u_2|-1} + \frac{\nu(u_3,5) \cdot \nu(u_3,5)}{|u_3|-1} \\
&= \frac{0 \cdot 0}{2-1} + \frac{0 \cdot 0}{2-1} + \frac{2 \cdot 2}{2-1} \\
&= 4.
\end{aligned}
$$

6. Calculate the observed disagreement D_o:

$$
\begin{aligned}
D_o &= \frac{1}{|R|} \sum_c \sum_k (o_{ck} \cdot \delta_{ck}) \\
&= \frac{1}{6} (\underbrace{8 \cdot 0}_{(1,1)} + \underbrace{0 \cdot 9}_{(1,5)} + \underbrace{0 \cdot 9}_{(5,1)} + \underbrace{4 \cdot 0}_{(5,5)}) \\
&= 0.
\end{aligned}
$$

7. Calculate α:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{0}{4.8} = 1.$$

The result is unsurprising, considering that A showed perfect agreement between observers. It is however important to note that $\alpha > 0$ because $D_o < D_e$; the observed *agreement* is greater than that expected by chance. Further, it is worthwhile to mention that α can also take negative values, specifically when $D_o > D_e$. Negative values are not used to infer reliability and can be interpreted as indicative of sampling errors or systematic disagreements; that is, observers agreeing to disagree (Krippendorff, 2004a, p. 222).

The above example demonstrated how α can be computed by hand. However, equations 3.2, 3.3 and 3.5 involve a considerable amount of unnecessary calculations caused mainly by the symmetric tuples in P. In the coming example, the following, more computationally efficient versions of δ_{ck}, D_o and D_e are used:

$$\delta_{ck} = (\frac{n_c}{2} + \sum_{g>c}^{g<k} n_g + \frac{n_k}{2})^2, \text{ where } k > c; \tag{3.6}$$

$$D_o = \sum_c \sum_{k>c} o_{ck} \cdot \delta_{ck}; \tag{3.7}$$

$$D_e = \frac{1}{|R|-1} \sum_c \sum_{k>c} (n_c \cdot n_k \cdot \delta_{ck}). \tag{3.8}$$

Equation 3.6 was obtained from Krippendorff (2004a, p. 233).

2ND WORKED EXAMPLE: Three observers were requested to judge three stimuli using our rating scale. Their responses are represented in the following 3 x 3 matrix, where columns represent stimuli and rows represent observers:

$$B : \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

α is determined as follows:

$$\begin{aligned} \alpha &= 1 - \frac{D_o}{D_e} \\ &= 1 - \frac{\sum_c \sum_{k>c} o_{ck} \cdot \delta_{ck}}{\frac{1}{|R|-1} \sum_c \sum_{k>c} (n_c \cdot n_k \cdot \delta_{ck})} \end{aligned}$$

$$
\begin{aligned}
&= 1 - \frac{o_{13} \cdot \delta_{13}}{\frac{1}{|R|-1} \cdot n_1 \cdot n_3 \cdot \delta_{13}} \\
&= 1 - \frac{(\frac{2 \cdot 1}{2}) \cdot (\frac{8}{2} + \frac{1}{2})^2}{\frac{1}{|9|-1} \cdot (8) \cdot (1) \cdot (\frac{8}{2} + \frac{1}{2})^2} \\
&= 1 - \frac{20.25}{20.25} \\
&= 0.
\end{aligned}
$$

Despite B having only one disagreeing observation, $\alpha = 0$ because $D_o = D_e$. Such peculiar results are often considered instances of the paradox problem of Feinstein and Cicchetti (1990), where chance-corrected metrics return a low value despite high levels of agreement. Krippendorff attributes this mainly to the lack of variance in the set of observations (Krippendorff, 2004a, pp. 236-237). The interested reader is referred to Hayes and Krippendorff (2007), Krippendorff (2011) and Krippendorff (2004a, pp. 211-243) for more in-depth information about this measure.

3.3.2 *Answering the data question*

The collected dataset (as represented in the heatmap in Figure 3.3) resulted[2] in an alpha of 0.644, indicating unreliable data. The value suggests a low level of consistent agreement among observers and the likely irreproducibility of the dataset if our 16 observers are replaced with others. The main reason behind this low alpha are the idiosyncratic responses given by expert observer p_{02}. As demonstrated in Figure 3.6, the observer can be seen choosing an answer that contradicts the responses of every other participant for 15 stimuli. Since alpha employs a weighting scheme that takes into account the disagreement magnitude of each pair of observations, the disagreements exhibited by p_{02} are given more weight, and consequently influence alpha's value more than others. Secondly—and more importantly—the fact that not one single participant agrees with these responses is further punished by the statistic as a strong indication of their irreproducibility. This particularly affects D_e, which heavily depends on the frequency of occurrence of each of the values in a pair of observations across observers. When considered together, both of these factors cause D_e and D_o to converge[3], as can be seen from their mathematical definitions in the previous subsection. If we were to hypothetically exclude this one participant's data from the analysis, alpha would increase to 0.78.

2 Krippendorff's alpha was calculated using a verified MATLAB function (Eggink, 2022).

3 This effect can also be seen to a more extreme extent in the second worked example.

This value would indicate tentative reliability and deviate only by 2% from the reliability threshold suggested by Krippendorff.

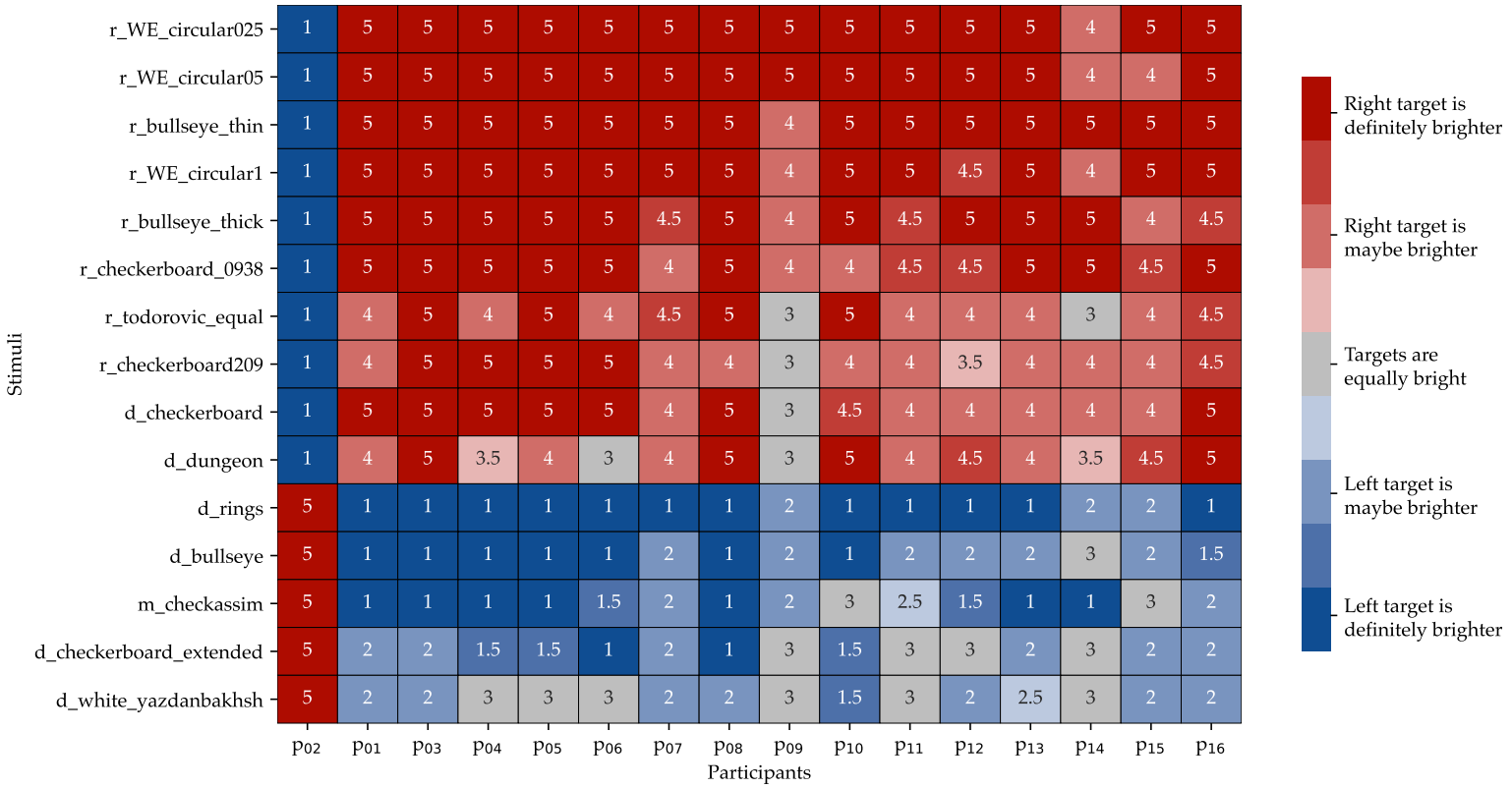

Figure 3.6: Snippet of the heatmap previously shown in Figure 3.3 demonstrating the stark disagreement in brightness judgments exhibited by expert participant p_{02}. The internal coding of responses is displayed in each cell. The responses of p_{02} are shifted to the first column. Alpha's value is mainly affected by two factors: the difference between observations in each observation pair and their frequency of occurrence across observers. Considering the pair $(1, 5)$ as an example, the difference δ_{15} is greater than that of any other pair. Further, the values 1 and 5 occur in extremely imbalanced ratios, such as $1 : 14$ in "r_WE_circular025". This behavior is consequently punished by alpha.

Alpha was calculated for the dataset in its entirety rather than for individual stimuli because of the aforementioned limitation of low variance. The stimuli were not categorized into smaller groups and tested for reliability since there is no commonly agreed upon basis for doing so. Nonetheless, a large subset of our stimuli consists of variations of White's effect. This subset, which includes all 15 White stimuli, resulted in an alpha of 0.691, exceeding the minimum threshold and indicating tentative reliability.

4 DISCUSSION

Quantitative experimental methods, such as matching, are limited in scope in the sense of being usually conducted using only a small number of stimuli and observers. Consequently, the directions of some known brightness effects cannot be reliably determined, as it is difficult to compare results across different studies that are often conducted under different experimental conditions using different stimuli and different participants. The goal of the present work was to firstly produce data on a large number of brightness stimuli, and secondly to investigate inter-individual differences in the perception of the direction of the induced effects. To accomplish this, a psychophysical experiment was conducted using 16 observers and 45 stimuli depicting a variety of brightness effects. Observers judged the direction of effect using a five-point rating scale. The results reveal that the stimuli varied in terms of the direction and the certainty of the effects they produce across observers. Further, some participants exhibited similar response patterns across the stimulus set. As a measure of the consistent agreement achieved among observers, Krippendorff's alpha was calculated to be 0.644, indicating low inter-observer reliability.

4.1 CONCORDANCE WITH OTHER STUDIES

Table 4.1 compares the observed directions of effects as perceived by our participants with those reported[1] in Murray (2020), Domijan (2015) and Robinson et al. (2007). Following previous usage (as in Figure 1.2), the symbols $>, <$ and $=$ refer to the direction of effect. All observed directions of effects are in line with those reported in the three studies apart from the snake control figure, the cube illusion and one version of the Todorović-Benary effect, where our participants on average perceived the targets as equally bright. This indicates that most human observers, with some exceptions, exhibit similar visual behavior and tend to generally agree on the direction of most brightness effects.

1 These are the directions against which the authors tested the performance of their predictive computational models. Robinson et al. (2007) do not explicitly state these in their paper, therefore, the comparison for the stimuli drawn from Robinson et al. (2007) is based on the directions reported in the original publications that first introduced these stimuli.

			Direction of effect		
Stimulus	Id	Complete name	In agreement	Observed	Reported
m_snake_control	1.1	Snake control figure	×	=	>
m_checkassim	1.2	Checkerboard assimilation	✓	>	>
m_simcon	1.3	Classic simultaneous contrast figure	✓	>	>
m_white	1.4	White's illusion	✓	>	>
m_koffka_adelson	1.5	Koffka-Adelson figure	✓	>	>
m_koffka_broken	1.6	Koffka ring, broken	✓	>	>
m_koffka_connected	1.7	Koffka ring, connected	✓	=	=
d_dungeon	2.1	Dungeon illusion	✓	<	<
d_cube	2.2	Cube illusion	×	=	<
d_grating	2.3	Grating illusion	✓	<	<
d_rings	2.4	Ring patterns	✓	>	>
d_bullseye	2.5	Bullseye display	✓	>	>
d_sbc	2.6	Contrast-contrast effect	✓	<	<
d_white	2.7	White's effect	✓	>	>
d_benary	2.8	Benary's cross	✓	<	<
d_todorovic	2.9	Todorović's illusion	✓	>	>
d_checkerboard	2.10	Checkerboard contrast	✓	<	<
d_checkerboard_extended	2.11	Checkerboard contrast extended	✓	>	>
d_white_yazdanbakhsh	2.12	White's effect-Yazdanbakhsh	✓	>	>
r_WE_thick	3.1	White's effect-thick	✓	>	>
r_WE_thin_wide	3.2	White's effect-thick-wide	✓	>	>
r_WE_anderson	3.3	White's effect-Anderson	✓	<	<
r_WE_howe	3.4	White's effect-Howe	✓	<	<
r_WE_zigzag	3.5	White's effect-zigzag	✓	>	>
r_WE_radial_thick_small	3.6	White's effect-radial-thick-small	✓	<	<
r_WE_radial_thick	3.7	White's effect-radial-thick	✓	<	<
r_WE_radial_thin_small	3.8	White's effect-radial-thin-small	✓	<	<
r_WE_radial_thin	3.9	White's effect-radial-thin	✓	<	<
r_WE_circular1	3.10	White's effect-circular-1	✓	<	<
r_WE_circular05	3.11	White's effect-circularo.5	✓	<	<
r_WE_circular025	3.12	White's effect-circularo.25	✓	<	<
r_sbc_large	3.13	Simultaneous brightness contrast-large	✓	>	>
r_sbc_small	3.14	Simultaneous brightness contrast-small	✓	>	>
r_todorovic_equal	3.15	Todorovic-equal	✓	<	<
r_todorovic_in_large	3.16	Todorović-in-large	✓	<	<
r_todorovic_in_small	3.17	Todorović-in-small	✓	<	<
r_todorovic_out	3.18	Todorović-out	✓	<	<
r_checkerboard_016	3.19	Checkerboard-0.16	✓	<	<
r_checkerboard209	3.20	Checkerboard-209	✓	<	<
r_checkerboard_0938	3.21	Checkerboard-0.94	✓	<	<
r_benary_cross	3.22	Benary cross	✓	>	>
r_todorovic_benary1_2	3.23	Todorović-Benary 1–2	×	=	<
r_todorovic_benary3_4	3.24	Todorović-Benary 3–4	✓	>	>
r_bullseye_thin	3.25	Bullseye-thin	✓	<	<
r_bullseye_thick	3.26	Bullseye-thick	✓	<	<

Table 4.1: Comparison of observed and reported directions of effects.

4.2 ADDRESSING THE LOW ALPHA

The set of stimuli varied to different extents across observers in terms of both the direction of effect and the confidence with which it was perceived. The variability in the confidence the observers had in their perceptual judgments is an expected but pertinent result, since brightness stimuli are known to vary in the strength of the effects they induce. On the other hand, the disagreement among observers on the direction of effect was consistent in the sense of being predominantly due to perceiving the targets of the stimuli as equally bright. That is, our participants rarely perceived the effects of the brightness stimuli in opposite directions. An exception to this is expert participant p_{02} whose responses influenced the result of the reliability analysis the most.

The responses given by expert participant p_{02} may call into question the appropriateness of using Krippendorff's alpha. The low value of 0.644 indicates the likely irreproducibility of the dataset, which is reasonable considering the peculiarity of some of the responses of p_{02} compared to those of all other participants. However, this low alpha could also be seen as contradictory to the results, since the (average) directions of effects as perceived by our participants are almost completely in agreement with those reported previously in other papers. In that sense, it could be argued that alpha is too "strict" in its treatment of outlying observations. Nevertheless, one fitting interpretation of this low value could be in this case that the brightness perception of expert participant p_{02}, whose responses were severely punished by the statistic, is likely to be unique.

4.3 LIMITATIONS

Due to a software error, stimuli 1.1, 3.6, 3.7, 3.8 and 3.9[2] were flipped vertically instead of horizontally. Consequently, the "flipped" versions of these stimuli showed no substantial difference from the original, as the placement and the surround of the left and right targets remained unchanged. This oversight went unnoticed and was only detected by the experimenter once the first two observers had already completed the experiment.

Excluding the author (p_{05}), only three participants (p_{01}, p_{06} and p_{07}) answered the last two catch trials correctly. All the other observers incorrectly responded with "the targets are equally bright". This could indicate that they may not have paid enough attention in later trials, or that the difference in luminance was too subtle for them to detect. Given that the observers did not spend more than four seconds on each

2 These were the stimuli that needed to be rotated. The mistake was flipping them first and then rotating them, when they should have been rotated first and then flipped.

trial, answering incorrectly is likely indicative of cognitive laziness, which could have skewed some results in the final trials.

4.4 CONCLUSIONS

The present work provided a so-far lacking overview of the directions of a large set of brightness stimuli as perceived by human observers. In spite of evident inter-individual differences, comparisons with other studies indicate that most of these brightness stimuli are on average perceived in the same direction by most human observers. The directions of effects as reported in this work could serve as a baseline against which modeling results can be reliably compared. Human brightness perception was found to vary inter-individually in terms of the direction of effect and the confidence with which it is perceived. This variability itself varies in turn depending on the brightness stimulus. Nevertheless, response patterns were identified among observers, suggesting that groups of people might share the same visual behavior. The low inter-observer reliability estimate indicates that our dataset is unlikely to be reproduced based on the responses of one participant, which suggests that brightness perception could be, in some cases, markedly individualistic.

REFERENCES

Abebe, M. A., Pouli, T., Larabi, M.-C., and Reinhard, E. (2017). Perceptual lightness modeling for high-dynamic-range imaging. *ACM Transactions on Applied Perception (TAP)*, *15*(1), 1–19. doi: 10.1145/3086577

Adelson, E. H. (1995). Checkershadow illusion. Retrieved from http://persci.mit.edu/gallery/checkershadow

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

Agostini, T., and Galmonte, A. (2002). Perceptual organization overcomes the effects of local surround in determining simultaneous lightness contrast. *Psychological Science*, *13*(1), 89–93. doi: 10.1111/1467-9280.00417

Aguilar, G., and Maertens, M. (2020). Toward reliable measurements of perceptual scales in multiple contexts. *Journal of Vision*, *20*(4), 19. doi: 10.1167/jov.20.4.19

Blakeslee, B., and McCourt, M. E. (2015a). Comments and responses to "theoretical approaches to lightness and perception". *Perception*, *44*(4), 359–367. doi: 10.1068/p4404re

Blakeslee, B., and McCourt, M. E. (2015b). What visual illusions tell us about underlying neural mechanisms and observer strategies for tackling the inverse problem of achromatic perception. *Frontiers in Human Neuroscience*, *9*. doi: 10.3389/fnhum.2015.00205

Blakeslee, B., Reetz, D., and McCourt, M. E. (2008). Coming to terms with lightness and brightness: Effects of stimulus configuration and instructions on brightness and lightness judgments. *Journal of Vision*, *8*(11), 3–3. doi: 10.1167/8.11.3

Cunningham, D., and Wallraven, C. (2011). *Experimental design: From user studies to psychophysics*. CRC Press.

Domijan, D. (2015). A neurocomputational account of the role of contour facilitation in brightness perception. *Frontiers in human neuroscience*, *9*, 93. doi: 10.3389/fnhum.2015.00093

du Buf, H. (2001). Modeling brightness perception. In *Vision models and applications to image and video processing* (pp. 21–36). Springer US. doi: 10.1007/978-1-4757-3411-9_2

Eggink, J. (2022). Krippendorff's alpha. MATLAB Central File Exchange. Retrieved from https://www.mathworks.com/matlabcentral/fileexchange/36016-krippendorff-s-alpha

Feinstein, A. R., and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical*

Epidemiology, *43*(6), 543–549. doi: 10.1016/0895-4356(90)90158-l

Gwet, K. (2021). *Handbook of inter-rater reliability: Volume 1: Analysis of categorical ratings*. Advanced Analytics, LLC.

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. doi: 10.1080/19312450709336664

Hurvich, L., and Jameson, D. (1966). *The perception of brightness and darkness*. Allyn and Bacon.

Kingdom, F., and Prins, N. (2016). *Psychophysics: A practical introduction*. Elsevier Science.

Krippendorff, K. (2004a). *Content analysis: An introduction to its methodology*. Sage Publications.

Krippendorff, K. (2004b). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433. doi: 10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. Retrieved from `https://repository.upenn.edu/asc_papers/43/`

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

Murray, R. F. (2020). A model of lightness perception guided by probabilistic assumptions about lighting and reflectance. *Journal of Vision*, *20*(7), 28–28. doi: 10.1167/jov.20.7.28

Robinson, A. E., Hammon, P. S., and de Sa, V. R. (2007). Explaining brightness illusions using spatial filtering and local response normalization. *Vision Research*, *47*(12), 1631–1644. doi: 10.1016/j.visres.2007.02.017

White, M. (1979). A new effect of pattern on perceived lightness. *Perception*, *8*(4), 413–416. doi: 10.1068/p080413