Improving Trial Selection for Maximum Likelihood Conjoint Measurement for Psychophysical Experiments

Abschlussarbeit
vorgelegt von

Rhea M. Widmer
Matrikelnummer: 0383316

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)
im Fach Wirtschaftsinformatik

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

Betreuer: Dr. Joris Vincent

Erstgutachter: Prof. Dr. M. Maertens
Zweitgutachter: Prof. Dr. G. Gallego

Berlin, 23. Januar 2025

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, January 23, 2025

Rhea M. Widmer

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit eigenständig ohne Hilfe Dritter und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe. Alle Stellen die den benutzten Quellen und Hilfsmitteln unverändert oder sinngemäß entnommen sind, habe ich als solche kenntlich gemacht.

Sofern generative KI-Tools verwendet wurden, habe ich Produktnamen, Hersteller, die jeweils verwendete Softwareversion und die jeweiligen Einsatzzwecke (z.B. sprachliche Überprüfung und Verbesserung der Texte, systematische Recherche) benannt. Ich verantworte die Auswahl, die Übernahme und sämtliche Ergebnisse des von mir verwendeten KI-generierten Outputs vollumfänglich selbst.

Die Satzung zur Sicherung guter wissenschaftlicher Praxis an der TU Berlin vom 8. März 2017 habe ich zur Kenntnis genommen.

Ich erkläre weiterhin, dass ich die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt habe.

Berlin, January 23, 2025

Rhea M. Widmer

Contents

List of Tables

List of Figures

1 Abstract

This thesis explores strategies to reduce the number of trials in psychophysical experiments using Maximum Likelihood Conjoint Measurement (MLCM) while maintaining the accuracy of estimated perceptual brightness scales. An experiment applied two trial reduction strategies to White's Illusion: the *a priori* (Static) Sampling Strategy, which excludes predictable trials based on prior knowledge, and the Runtime (Dynamic) Sampling Strategy, which dynamically eliminates deterministic trials during data collection. Experimental data from 10 participants was collected and used to estimate perceptual scales under three conditions: full trials (control), *a priori* (Static) Sampling, and Runtime (Dynamic) Sampling. The scales from reduced trials were compared to the control scales using Root Mean Squared Error (RMSE) as a metric. Results show that both strategies produce reliable estimates, with the Runtime (Dynamic) approach yielding more accurate scales on average. This shows the potential of trial reduction strategies to optimize experimental efficiency without significantly sacrificing data quality.

2 Zusammenfassung

In dieser Arbeit werden Strategien zur Steigerung der Effizenz in psychophysikalischen Experimenten mit Maximum Likelihood Conjoint Measurement (MLCM) und deren Einfluss auf die Genauigkeit der daraus resultierenden geschätzten Wahrnehmungsskalen untersucht. In einem Experiment, bei dem entschieden werden muss, welcher von zwei Stimuli bei *White's Illusion* heller wirkt, wurden zwei Strategien zur Reduzierung von Versuchen auf die *White's Illusion* angewandt: die statische *a priori*-Sampling-Strategie, die vorhersehbare Versuche aufgrund von Vorwissen ausschließt, und die dynamische Laufzeit-Sampling-Strategie, die vorhersehbare Versuche während der Datenerhebung dynamisch eliminiert. Es wurden experimentelle Daten von 10 Teilnehmer*innen gesammelt und zur Schätzung der Wahrnehmungsskalen unter drei Bedingungen verwendet: vollständigs Sampling (Kontrollgruppe), statisches *a priori*-Sampling und dynamisches Laufzeit-Sampling. Die Skalen aus den reduzierten Versuchen wurden mit den Kontrollskalen unter Verwendung des Root Mean Squared Error (RMSE) als Metrik verglichen. Die Ergebnisse zeigen, dass beide Reduktions-Strategien zuverlässige Schätzungen liefern, wobei der dynamische Laufzeit-Samplingansatz im Durchschnitt genauere Skalen liefert. Dies zeigt das Potenzial von Strategien zur Versuchsreduzierung, die die experimentelle Effizienz steigern, ohne die Datenqualität wesentlich zu beeinträchtigen.

3 Introduction

Brightness perception is a field within visual sciences focused on measuring, explaining, and quantifying how we perceive brightness. It has two facets to it: on the one hand, there is the physical properties, most importantly luminance, of an object which can be measured photometrically, on the other hand there is the individual perception that one experiences when this physical occurrence is observed by the human eye and processed by the human brain. Luminance is a photometric measure of the luminous intensity per unit area of light traveling in a given direction. It describes the amount of light that passes through, is emitted from, or is reflected from a particular area. Luminance is influenced by how much light is falling on the surface of the object (illuminance) and the percentage of light reflected back by the object, coined by material properties (reflectance). The remaining light being reflected into our vision is luminance (Kingdom, 2014). Unlike luminance, brightness is not a measurable physical quantity. Brightness, which is defined as perceived luminance, is subjective and is not only influenced by the luminance emitted by the object.

This is where context and contrast come into play. According to Kingdom (Kingdom, 2014), context and especially contrast are a major factor on how bright humans perceive an object to be. The influence it has on the brightness perceived by us can be seen in visual illusions such as the Simultaneous Contrast Illusion (Adelson, 2000). Two equi-

Figure 1: Simultaneous Brightness Contrast Illusion: Even though both gray inner squares embedded in the bigger, outer squares have the same luminance, the left one is usually perceived as more bright than the right one due to the difference in background luminance. This illusion is created by the difference in contrast (Adelson, 2000).

luminant patches are being displayed next to each other in different contexts: on the left side, the patch is placed on a dark gray background, on the right side on a light gray one. The left patch has higher luminance than its background and vice versa. The difference of the luminance of their surrounding area leads the observer to perceive the left patch as brighter than the right one despite their physical equiluminance. Another example of this phenomenon can be found in White's Illusion (Vincent et al., 2023). In this illusion, two gray patches with identical luminance placed on alternating black and white bars are perceived differently due to deviation in context, i. e. the background they are displayed on (Adelson, 2000).

It is intuitive that there is some kind of link between luminance and the brightness perceived: the higher the luminance of an object is, the brighter it will be perceived. The exact relation they have to each other is not known though. So there seems to be

Figure 2: White's Illusion: The left target appears more bright than the right one, despite both having identical luminance. The difference in perception arises because the targets are presented in different contexts: the left one is on black background in the black and white grid, and the right one is on white background (Vincent et al., 2023).

some process of combining the physical incidence of light with the context, which is then encoded into a brightness perception. This encoding process, turning the physical stimulus observed into our perception of it, can be mathematically described in a perceptual encoding function, which displays brightness as a function of luminance. These perceptual encoding functions differ from person to person and are also dependent on the surroundings. An example of two encoding functions can be seen in Figure 3.

The illusions introduced before are examples in which the brightness perceived is not only dependent on the luminance of the object, but by the context in which the target is embedded as well. This leads to different perceptual encoding functions for different contexts. In Figure 3, one sees the discrepancy of brightness perceived when one is being presented equiluminant gray targets in different context can be modeled as two separate encoding functions. The target placed on a black background is perceived as more or equally bright than the target in a white setting consistently over luminance levels. This divergence is illustrated by the vertical line connecting circular markers on both functions in the graph.

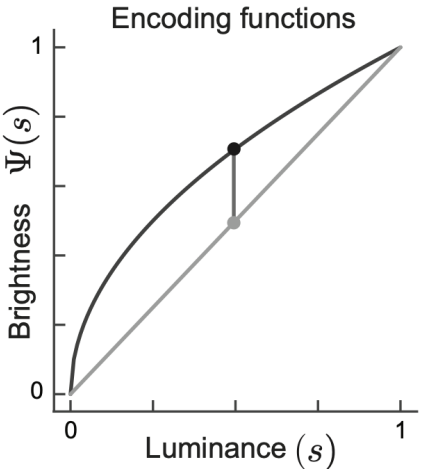

Figure 3: One possible relation of the physical luminance of a target (x-axis) and the perceived brightness observed as a function of it (y-axis). White's Illusion seen in Figure 2 (p. 9) creates a brightness difference between equiluminant gray patches placed in the black and white phase of a square-wave grating. The gray graph is the brightness perceived of a target set on a white bar, while the black graph is the brightness perceived of a target of equal luminance but on a black bar. The target on white is consistently perceived as less bright than a target on black with the same luminance. This difference is illustrated by the vertical line connecting circular markers on both functions (Vincent et al., 2023).

We do not have a way to directly measure such brightness encoding functions. The next best option is to use psychophysical experiments and interpret the outcomes. Manipulating the physical variables, luminance and context, of the illusion and observing the perceived brightness of the participants, we hope to be able to make conclusions concerning this mechanism. Vincent et al. (2023) conducted an experiment in order to estimate perceptual encoding functions where White's Illusion was shown to participants with varying target luminance and context (placement on a black or white bar).

The outcome of the experiment was processed and used to estimate perceptual scales. The estimated perceptual scales that were constructed based on the experimental data can be seen in Figure 4. Each panel contains two scales, one for each context (black or white). The scales are monotonically increasing, showing the link of higher luminance leading to higher perceived brightness. The black context scale values are higher than the white context ones for almost all data points, which shows the influence of context on the perception. This shows the effect of White's Illusion. The targets of equal luminance are consistently perceived as having the same or higher brightness on a black background than the ones on white background. The perceptual scales do not only depend on luminance and context, but are also different for each individual.

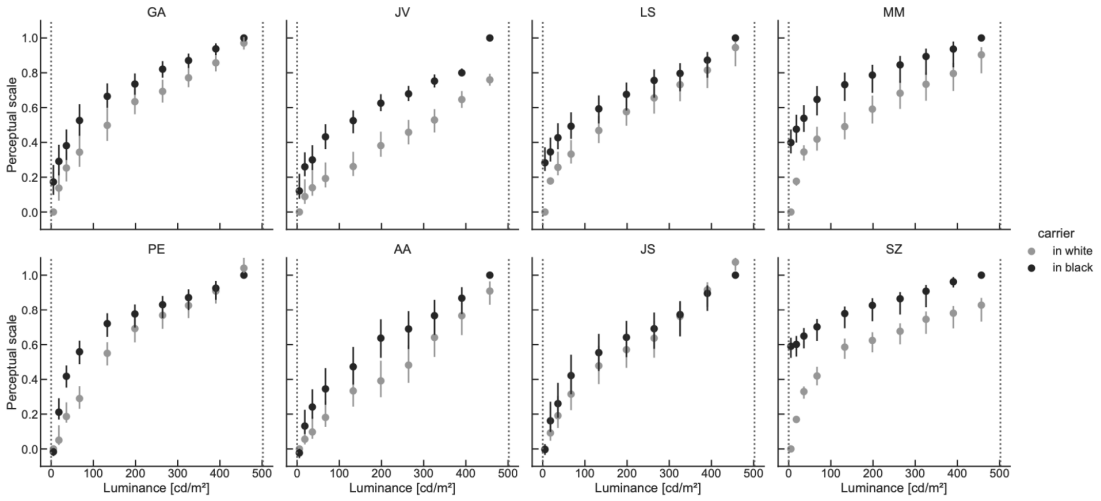

Figure 4: Perceptual scales estimated using data from an experiment conducted by Vincent et al. (2023). Each panel contains the data of one participant, in total 8. As in Figure 3, gray dots represent of targets on white background and the black dots are on black background. The error bars are based on 95% confidence intervals. For comparability, the scales are normalized, anchoring the lowest luminance target value of the white context scale to 0 and dividing the scale values by the maximum value per participant. This results in all scale values being between 0 and 1 for all participants.

In order to produce such an estimated encoding function, a statistical scaling method is required which estimates perceptual encoding functions from the experimental deci-

sion data. The method used by Vincent et al. (2023) is Maximum Likelihood Conjoint Measurement (MLCM), developed by Ho et al. (2008) and Knoblauch and Maloney (2012). MLCM uses the relative frequencies of choice and estimates the scales so that the likelihood of having this experimental data as an outcome is maximized. The data used to estimate the scales was obtained in a human experiment with eight participants, four naive and four expert ones (Vincent et al., 2023). The participants were being shown multiple versions of White's Illusion, all with different target luminance difference and context (e.g. both targets "on black" or one target "on white" and one "on black"). There were ten target luminance levels being used. The stimuli were presented and the participant had to decide which one of the targets is perceived more bright by them using a forced choice experiment setting. Each stimulus, containing two targets, was presented 15 times to the participant in randomized order of the trials.

The brightness judgments of each participant can also be visualized as choice probabilities. The heat map in Figure 5 (p. 13) shows an overview of these relative frequencies of one participant. In the cells with relative frequency of 1 or 0, the participant has consistently judged the left target as brighter over all trials or vice versa. This implies that there is no room for doubt which target is brighter and thus renders them more or less deterministic. The trials where the participants did not manage to decide consistently are the ones which lie between 1 and 0. These seem to be difficult to decide on, meaning that the perceived brightness of the two targets is on a very similar level. Almost all values that are not 1 or 0, are being found in the upper right quadrant, where the two targets are displayed in different context. In most of the heat maps produced by the experiment, a lot of cell values are either 1 or 0, as seen in Figure 5. This is especially the case for stimulus pairs with high difference in luminance. An extreme example would be one white target and one black target, both on the same background. Here, basically all participants would judge the white target as brighter than the black target, while the context doesn't interfere with perception as the background is the same.

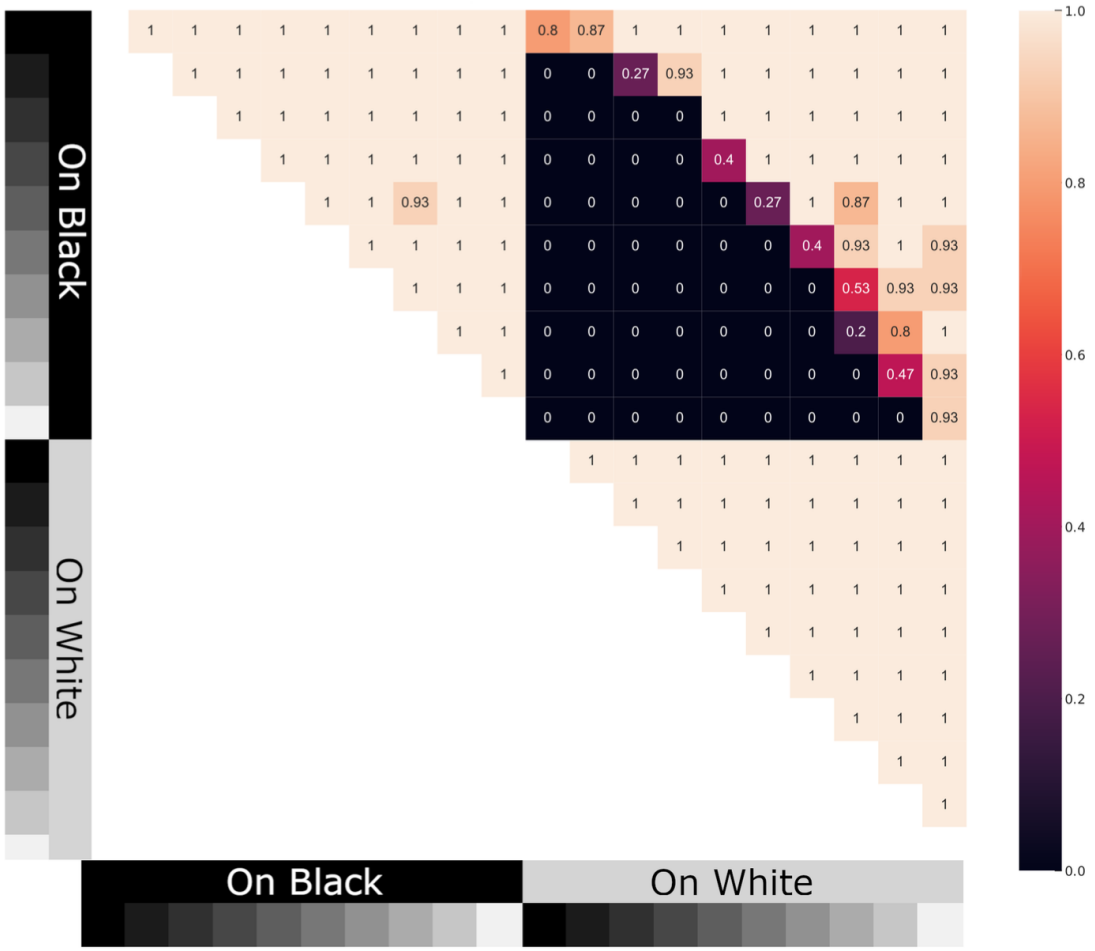

Figure 5: Heat map showing the relative frequency of responses for White's Illusion trials for one participant. The x- and y-axes represent the different contexts and luminance levels, with the stimuli on these axes being compared against each other. Each cell value indicates relative choice frequency of the participant choosing one stimulus over another across 15 repeated trials, totaling 2850 trials. The color scale corresponds to relative choice frequency, the extremes being white, equaling 100% and black 0% (Zabel, 2023).

This observation leads to the question, whether these deterministic trials carry information relevant for the scale estimation at all. Leaving these ”easy” trials out could optimize resource usage and thus enable to obtain more data in the same experimental time. In the original experiment, as conducted by Vincent et al. (2023), the number of trials is very high because of the statistical scaling method used. MLCM, as developed by Knoblauch and Maloney (2012), depends on all possible decision data as input. As mentioned before, MLCM usually requires all possible stimulus pairings being shown, and it is unclear if MLCM will still produces reliable outcome when using data from a reduced set of trials. Our task is to determine whether the outcome of MLCM is reliable enough to use for research when applied to data that does not include all possible trials. To do so, we propose to reduce the experiment to the “hard” trials, which carry actual informational value, since the “easy” trials are predictable and thus trivial.

Jan Zabel's thesis (Zabel, 2023) made a first step in this direction. In a simulation, he conducted a simulated experiment in order to develop trial selection strategies for data acquisition optimization using MLCM. Zabel used a pair of perceptual encoding functions depending on the luminance level for the two contexts (on black and on white) as the ground truth function that he built the simulation upon. In this setup, a simulated observer replaced a human participant and the “human factor” is emulated by a Gaussian noise level which brings a slight randomness into the simulation. Generated unique trials showing White's Illusion with varying luminance levels and context were created, using 10 luminance levels and two contexts (on black and on white), just as in Vincent et al., 2023. Based on the ground truth functions and a noise level, Zabel presented the trials to the simulated observer.

The decisions retrieved in this simulation were processed like actual data from a human experiment and fed into MLCM in order to obtain the estimated perceptual scales. Finally, the ground truth functions were compared to the estimates resulting from the simulated experiment, indicating if decision making has changed due to the reduced number of trials.

To reduce trial size, two sampling strategies were developed:

1. The *a priori* (Static) Sampling Strategy uses information previously acquired in experiments to filter the trials with low information density. Based on the experience resulting from previous experiments, trials which seem to have been “easy” decisions for the previous participants are excluded. In Figure 5 (p. 13), one can see that the participant was very consistent in trials where the targets were displayed in the same contexts, i. e. both on black or both on white and in the trials which are in different contexts but with big luminance difference. Zabel (2023) used a luminance difference of >20% for for the same context and >50% for different contexts.

2. The Runtime (Dynamic) Sampling Strategy excludes “easy” trials in the runtime of the experiment. This is implemented by splitting the experiment in two phases: The initial stage and the sampling stage. In the initial stage, the full set of trials is presented, reducing the number of repetitions from 15 to 7. The trials which did

not have a choice probability of 100% or 0% are excluded and the reduced set of trials is presented to the simulated observer for the remaining 8 repetitions in the following sample stage.

The Runtime Sampling Strategy's effectiveness, even without replacing removed trials with predictable results, was unexpected, as MLCM remained unaffected by missing trials. The *a priori* Sampling Strategy, using previously collected data to exclude quasi-deterministic trials, was anticipated to affect accuracy and precision and performed as expected. Therefore, simulation results indicated that the number of unique trials in an experiment can be reduced without significantly affecting the quality of the estimated perceptual scales or with less than a 10% reduction in accuracy in comparison to the ground truth functions.

The outcome of Zabel's simulation raises the question if the results are transferable to experiments with human participants. This is the research question of this thesis: Can Zabel's results be reproduced in experiments with human participants? To answer this question, I conducted an experiment using the trial selection strategies developed by Zabel and compare the estimates with the perceptual encoding functions deducted from data using the full set of trials.

To minimize the influence of the human factor on the reliability of our results, it is important to put thought into a good experimental build and ensure high internal and external validity. Additionally, we need a way to replace the ground truth function used in Zabel's work, since the true perceptual function is the very thing we are trying to estimate and thus is not accessible. Without it, we lack a measure to asses the quality of our estimates produced with the reduced sets of trials.

For comparability, each participant will have to conduct the experiment once with the full, original set of trials to estimate a ground truth function and once with the reduced set of trials of each sampling strategy. These experiment variants will be implemented in separate experimental sessions. The order of these sessions is a variable which might influence the perceptual scales that are being put out by MLCM. This brings order effects into play since it might influence the outcome whether the full set of trials is displayed first and the reduced set after or vice versa and thus will have to be taken into consideration in our experimental build. Order effects refer to the influences on participants' responses that result from the order in which the stimuli are presented. There are several effects which become relevant to us as soon as we see order effects. One possible consequence are practice effects since participants might have more consistent decision making once they are familiar with the experiment. For us, this could mean that a participant which worked on one modus first would be more consistent in the second session because they got used to the experiment. This undermines comparability of the sessions. Another possibility is the occurrence of fatigue effects of which we are talking about if participants perform worse on tasks as they become tired or bored over time. Since the time frame of the experiment will be extended by adding the reduced set of trials to the full set of trials, the probability of the occurrence of fatigue effects rises. Also, they might have an easier time deciding on the reduced set of trials because of the lower number of decisions they have to take. One more problem that comes to mind

is the impact that using only hard trials might have on the participants: They might become fatigued more quickly and motivation might drop when only using "hard" trials. This leads to a reasonable doubt concerning the independence of the trials. The first trials presented might be decided on using full capacity of concentration while the later ones are more randomly decided out of fatigue or frustration.

The goal of this thesis was to find out how reliable the data of the human experiments actually is after trial amount reduction and which methods provide the best results. Since the perceptual encoding functions are individual to each participant, my assumption was that the Runtime (Dynamic) Condition will produce results closer to the full data estimations.

4 Methods

My goal was to test if the two sampling strategies introduced by (Zabel, 2023) produce reliable results in human experiments. The strategies' results are being compared to the estimates produced by the full set of trials as a control condition. Furthermore, it is to be determined which of the strategies performs better, i. e. produces an estimate closer to the one deducted from the full set of trials.

4.1 Sampling Strategies

The *a priori* (Static) Sampling Strategy selects trials which are not presumed to carry relevant information before any trials are presented to a participant or simulated. This is done by Zabel by comparing luminance levels and omitting trials with significant luminance differences. It is expected that participants will largely agree on trials with large luminance differences within the same context (e.g., both targets are "on white" or "on black"). A luminance difference of more than 20% within the same context is used as the threshold for showing trials. For different contexts (one target "on white" and one "on black"), a threshold of more than 50% is applied.

The other option is the Runtime (Dynamic) Sampling Strategy: This strategy samples during the presentation of trials to a participant or during simulation, sorting out the "easy" trials so only the "hard" trials with actual informational value remain for each individual participant. This process is divided into two stages: the initial stage and the sample stage.

In the initial stage, all trials are shown for half of the 14 total repetitions regardless of luminance differences. This stage provides an initial impression and helps identify trials with high agreement results (0% or 100%), which are not repeated. In the sample stage, trials with low agreement results (>0% and <100%) are shown an additional seven times.

In Zabel's thesis, the reduction methods' results were used to estimate perceptual encoding functions using MLCM and compared to the ground truth perception functions that were used as the simulation base (Zabel, 2023). Since we also need such a ground truth to compare our outcome to, the experiment will not only be conducted with the two reduced trial methods but also in the original way used by Vincent et al. (2023). The full set of trials is presented to the participants as well, totaling 14 repetitions as well as the reduced trials. This results in three conditions of the experiment:

1. Runtime (Dynamic) Sampling reduced condition,
2. *a priori* (Static) Sampling reduced condition,
3. control condition showing the full set of trials.

The set of trials of each condition is being repeated for 14 times, the dynamically reduced one uses the full set of trials 7 times and the reduced set based on the first 7 repetitions for another 7 repetitions. The sets of trials of the other two conditions remain the same over all repetitions.

As for how the different conditions are spread in between participants, two options were considered.

1. One participant group works only with the full set of trials and one group that takes the experiment with the reduced sets of trials. In this scenario, there would be no baseline to compare the groups with each other which endangers comparability and thus external validity, especially since the perception scales differ from person to person.

2. Each participant conducts the experiment with the full as well as the reduced sets of trials, i. e., a between participants design. This comes with the upside of being able to compare the outcome of the reduced sets with the ground truth of the full set of trials on the same participant. On the downside, it opens the door to many other questions concerning order effects and internal validity in general.

Option 2 is preferable since the threats to external validity of the first option render it useless while the reduction of damage to internal validity caused by the second possibility will be discussed in the following.

4.2 Sessions and Trials

We conduct an experiment in which the participants are shown different versions of White's Illusion, varying in the ten luminance levels for the two target patches and the two contexts they can be placed in. The two varying parameters of the targets are luminance level and target placement (context). There are 10 luminance levels being used, spaced linearly between 0.1 and 0.9, and the two contexts mentioned before: the target being placed in the black or white phase of the grating respectively. This results in a total of 20 combinations.

Each participant will produce data for all three conditions, Runtime (Dynamic) Sampling and *a priori* (Static) Sampling, and control condition (full set of trials). This is divided into four sessions, one *a priori* (Static) Sampling session with 14 repetitions, one control condition session with 7 repetitions to serve as base for Runtime (Dynamic) Sampling reduction as well as the first half of repetitions of the control condition. Therefore it needs to be held chronologically before the Runtime (Dynamic) Sampling session. These first blocks of full trials are replenished by a second full session with another 7 repetitions. The session with the 7 remaining repetitions of Runtime (Dynamic) Sampling trials is conducted separately. These are afterwards combined with the results of the first full session in order to estimate the scale. The *a priori* (Static) Sampling session can be held independently as it is not dependent on another sessions data.

In the control condition, the participants are presented each target paired with all the other targets, resulting in $(20 \times (20 - 1))/2 = 190$ stimuli presented, each repeated 14 times in total over two sessions à 7 repetitions. Hence, each participant will be making 190 x 14 = 2660 decisions in the full set of trials.

For the Runtime (Dynamic) Sampling condition session, there are 7 repetitions of the full set of 190 trials, 7 of the approx. 140 dynamically reduced ones (depending from

participant to participant but not varying greatly), resulting in approximately 2310 trials in total.

The third session type is the *a priori* (Static) Sampling session with 14 repetitions of the 114 trials selected beforehand, which results in 1596 trials.

Each of the four sessions, lasting between 20 minutes and one hour, is held on different days to ensure that the participant is not fatigued by too many decisions to be taken consecutively.

4.3 Participants

A total of 10 individuals participated in the experiment. 3 were expert participants who have completed the experiment before (GAA, JXV and RMW). 7 individuals were naive participants with no prior experience which are compensated financially (12,50€ per session, 4 sessions in total).

Regarding the order of sessions, the participants were split into two groups, consisting of one group of 5 participants (GAA, JBL, SXL, WSS, MJB) that started with the *a priori* (Static) Sampling session and the other 5 (RMW, JXV, LFD, LSN, LYF) which started with the first of two full sample sessions. An overview can be found in Table 1.

Full Sample Session First	*a priori* (Static) Sampling Session First
JXV (non-naive)	GAA (non-naive)
RMW (non-naive)	JBL
LFD	SXL
LSN	WSS
LYF	MJB

Table 1: Participant group overview of the 10 participants, divided in two equally sized groups depending on which session was held first. The non-naive participants are marked as such.

4.4 Apparatus and Procedure

The sessions were held in a room darkened by black-out curtains. The participants were being shown the stimuli on a calibrated screen. They positioned themselves on a chin rest in a constant distance of approximately 76 cm to the screen in order to ensure comparability. The participants were informed about the experiment in the beginning of the first session and had a few practice trials in the beginning of each session to become familiar with the experimental setup. For each trial, the participant had to decide which target appeared more bright to them and pressed the left or right button on the button-box accordingly. The push of the button triggered the display of the next stimulus. There was no time limit set and there was no option to skip a trial.

The participants completed the experiment once with the full set (control condition) and once with each of the reduced sets of trials (*apriori* (Static) Sampling and Runtime

(Dynamic) Sampling). From the data of all of the experiment sessions, perceptual scales were estimated using MLCM. The data of each condition was prepared combining the data of the sessions for the dynamic and full condition. This data was then fed into MLCM which estimated two scales per participant and condition: one for black and one for white context. The scales estimated from the full set of trials were used as the ground truth function to which the scales derived from the reduced sets of trials were compared to. From this comparison we determined how high the error is and furthermore derived whether using the data from the reduced set of trials is sufficient versus conducting the full experiment.

4.5 Comparison of Estimated Scales

The obvious way to compare the estimated scales is to juxtapose them visually. Since this is very crude and might overlook some of the results, it is appropriate to introduce statistical metrics. One of the most commonly used error metrics is the Root Mean Squared Error (RMSE), see equation 1.

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \tag{1}$$

In equation 1, y represents the original variable (estimate from full condition) that the estimator $\hat{y}$ (estimate from sampling strategy) is being compared to. For each perceptual scale value pair $(y_i, \hat{y}_i)$ that is being compared, the difference is being calculated and squared to eliminate the algebraic sign. The mean of this squared error from reduced trial scales to the control condition scale is then calculated over all pairs by dividing the error squares by N, the amount of luminance levels, and the square root taken to make it intuitively interpretable. It is a widely used one for several reasons. It is very intuitive in interpretation since it has the same unit as the target variable. Furthermore, it is sensitive to larger errors because the value differences are being squared before calculating the mean. Overall it is easy to calculate, compare and interpret which is why we use it here to compare accuracy of the estimations.

5 Results

The goal of this study was to determine which is a method fit to reduce resource usage in human experiments. To assess this, each participants' decision data was used to estimate a pair of scales (one per context) for each of the three conditions (*a priori*, dynamic and full). The resulting scales, six per participant, were plotted, see Figure 6 and 7. The ground truth we are comparing our two sampling strategy estimates to is the scale with the round markers, calculated from the data from the two full sessions.

Both sampling strategies display White's Effect, since the scales for the target placement on black context (black markers in plot) are perceived as having the same or higher brightness than the equiluminant targets on white context (gray markers in plot) for almost all participants and conditions. The scales also closely resemble the ones produced by Vincent et al. (2023) for all three conditions (see Figure 4). The scales are, as expected, very different across participants. To illustrate, participant SXL's scales are quasi-linear, while the scales of LFD are non-linear with a steep slope in low luminances, followed by a suggestion of a saddle point in medium luminance level and a slight slope upwards as approaching maximum luminance. This indicates that the individuality of the perceptual encoding functions was captured by the estimations.

While variance across participants is high, the results for each participant are mostly consistent across conditions. For participant SXL, all three conditions produce very similar results which is reflected in all plots lying very close together. For LFD, LSN, LYF, RMW and GAA, all three conditions produced scales of similar general shape but with greater variance as opposed to SXL. For MJB, JXV, JBL and WSS, one can tell that while one sampling method produced results very close to the control condition, the other one lies far off.

For MJB, JBL and WSS, the *a priori* (Static) Condition produced a shifted scale, in which the general brightness perceived is higher than in the other two conditions in most data points. Even the perceived brightness for the white context in the *a priori* (Static) Condition is higher than the one in black context of the other two conditions for most data points. In JXV's case, while also having a seemingly bigger deviation from the control group for the *a priori* (Static) Condition, all three scales show that the targets in black context are perceived as brighter than the ones in white context. For MJB, JBL, WSS and JXV, it is apparent from the plots that in the *a priori* (Static) Condition, estimation performed worse than in the Runtime (Dynamic) Condition. This is not always the case, for the other participants it is just not interpretable as intuitively on first view, like participant LFD where both sampling methods deviate from the control condition but in different directions which makes intuitive interpretation difficult. Since it is not always apparent, which reduction method is closer to the full method estimation, it is necessary to calculate errors to describe how far the scales from the trial selection strategies are from the control condition.

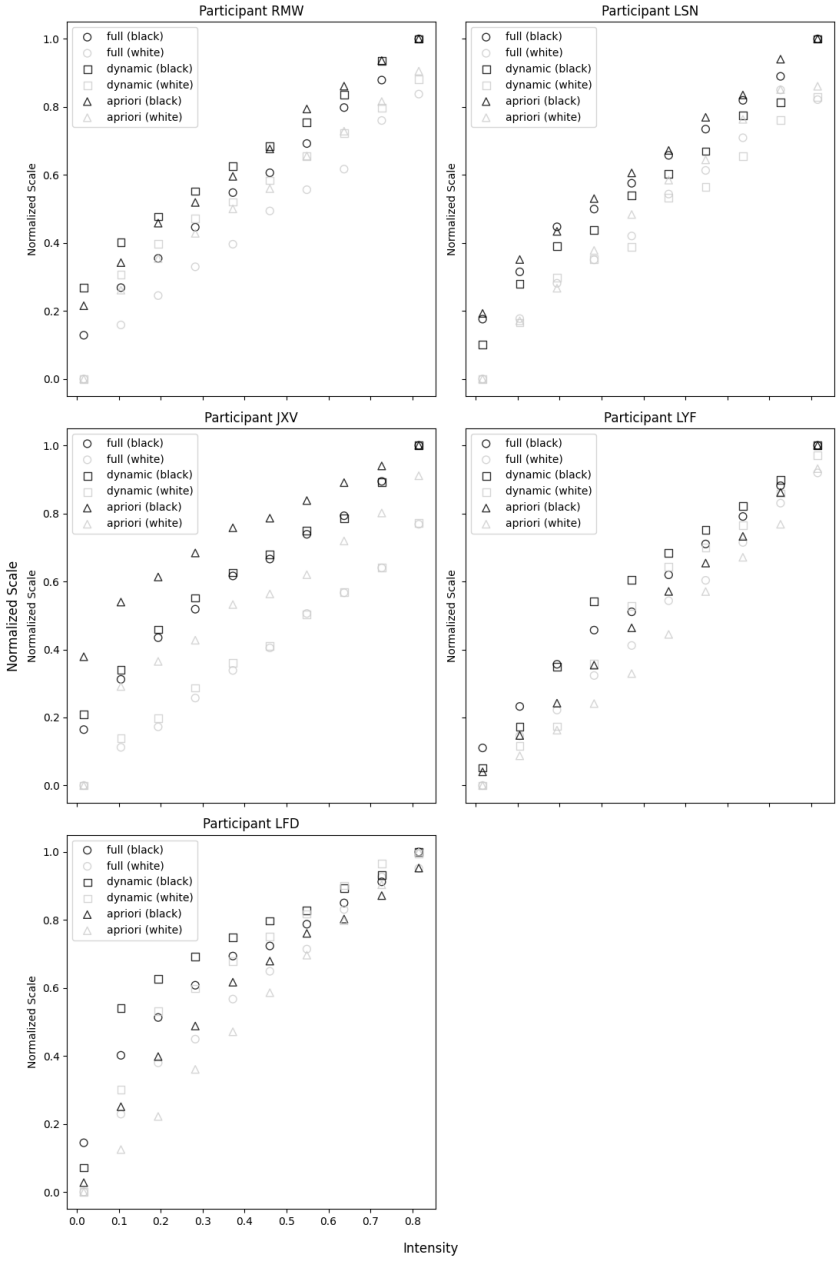

Figure 6: Estimated perceptual scales for participants that started with the control condition. JXV and RMW are non-naive participants. Each participant's scales for the three session types are displayed in one panel. White context is displayed in gray, black context in black. On the x-axis, the intensity, i. e. luminance of the target is displayed, while the y-axis represents the normalized perceived brightness. The full condition uses circular markers, the dynamic condition squares and the *a priori* condition triangles.

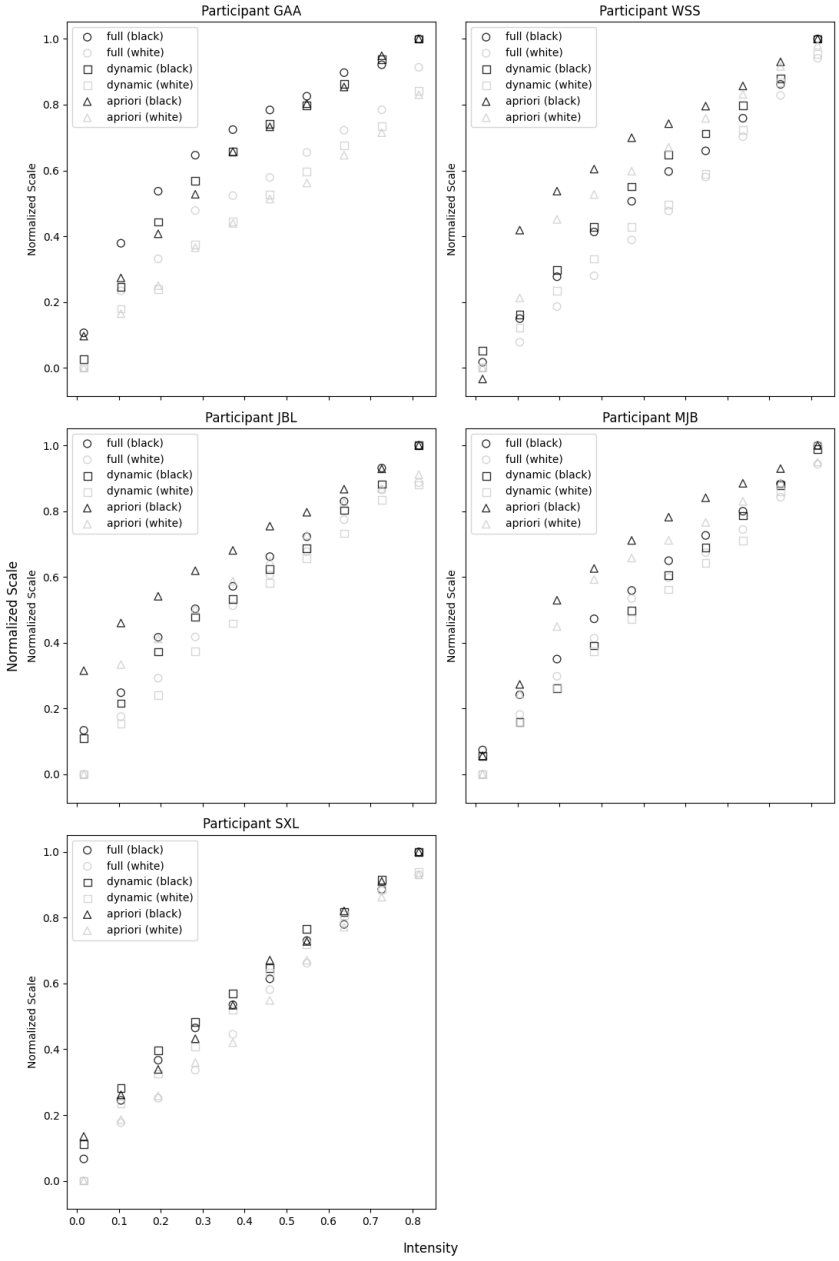

Figure 7: Estimated perceptual scales for participants that started with the *a priori* (Static) Condition. GAA is the only non-naive participant. Each participant's scales for the three session types are displayed in one panel. White context is displayed in gray, black context in black. On the x-axis, the intensity, i. e. luminance of the target is displayed, while the y-axis represents the normalized perceived brightness. The full condition uses circular markers, the dynamic condition squares and the *a priori* condition triangles.

The comparison of the errors of the two sampling conditions $RMSE_d$ and $RMSE_a$ in Figure 8 reveals that six participants had a bigger error in the *a priori* (Static) Condition. These are the data points in the plot which lie above the dashed equilibrium line. While the remaining 4 participants had a bigger error in the Runtime (Dynamic) Condition, the data points lie closer to the dashed line. This means that the errors do not diverge as much. This is confirmed by the comparison of the mean RMSE over all participants: $\bar{RMSE}_d = 0.0538 < \bar{RMSE}_a = 0.0887$, see Table 2. Figure 9 offers another visualization with the RMSE plotted in two histograms, one for each reduced sampling condition in comparison to the full condition. On a frist glance, one can already see that the RMSEs for the Runtime (Dynamic) Condition, on the left panel, is less widely spread and closer to 0. Especially for the non-naive participants JXV, GAA and RMW, the RMSE of the Runtime (Dynamic) Condition ($RMSE_d = 0.062267$) is significantly lower than the *a priori* (Static) one ($RMSE_a = 0.102433$). Across the participants without experience, the Runtime (Dynamic) one, while being less extreme in difference, still has lower error.

Participant	$RMSE_d$	$RMSE_a$	$(RMSE_d - RMSE_a)$
GAA (non-naive)	0.0678	0.0756	-0.0078
JBL	0.0341	0.0984	-0.0643
JXV (non-naive)	0.0197	0.1509	-0.1312
LFD	0.0866	0.0837	0.0029
LSN	0.0473	0.0316	0.0157
LYF	0.0605	0.0652	-0.0047
MJB	0.0466	0.1059	-0.0593
RMW (non-naive)	0.0993	0.0808	0.0185
SXL	0.0426	0.0276	0.0150
WSS	0.0335	0.1672	-0.1337
average	0.0538	0.0887	-0.0349

Table 2: RMSE of each participant. In the second column for the Runtime (Dynamic) and in the third column for the *a apriori* (Static) reduction method. The difference of the two is calculated in the fourth column. Below are the three corresponding arithmetic means. A lower RMSE corresponds to higher accuracy.

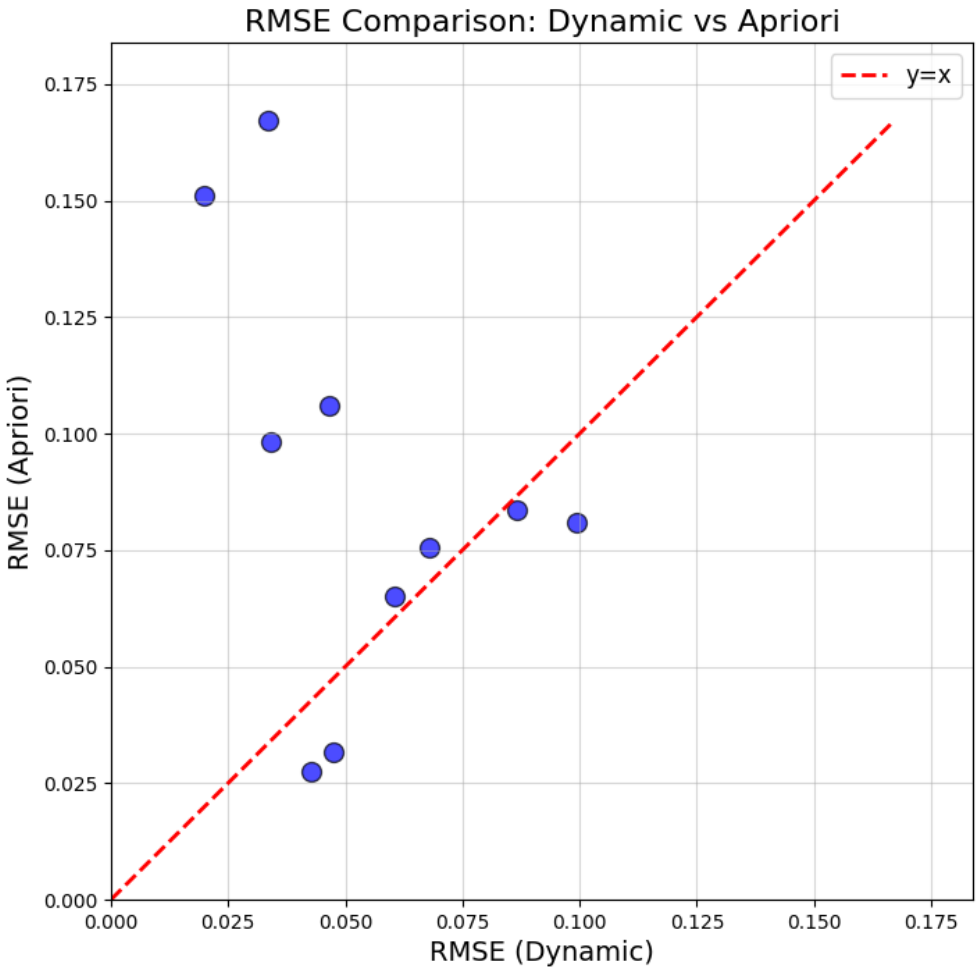

Figure 8: Comparative scatterplot of Runtime (Dynamic) (x-axis) and *a priori* (Static) (y-axis) conditions RMSE. Each data point represents one participant. The dashed red line marks the equilibrium of $RMSE_d = RMSE_a$.

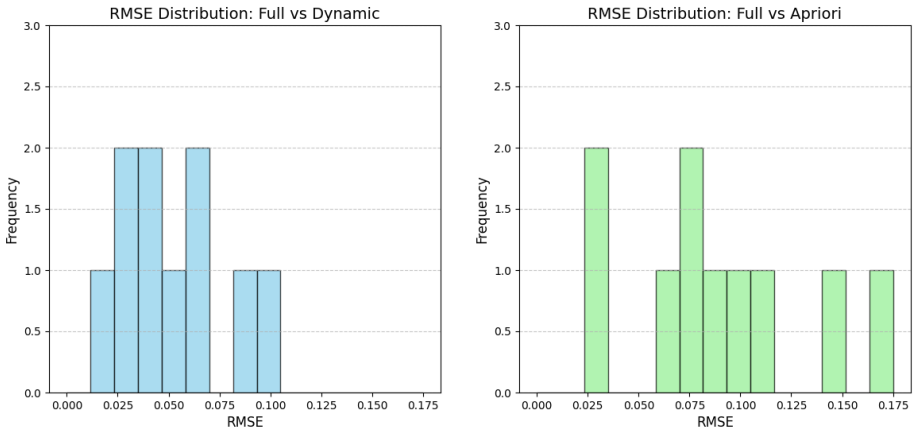

Figure 9: In the histograms, the Root Mean Squared Error (RMSE) of the two reduction methods in comparison to the full method for all 10 participants are displayed. On the left side, the RMSE of the Runtime (Dynamic) Condition to the Full Condition, on the right side, the RMSE of the *a priori* (Static) Condition to the full Condition. The x-axis represents the RMSE, while the y-axis represents the absolute frequency. A lower RMSE corresponds to higher accuracy.

The participants can also be viewed as two groups regarding session order: Participants GAA, JBL, SXL, WSS and MJB all completed the *a priori* (Static) session first, the remaining participants (RMW, JXV, LFD, LSN, LYF) started with a full session that was used to dynamically reduce the trials. While the *a priori* (Static) Conditio still consistently performed worse than the Runtime (Dynamic) one, the incline differed a bit: The group that started off with the *a priori* Condition had $RMSE_d = 0.04492$ and $RMSE_a = 0.09494$, an error more than twice as big in the *a priori* (Static) Condition, while the group starting with the full session had $RMSE_d = 0.06268$ and $RMSE_a = 0.08244$, with a smaller yet still significant difference between trial reduction methods. The significantly higher error of the *a priori* (Static) Condition in the group that started with the *a priori* session could be due to practice effects. This would mean that they answered the trials less consistently in the first session held due to a lack of experience.

The result of the different sessions can also be displayed in heat maps, examples can be found in Figure 11 and 10. Figure 10 shows the ground truth frequencies, collected over the two full sessions of participant GAA. Figure 11 shows the choice frequencies of the same participant in the *a priori* reduced session, which is also indicated by some cells of the heat map missing in comparison to the full one of Figure 10, since these trials have been removed prior to the experiment session.

6 Discussion

The goal of this thesis was to determine whether the number of trials in Maximum Likelihood Conjoint Measurement (MLCM) experiments can be reduced while only impacting the resulting scales minimally. MLCM, a statistical scaling method developed by Ho et al. (2008) and Knoblauch and Maloney (2012), was used to estimate perceptual encoding functions. To test this, we conducted an experiment using White's Illusion similar to the one of Vincent et al. (2023). Two sampling strategies, developed and simulation-tested by Zabel (2023), were applied to reduce trial size:

1. The *a priori* (Static) Sampling Strategy which excluded trials that were expected to always be answered the same because of big luminance differences

2. The Runtime (Dynamic) Sampling Strategy which samples during the presentation of trials to a participant, excluding the trials that were consistently answered the same after the initial half of repetitions and only showing the remaining trials in the second half.

Zabel has successfully tested these sampling strategies using a simulation in his thesis (Zabel, 2023). It was still in question if his results would be reproducible in human experiments though. Each participant conducted the experiment in three conditions: The *a priori* (Static) reduced condition, the Runtime (Dynamic) reduced condition and a control condition using the full set of trials as used by Vincent et al. This served as replacement for the ground truth function used by Zabel. The results of these experiment sessions have been used to estimate perceptual scales for each condition and participant with MLCM. The resulting scales of the sampling strategies were then compared to the scales of the full condition using the Root Mean Squared Error.

Sampling Strategies	added $RMSE_s$	added $RMSE_h$
Full	0	—
Static	(0.0701-0.0696) = 0.0005	0.0887
Dynamic	(0.0737-0.0696) = 0.0041	0.0538

Table 3: Comparison of increases in Error. The second column contains the added RMSE of the sampling conditions versus full condition of Zabel's simulation (Zabel, 2023), the third column for the human experiment conducted by me.

Zabel used a ground truth function as the base of his simulation and also used it to compare the estimation of the two sampling strategies as well as the estimate resulting from the simulation of the full experiment to. As we do not have the actual perceptual encoding function of our participants—this is what our goal is to estimate—I used the full condition as a replacement. The full condition scales that we are using as replacement for the ground truth function are already an estimate. Zabel had an RMSE of 0.0696 for the full condition versus the ground truth function (see Table 3). This can be interpreted as an error of 6.96 percent due to the normalized scale values. Compared to the RMSEs

of the statically (*a priori*) and dynamically (Runtime) reduced conditions which are $RMSE_a = 7.01\%$ and $RMSE_d = 7.37\%$, the difference was only marginal which implied that reducing the trials to only the "hard" ones did not impact the quality of estimation significantly. The increase in error of the Runtime (Dynamic) Condition $RMSE_d - RMSE_f = 0.0041$ was bigger than the one of the *a priori* (Static) Condition $RMSE_a - RMSE_f = 0.0005$. In the scales estimated from our experimental data, we saw average RMSEs of $\bar{RMSE}_d = 0.0538$ and $\bar{RMSE}_a = 0.0887$ over all participants. This can be read as an error of 5.38% and 8.87%. While being larger than the ones from the simulation, they are of comparable magnitude as the increases in error of the reduced conditions from Zabel's results. Most importantly, even the higher error of the *a priori* (Static) Condition is quite low. From this we can conclude that in human experiments, the reduction strategies still produced reliable estimates comparable to the ones from Zabel's simulations in quality.

While Zabel had slightly better results using the *a priori* (Static) reduction method than with the Runtime (Dynamic) one, we could not replicate these results in human experiments. It was the other way around, with the Runtime (Dynamic) Condition producing slightly better results than the *a priori* (Static) sampling method. This could be due to the individuality of the participants versus the simulation. While it is easy to exclude trials for a simulated observer which always uses the same ground truth function, the estimation of the flexible dynamic approach is more fit for being used on different individuals. Over all participants, the mean RSME of the *a priori* (Static) reduction method was almost 65% higher than the Runtime (Dynamic) one, which indicates that the Runtime (Dynamic) one is to be preferred. This is true for the experts well as the naive participants that completed the experiment and for both groups regarding session order.

One reason for this could be the difference in trial amount, while the Runtime (Dynamic) Condition session had about 2310 trials over two sessions in total per participant while the *a priori* (Static) session had 1596. This could have a direct influence on the quality of estimation since there is less data to use in the statistical methods applied.

Furthermore, the *a priori* (Static) trials were all presented in one session, while the Runtime (Dynamic) ones were split over two: the first session with 1330 trials and the second, dynamically reduced one with around 1000, depending on the participant. This is the case because the Runtime (Dynamic) Condition is divided into two stages: The initial stage, showing the full set of trials and the sampling stage, showing only the trials which did not show consistent decisions in the initial stage. These two stages are divided over two sessions. The shorter Runtime (Dynamic) sessions could lead to a higher level of concentration for the participants and thus indirectly to a more accurate estimation. To tackle this issue, it could be tested how the Runtime (Dynamic) Condition would perform if the data was collected in a single session in future studies.

The fact that there was four sessions to be completed lead to the issue of order effects. This was tackled by splitting the participants into two groups: one starting with the *a priori* (Static) session and one starting with the Full Session serving as first half of the data for both the Runtime (Dynamic) and Full condition. Repeating the experiment with more participants would allow more groups and thus more insight on the impact

of order effects on the quality of estimations.

The efficiency of the experiment is also impacted by the threshold that determines which trials are excluded. The Runtime (Dynamic) reduction method was very strict, using 100% and 0% as thresholds for choice probability, while allowing one missed click could reduce the amount of trials presented in the second part of the Runtime (Dynamic) data collection by much more, making it even more time efficient.

Several of the participants mentioned that trials were prone to provoke "wrong" decision making when a very bright target is placed in white context or a very dark one on black. Since the decisions are taken very quickly, the participants might not realize it is a light target on white and assume its a dark one on black or vice versa. This makes regular careless errors probable. These could be avoided by elimination of those stimuli prone to provoke this error *a priori* since the dynamic reduction method will not eliminate them due to the influence of the careless errors on the choice probabilities. Thus, one could test whether its beneficial to combine the two reduction methods, excluding the extreme luminance differences *a priori* (Static) and then further reducing the trials using the Runtime (Dynamic) Sampling method.

This thesis tested the reliability of trial reduction strategies for Maximum Likelihood Conjoint Measurement (MLCM) in human experiments using White's Illusion. It investigated whether the dynamic (Runtime) and *a priori* (Static) sampling strategies developed by Zabel (2023) could maintain the accuracy of perceptual scale estimates while reducing the number of trials. The results confirmed that both strategies produce reliable estimates, with the Runtime (Dynamic) strategy outperforming the *a priori* (Static) one, as evidenced by lower Root Mean Squared Error (RMSE) values. Despite this, both strategies yielded low RMSEs overall, indicating that both of them are effective alternatives to the full trial set. Future studies could further optimize these methods to enhance accuracy and resource efficiency. Combining both strategies may offer additional benefits.

References

Adelson, E. (2000). Lightness perception and lightness illusion. In *The new cognitive neurosciences* (2nd ed.). M. Gazzaniga.

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science*.

Kingdom, F. A. A. (2014). *Brightness and lightness*. EBSCO.

Knoblauch, K., & Maloney, L. (2012). Maximum likelihood conjoint measurement. In *Modeling psychophysical data in r*. Springer.

Vincent, J., Maertens, M., & Aguilar, G. (2023). What Fechner could not do: Separating perceptual encoding and decoding with difference scaling. *Journal of Vision*.

Zabel, J. (2023). *Optimizing data acquisition for maximum likelihood conjoint measurement*. [Bachelor's thesis, TU Berlin].

7 Appendix

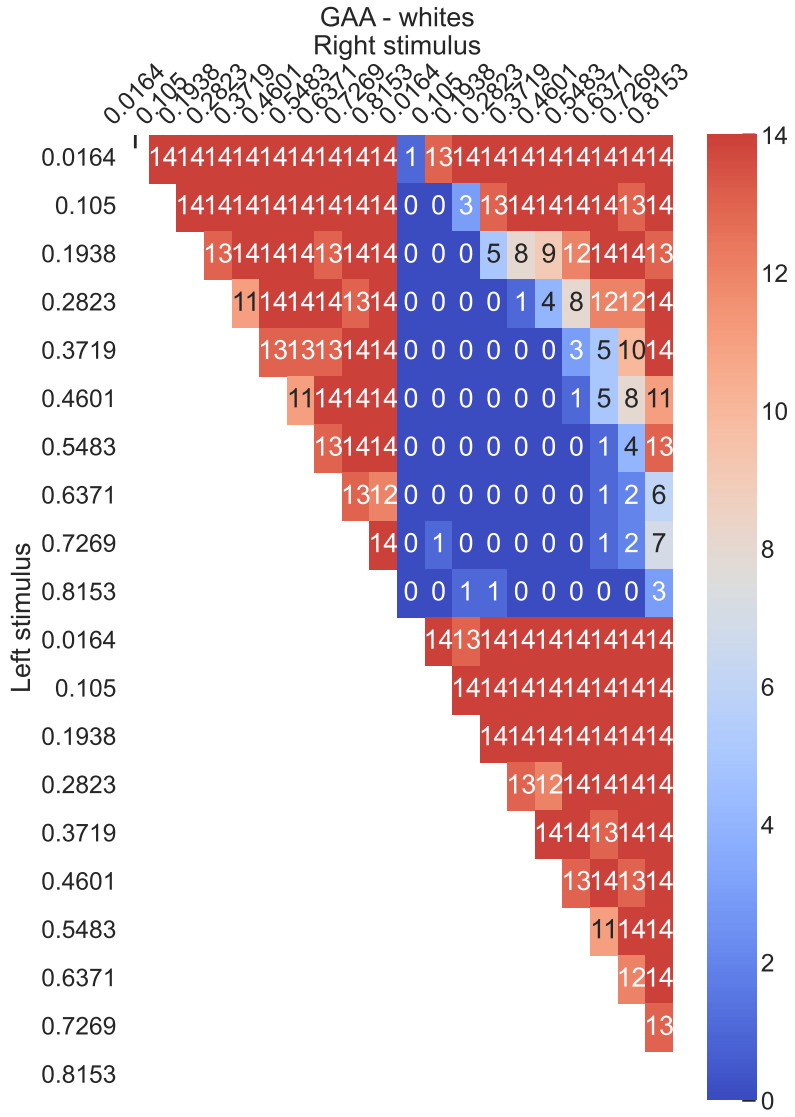

Figure 10: Exemplary heat map showing the absolute frequency of responses for White's Illusion trials for one participant over the two full sessions. The x- and y-axes represent the different contexts and luminance levels, with the stimuli on these axes being compared against each other. Each cell value indicates how often the participant chose one stimulus over another across 14 repeated trials.

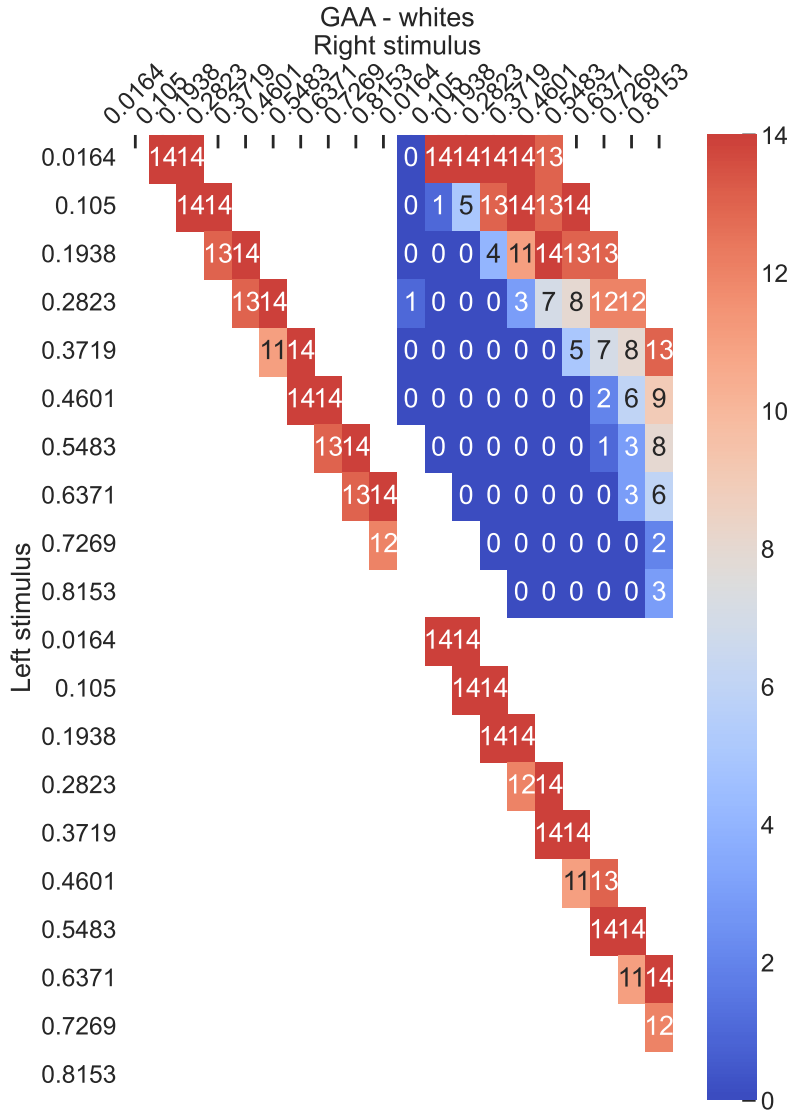

Figure 11: Exemplary heat map showing the absolute frequency of responses for White's Illusion trials for one participant in the *a priori* session. The x- and y-axes represent the different contexts and luminance levels, with the stimuli on these axes being compared against each other. Each cell value indicates how often the participant chose one stimulus over another across 14 repeated trials.

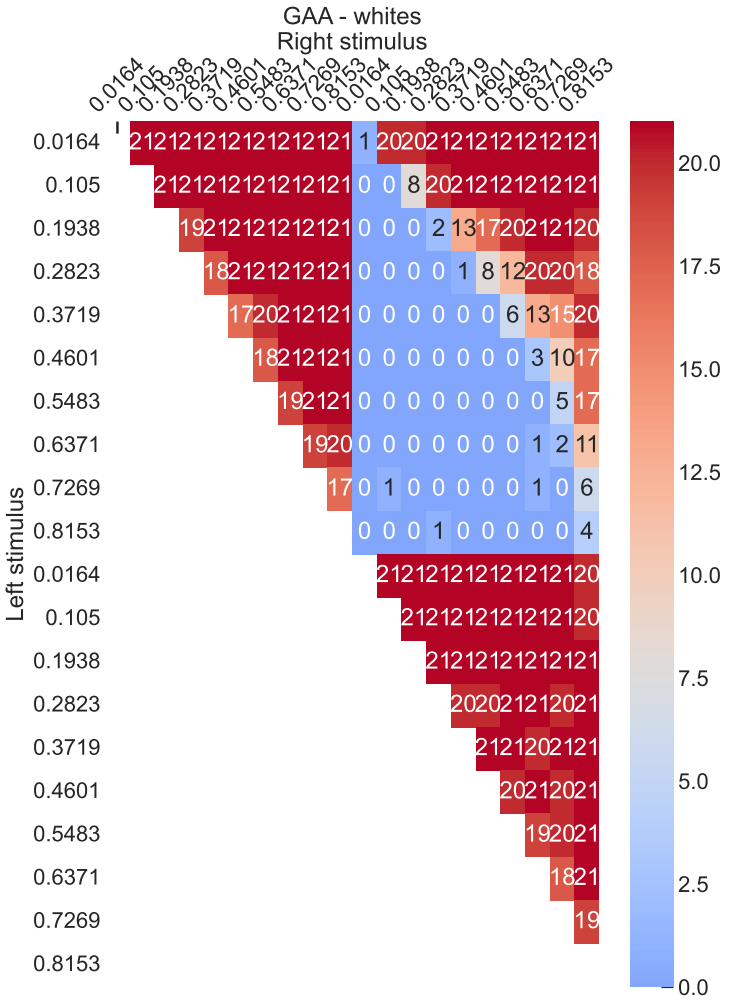

Figure 12: Participant GAA: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

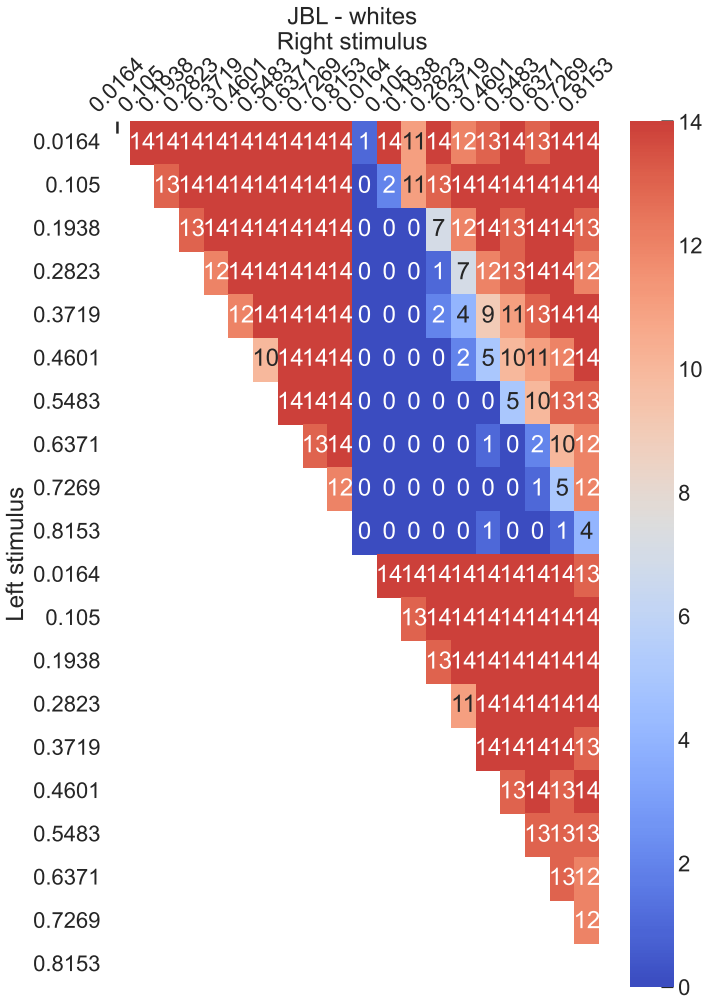

Figure 13: Participant JBL: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

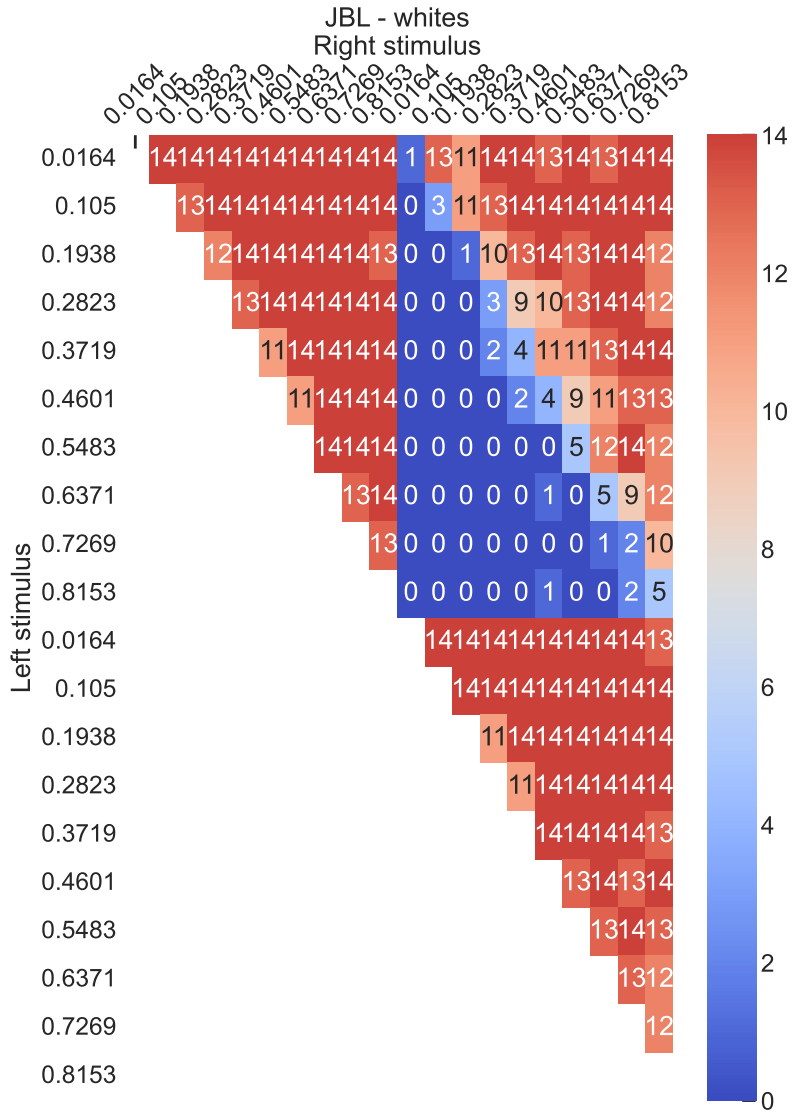

Figure 14: Participant JBL: Heatmap of absolute frequencies of choice in control condition.

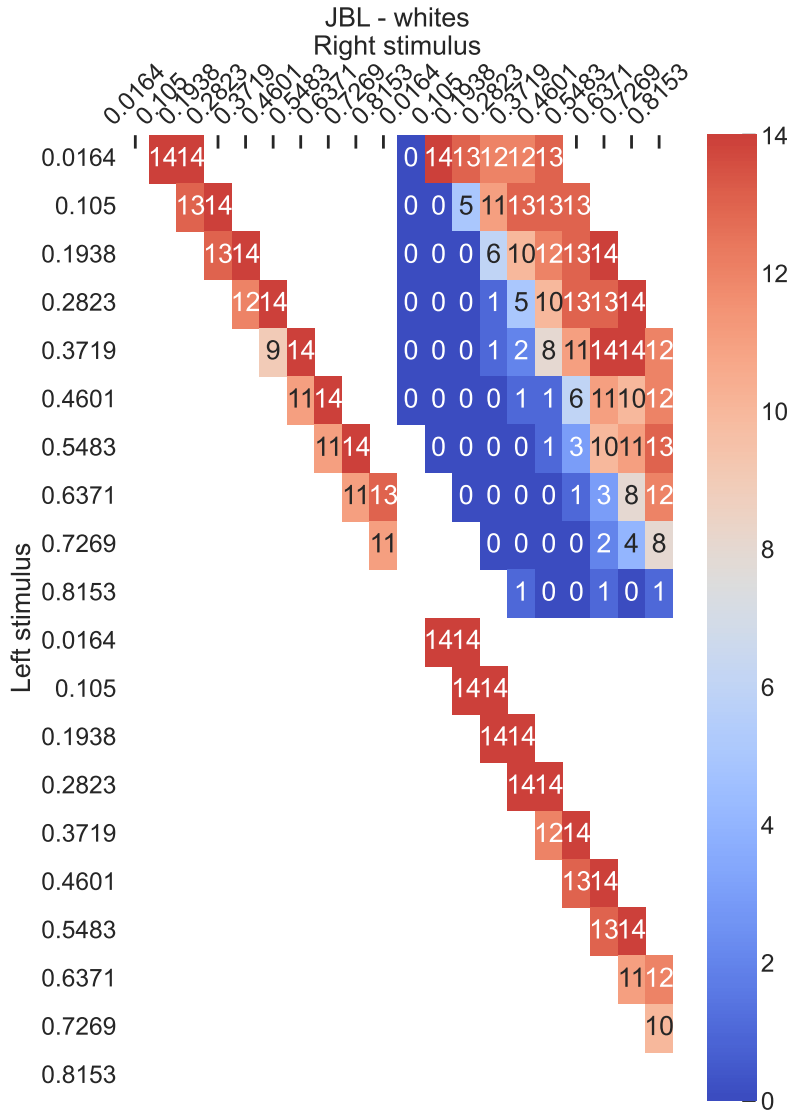

Figure 15: Participant JBL: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

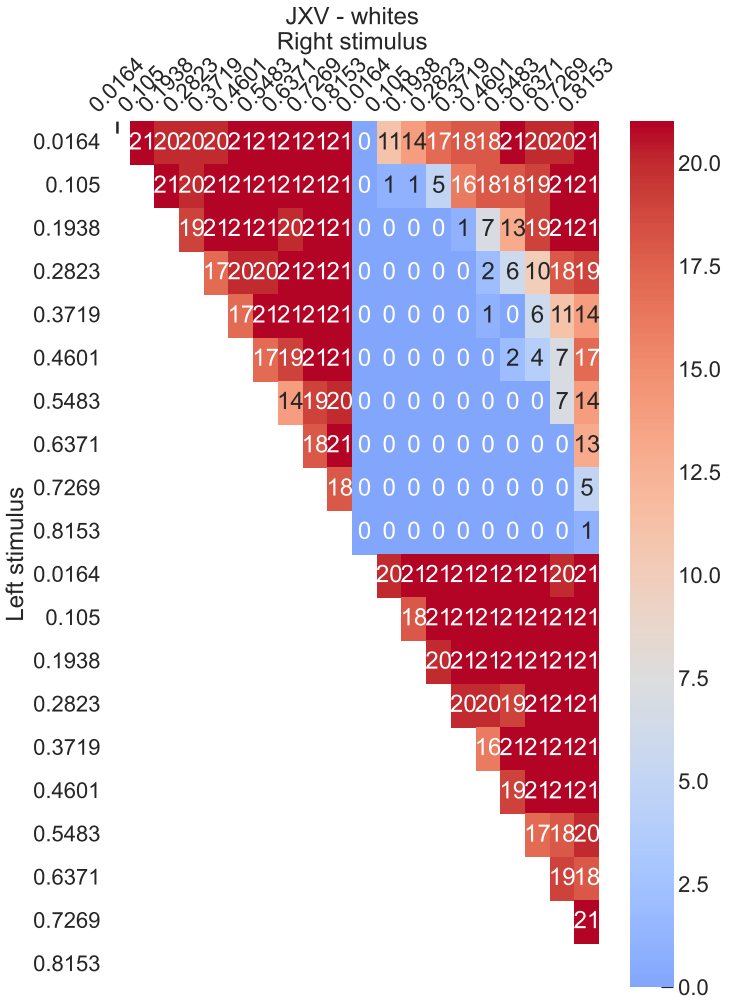

Figure 16: Participant JXV: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

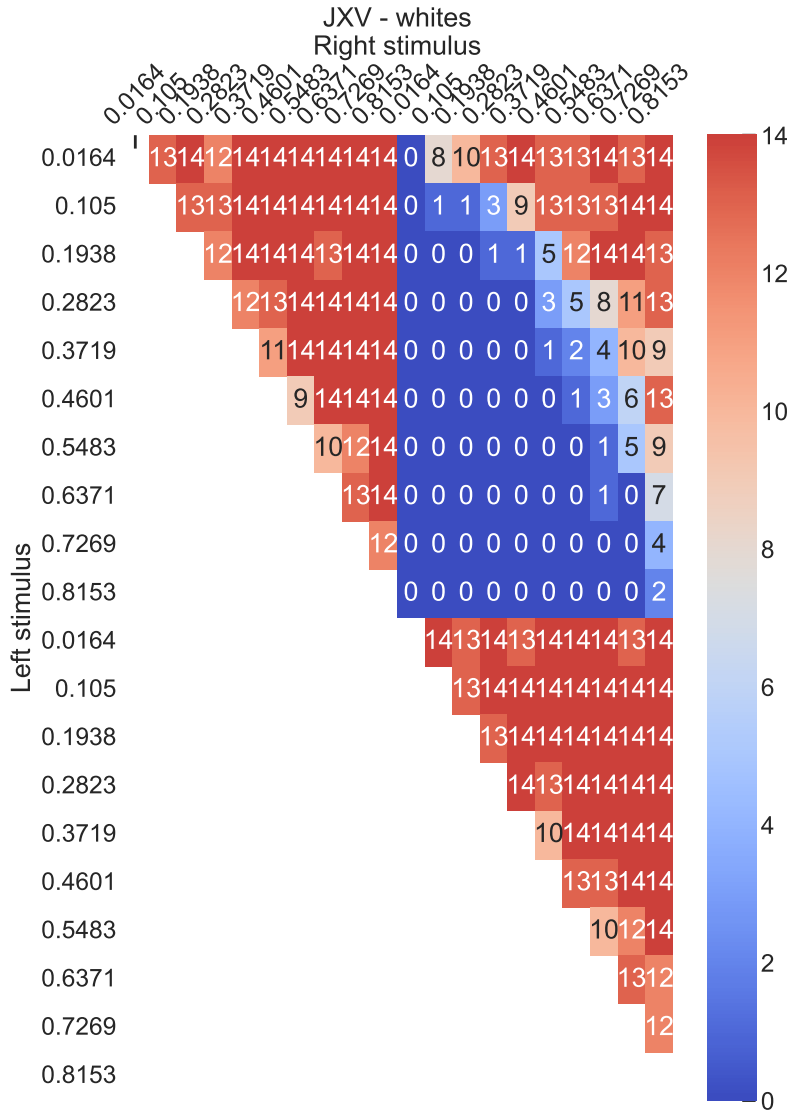

Figure 17: Participant JXV: Heatmap of absolute frequencies of choice in control condition.

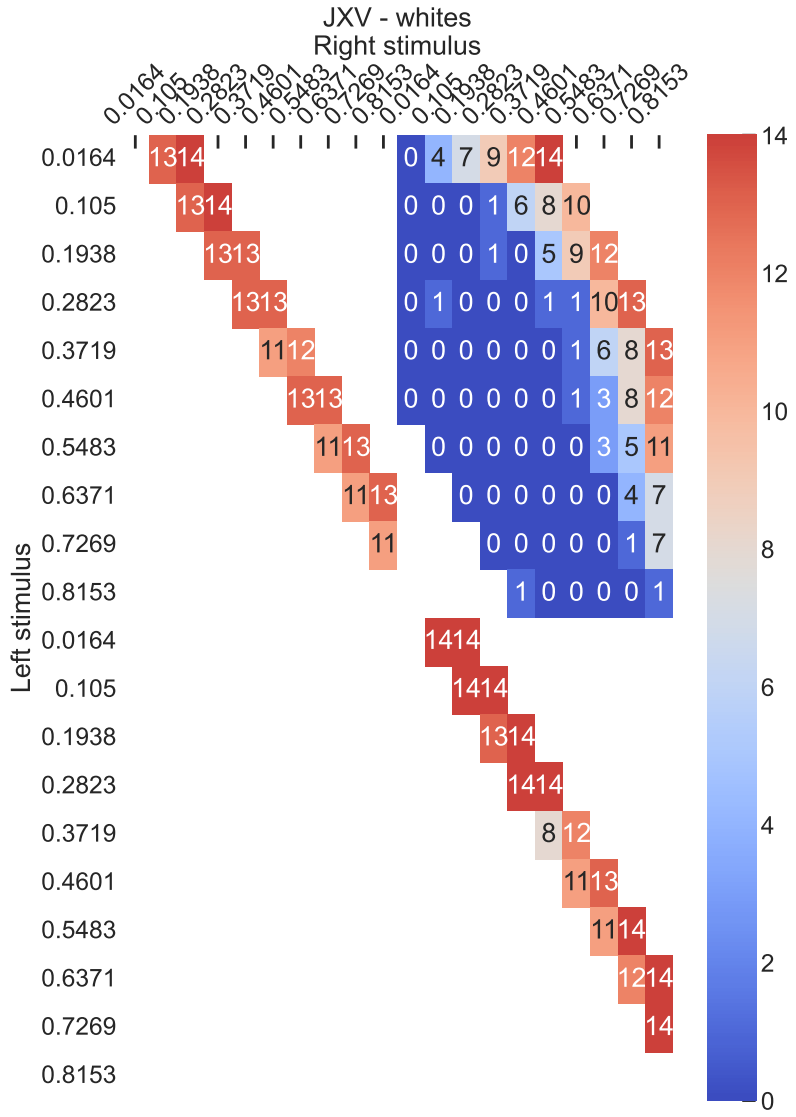

Figure 18: Participant JXV: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

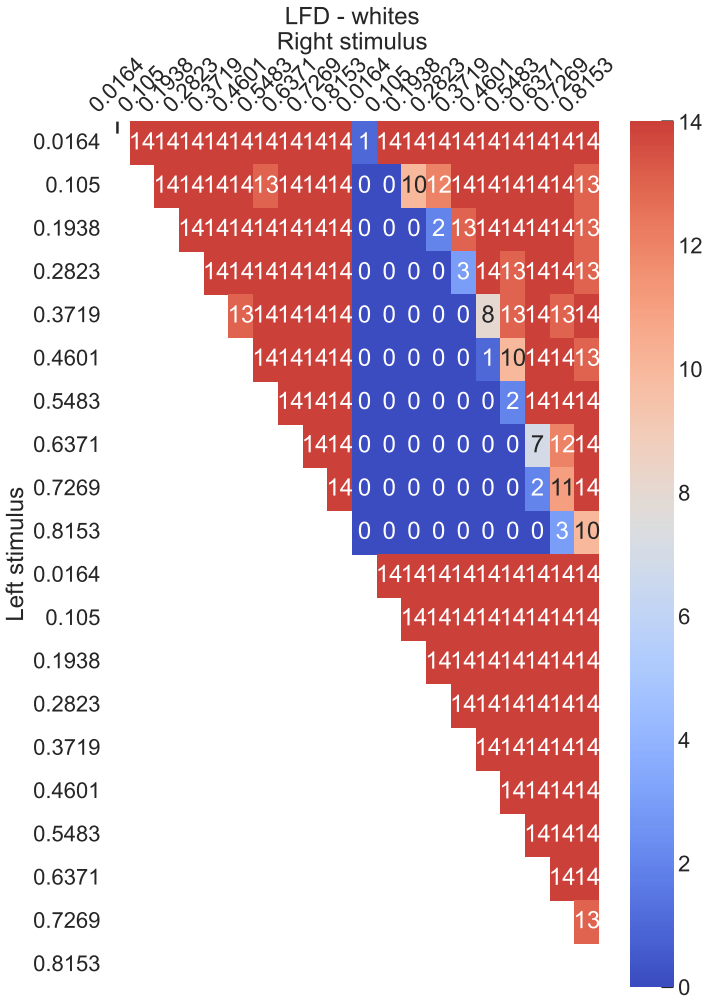

Figure 19: Participant LFD: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

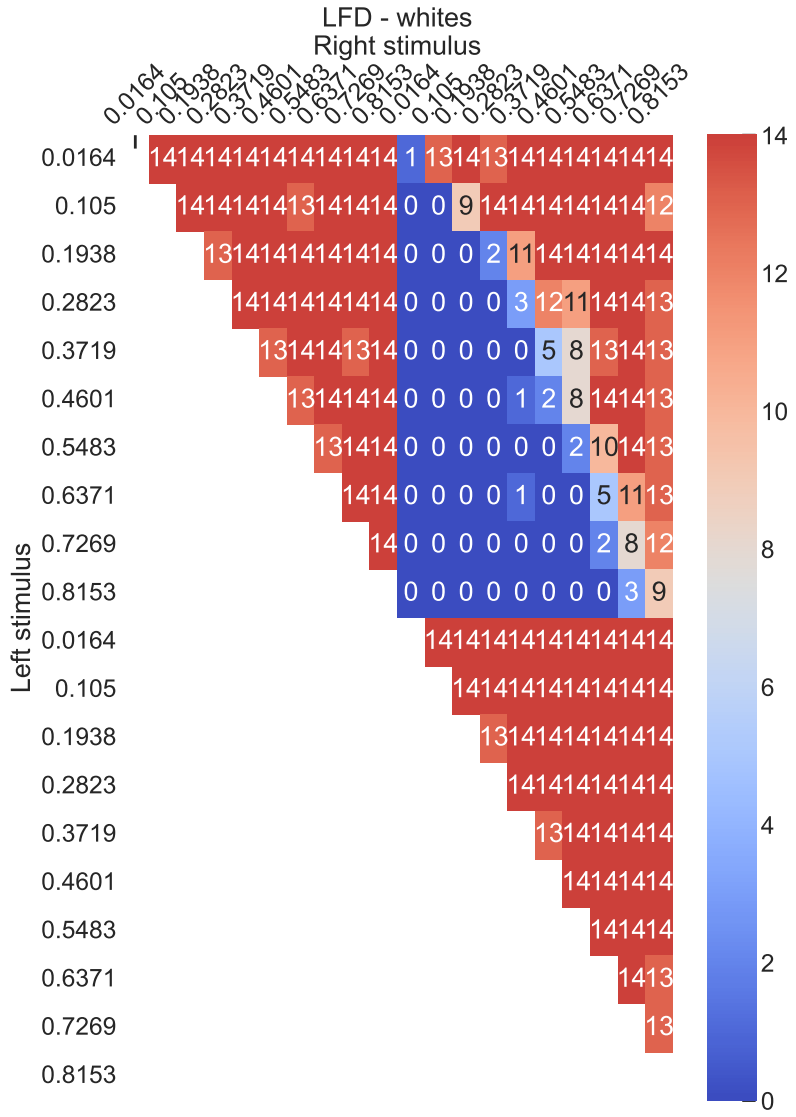

Figure 20: Participant LFD: Heatmap of absolute frequencies of choice in control condition.

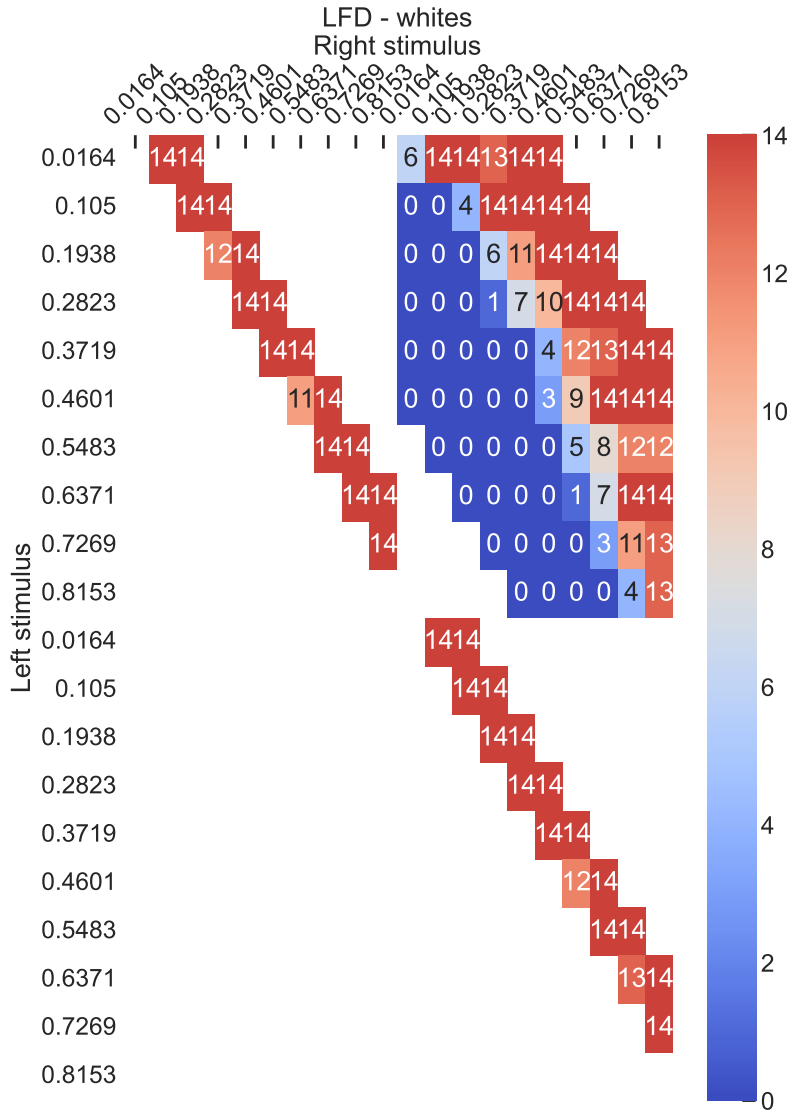

Figure 21: Participant LFD: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

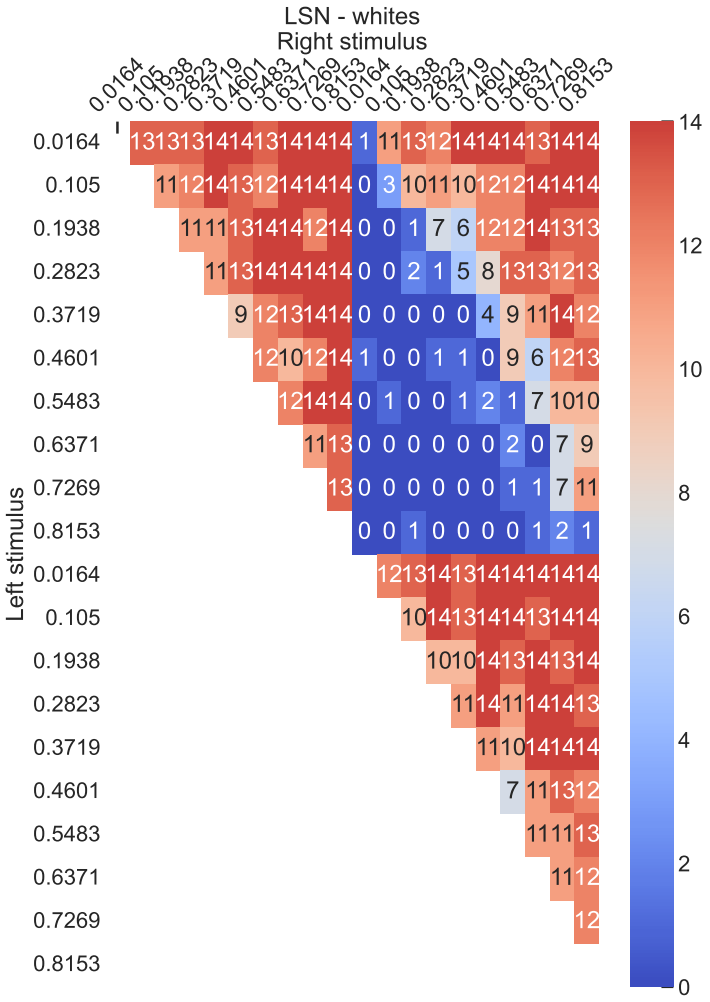

Figure 22: Participant LSN: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

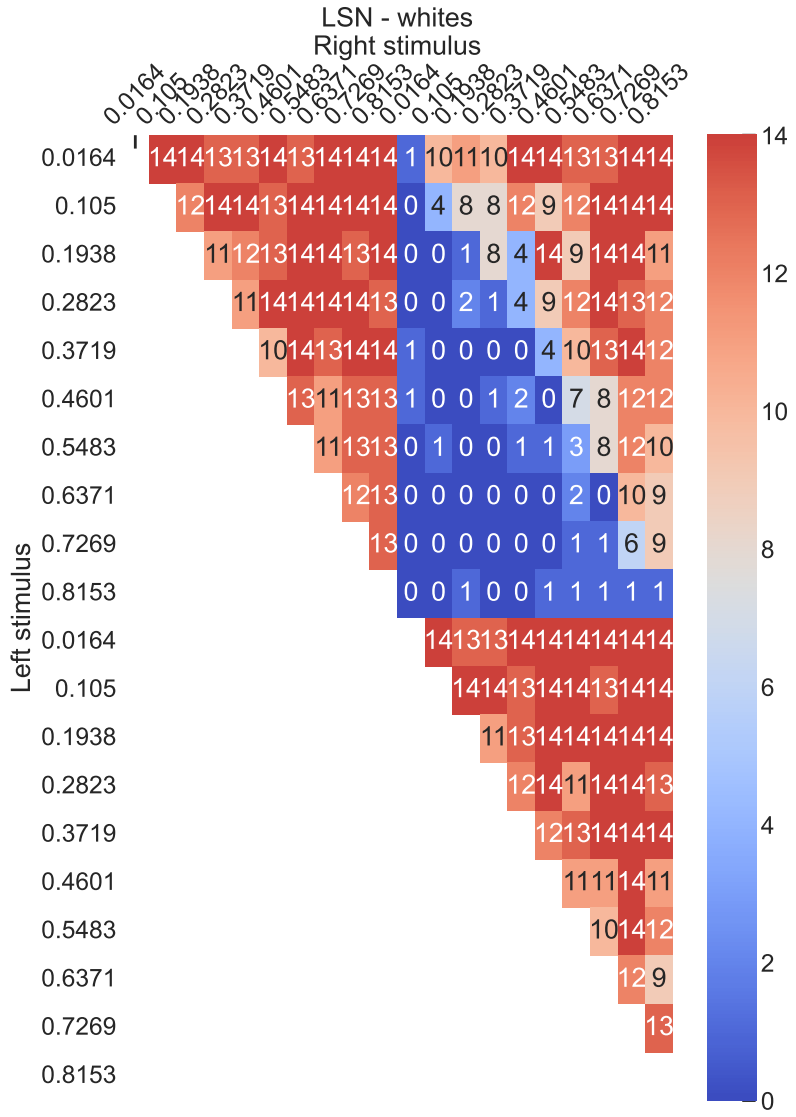

Figure 23: Participant LSN: Heatmap of absolute frequencies of choice in control condition.

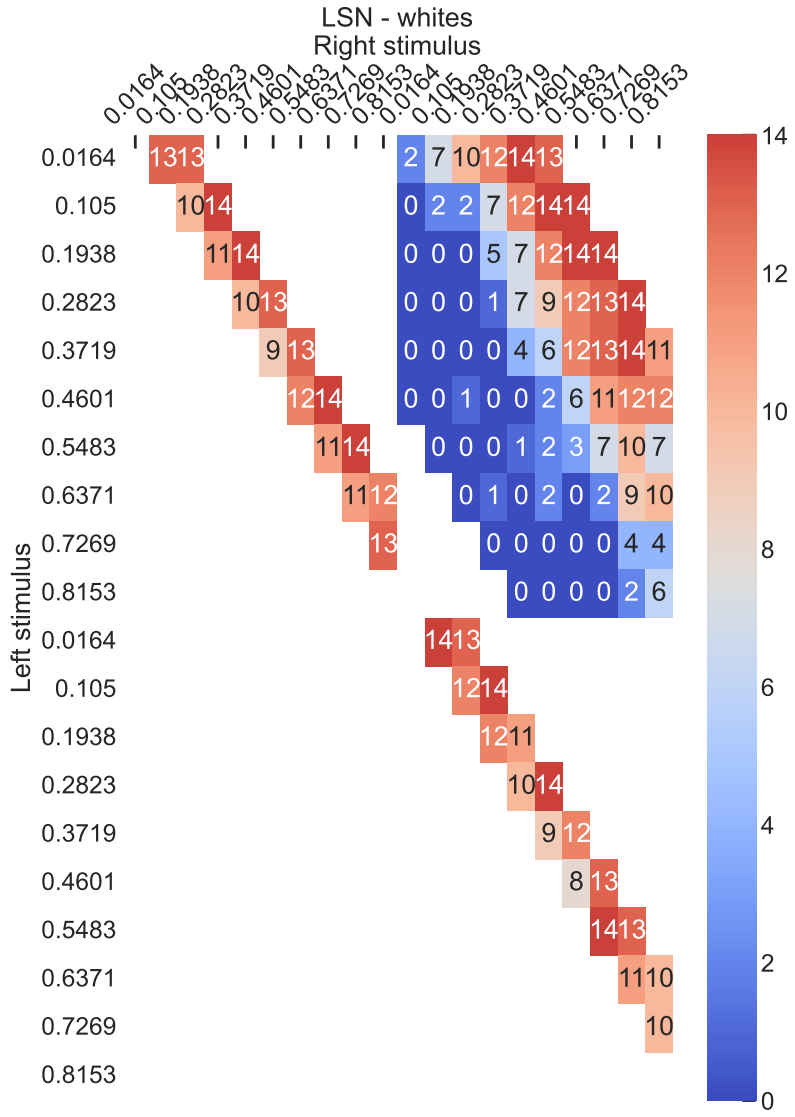

Figure 24: Participant LSN: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

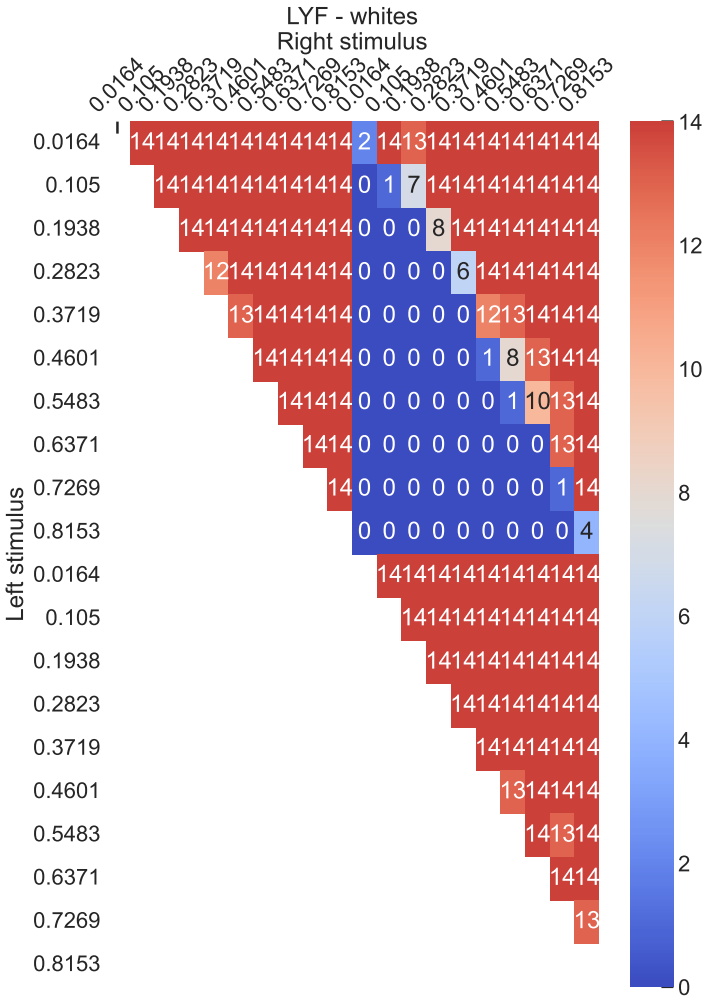

Figure 25: Participant LYF: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

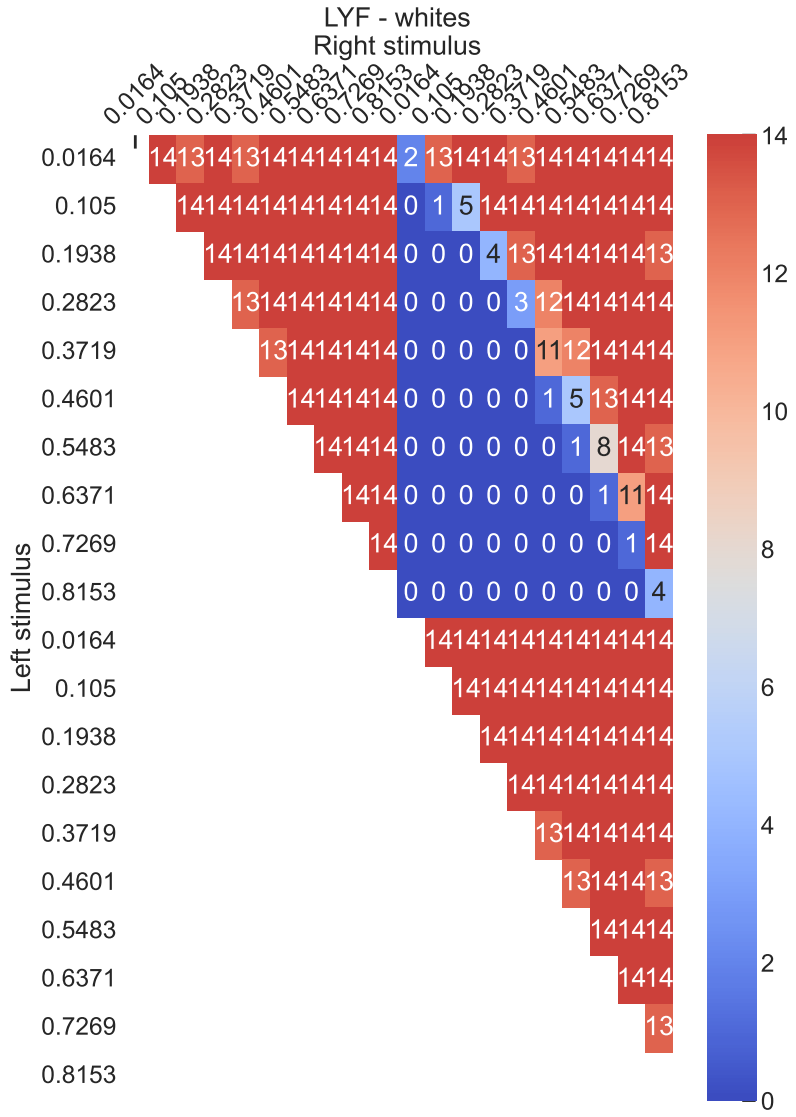

Figure 26: Participant LYF: Heatmap of absolute frequencies of choice in control condition.

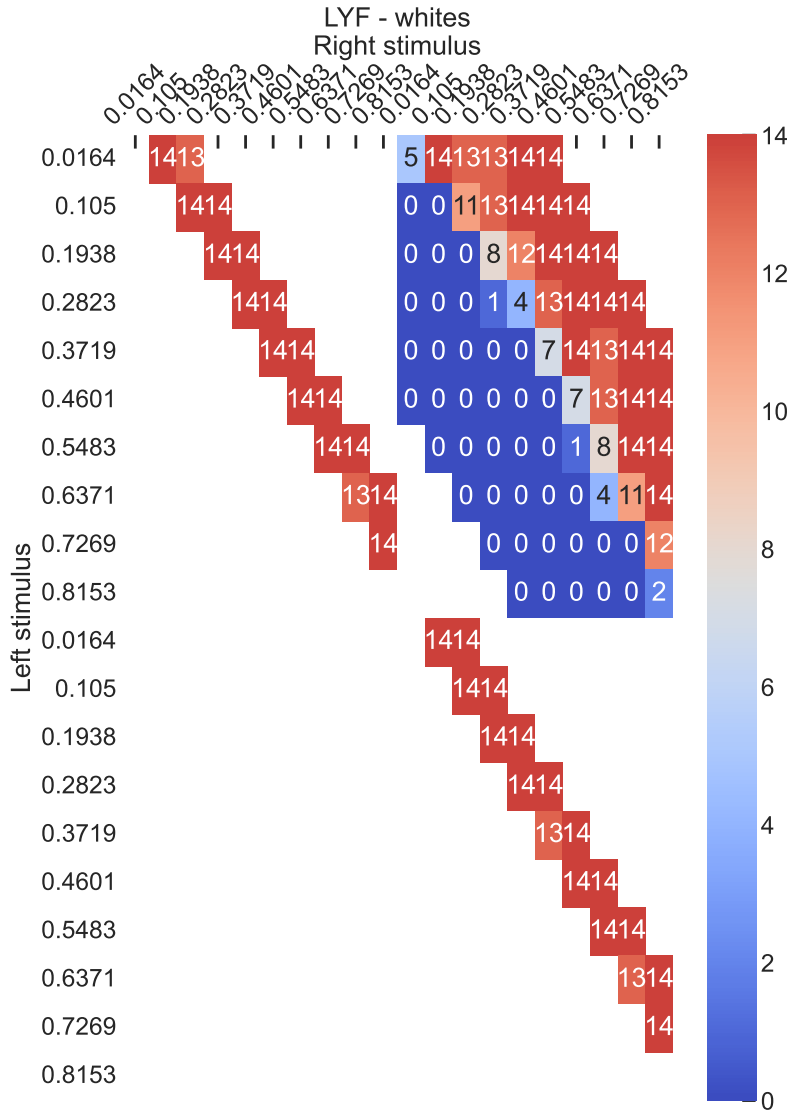

Figure 27: Participant LYF: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

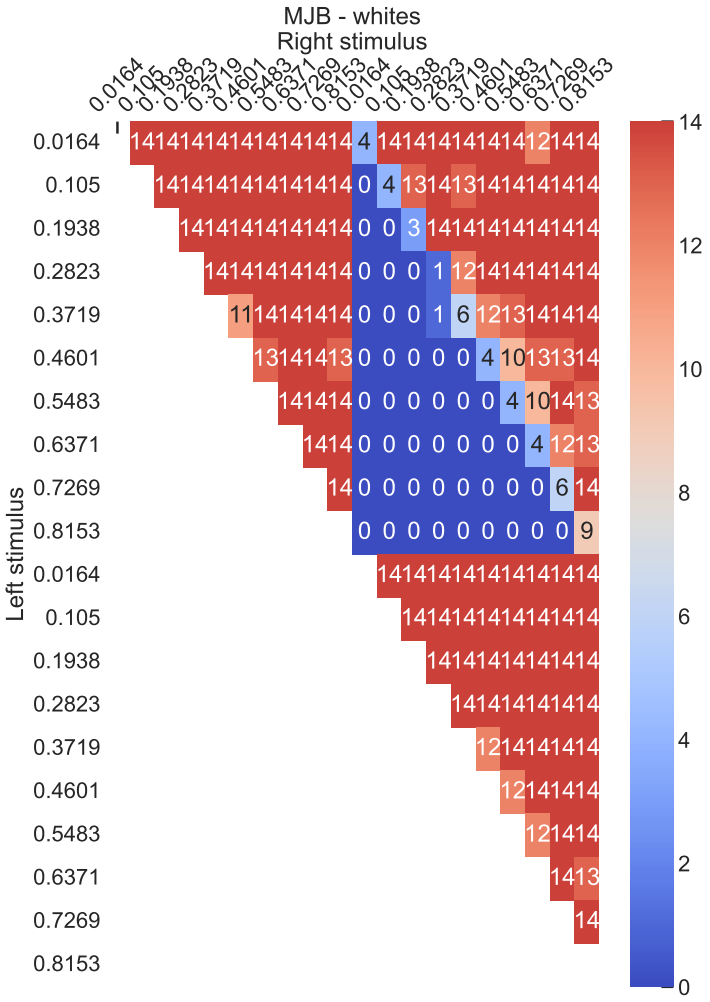

Figure 28: Participant MJB: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

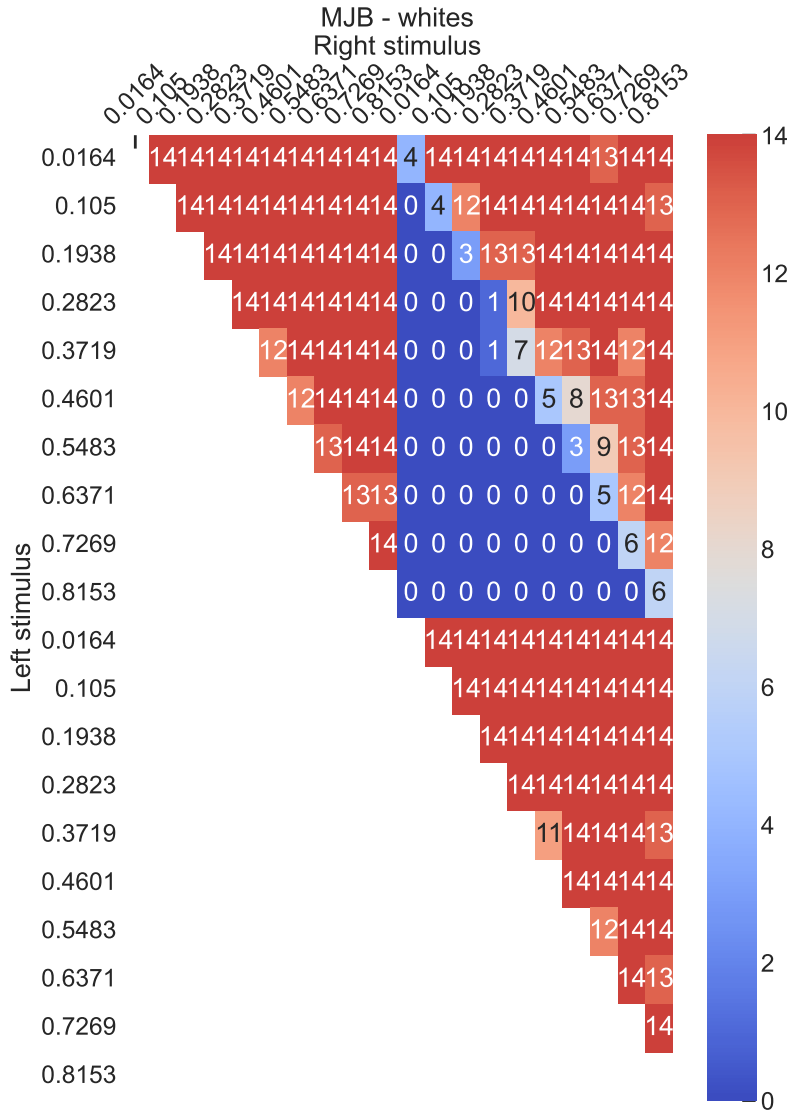

Figure 29: Participant MJB: Heatmap of absolute frequencies of choice in control condition.

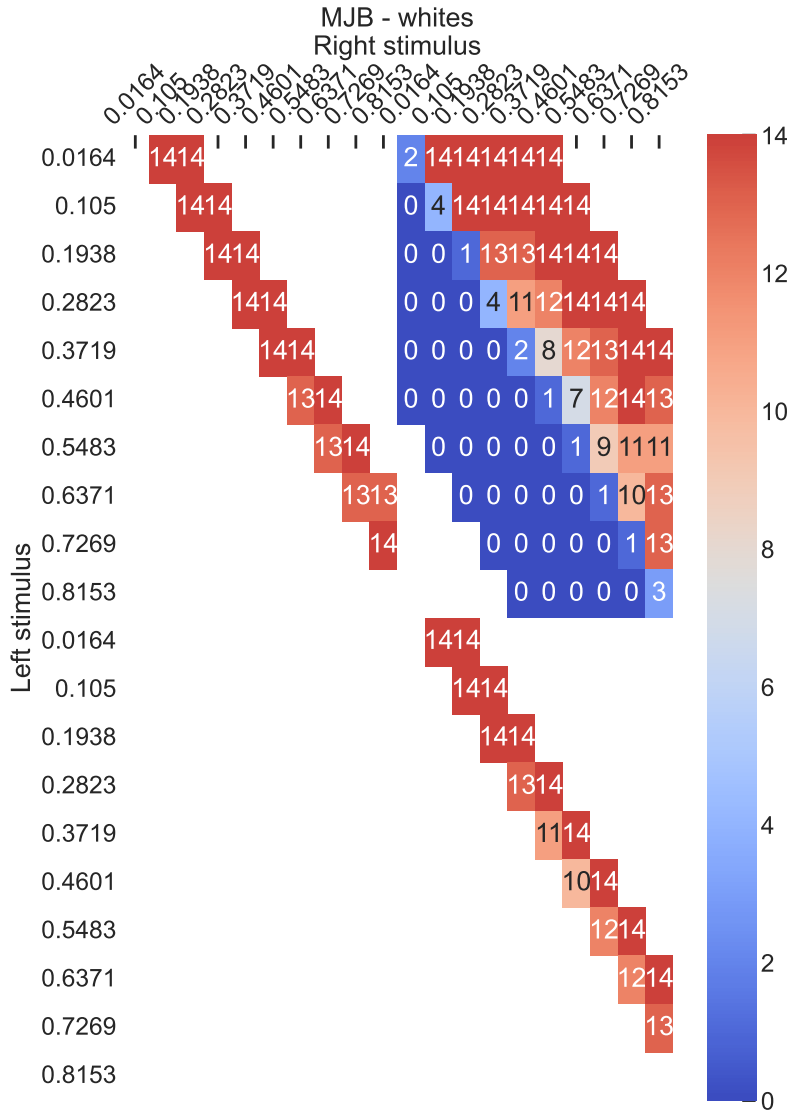

Figure 30: Participant MJB: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

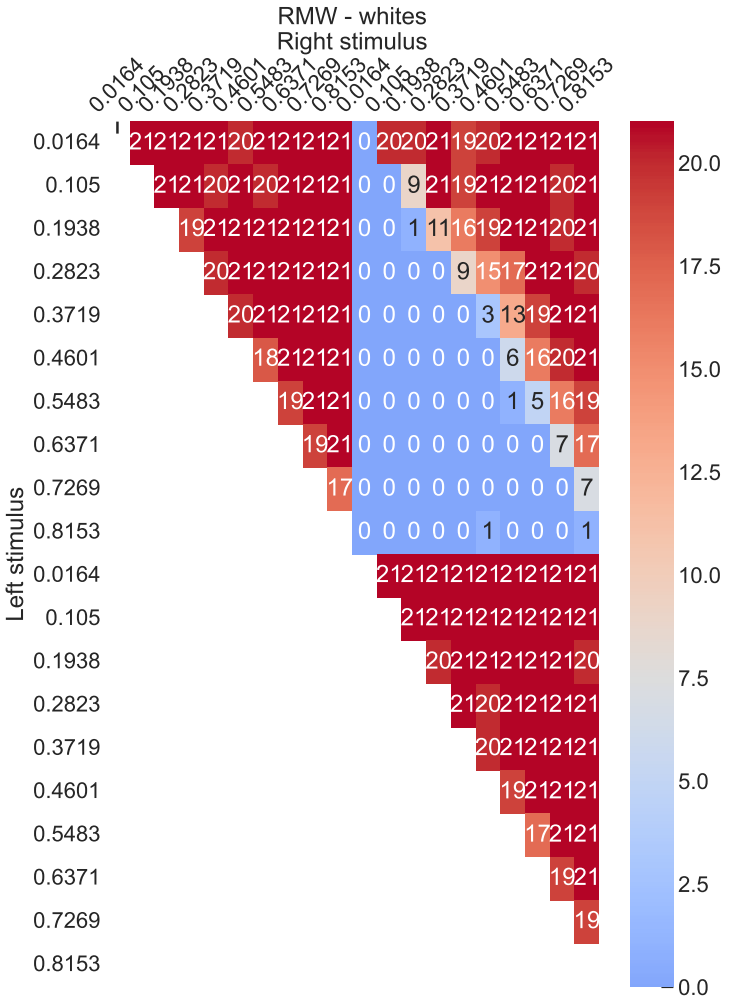

Figure 31: Participant RMW: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

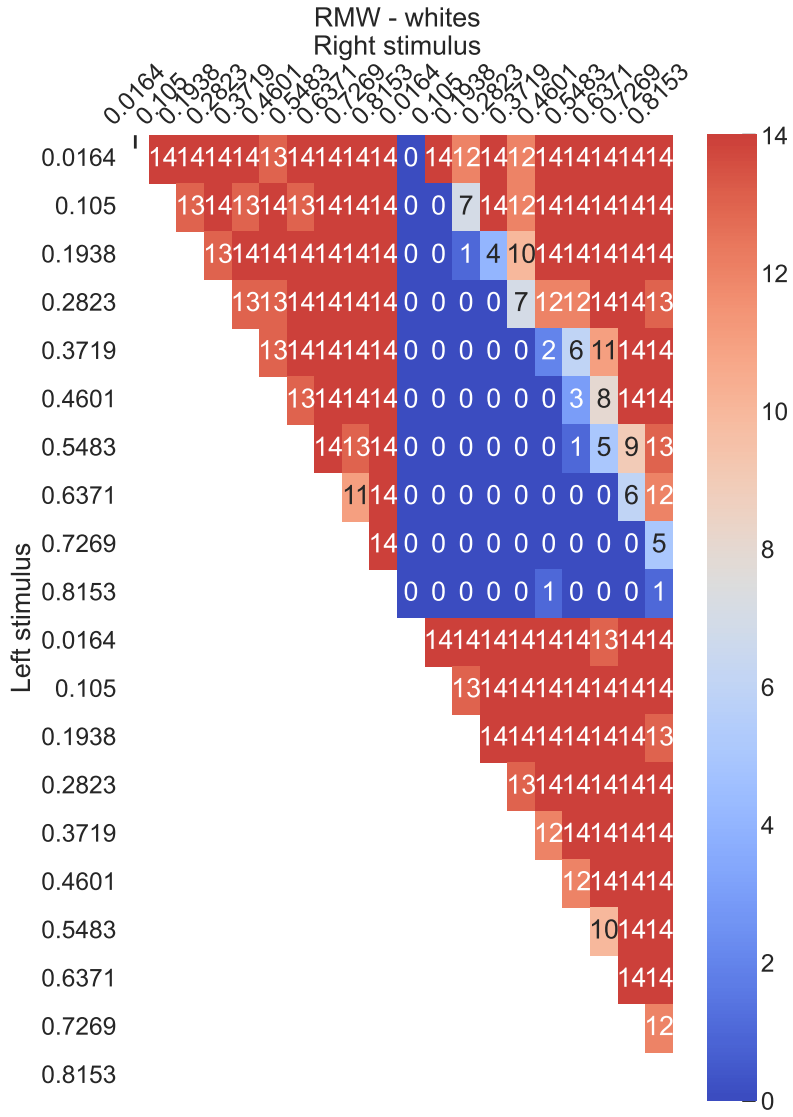

Figure 32: Participant RMW: Heatmap of absolute frequencies of choice in control condition.

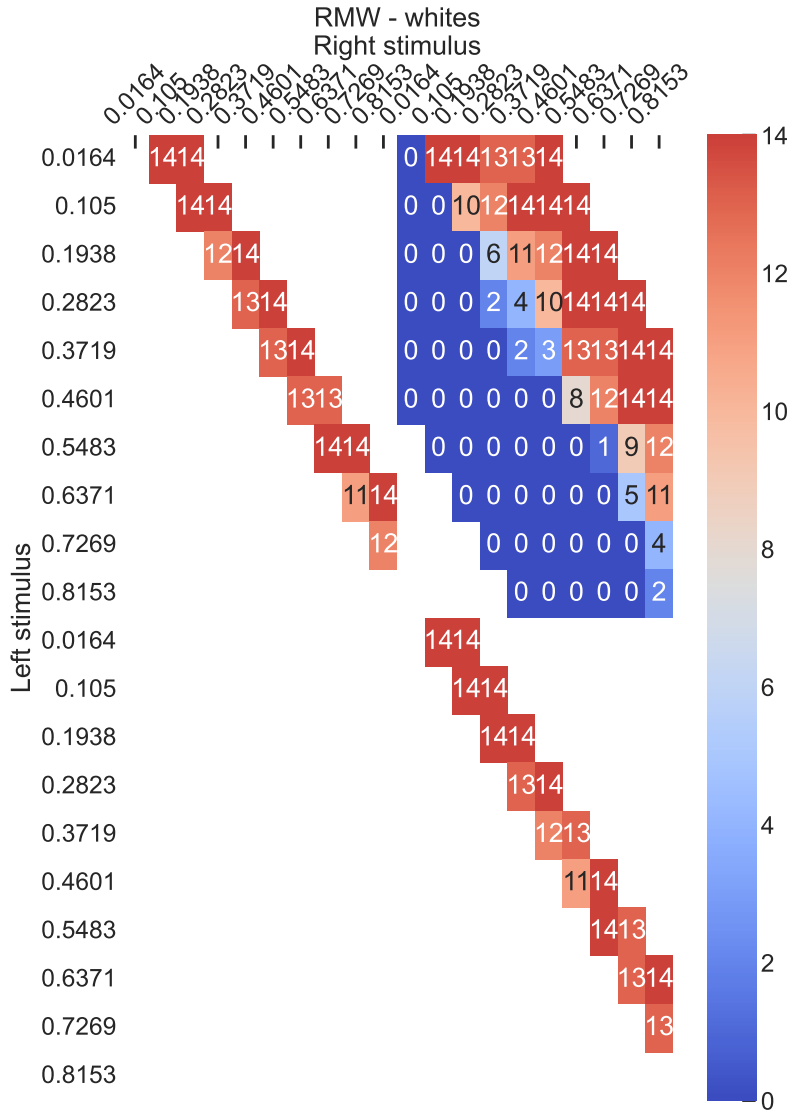

Figure 33: Participant RMW: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

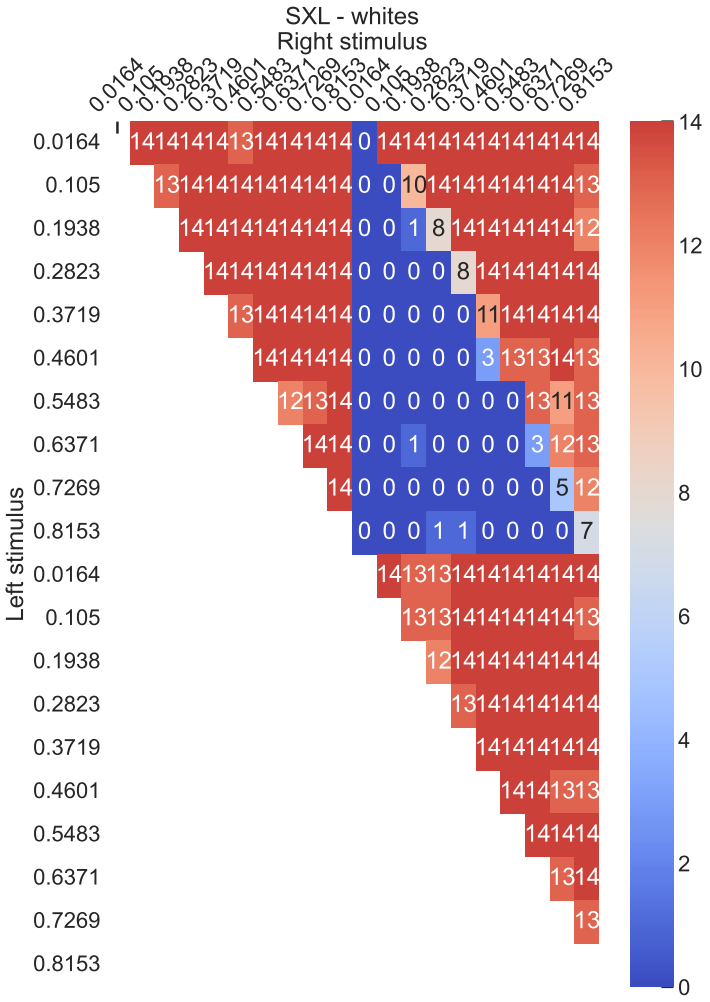

Figure 34: Participant SXL: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

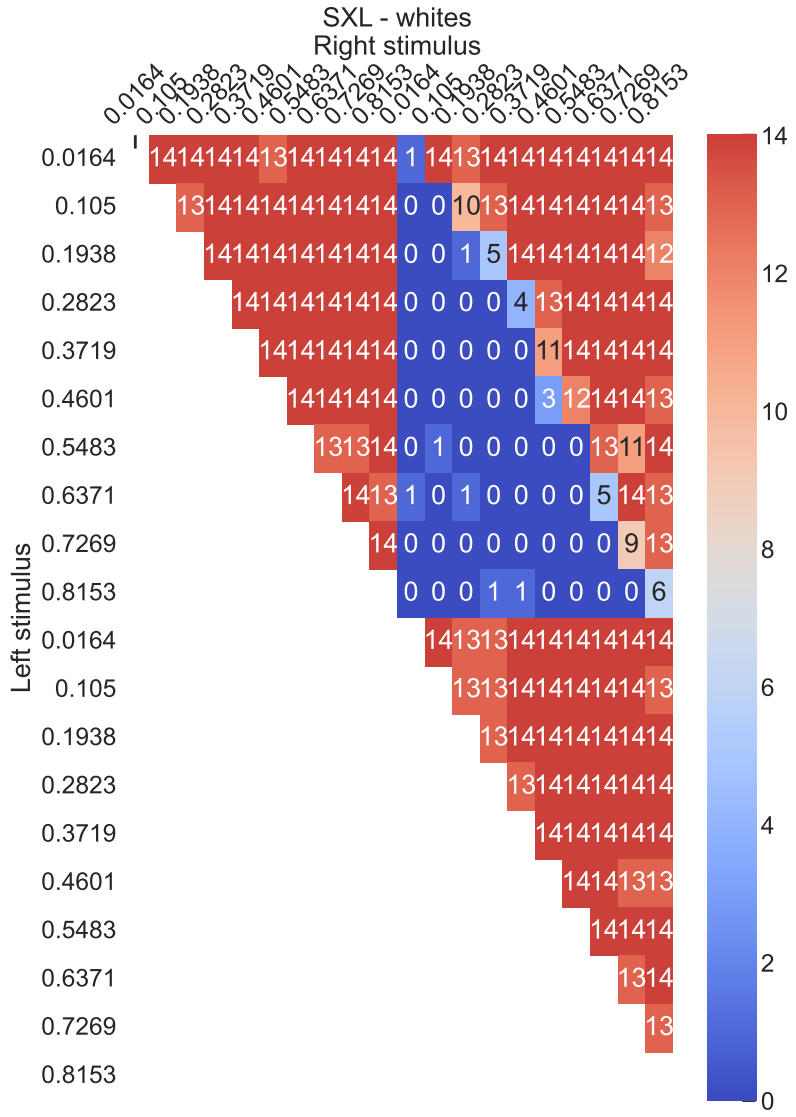

Figure 35: Participant SXL: Heatmap of absolute frequencies of choice in control condition.

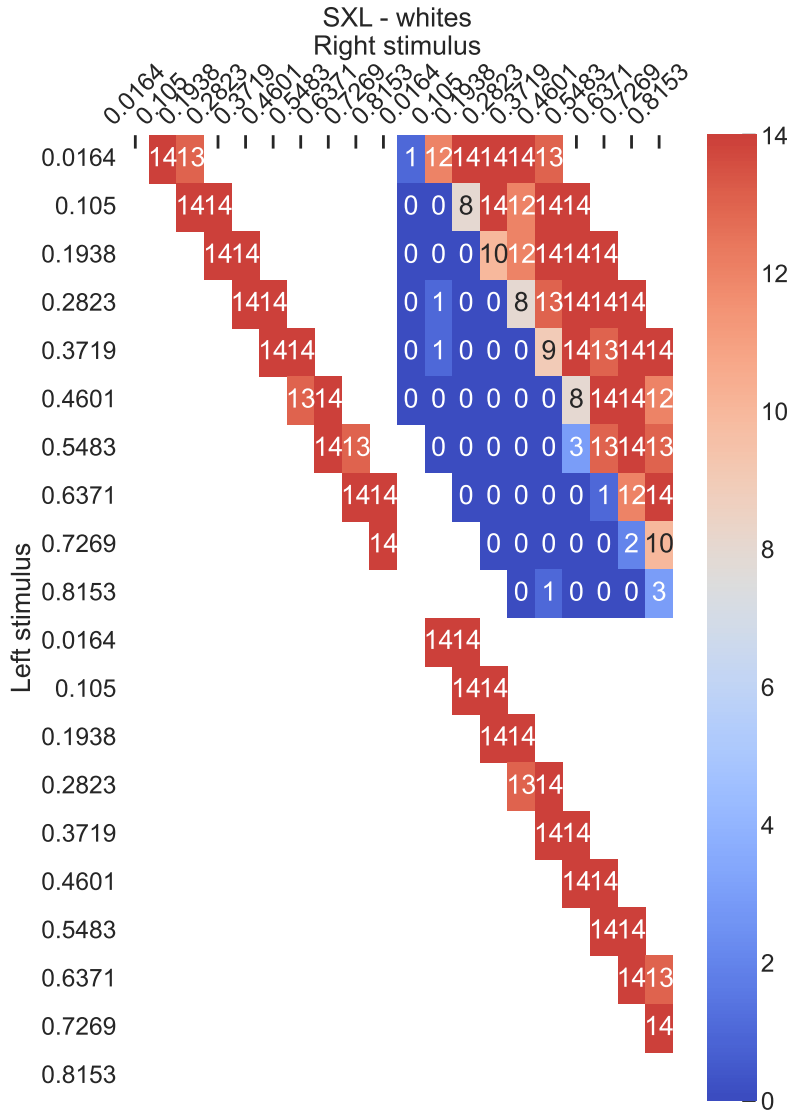

Figure 36: Participant SXL: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.

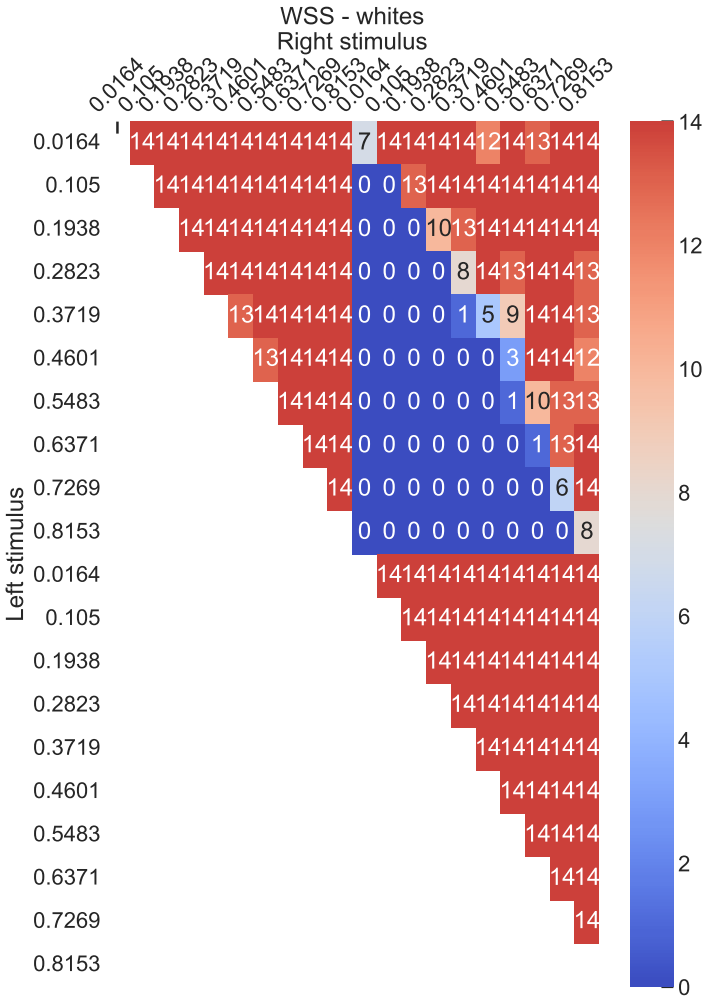

Figure 37: Participant WSS: Heatmap of absolute frequencies of choice in Runtime (Dynamic) Sampling condition.

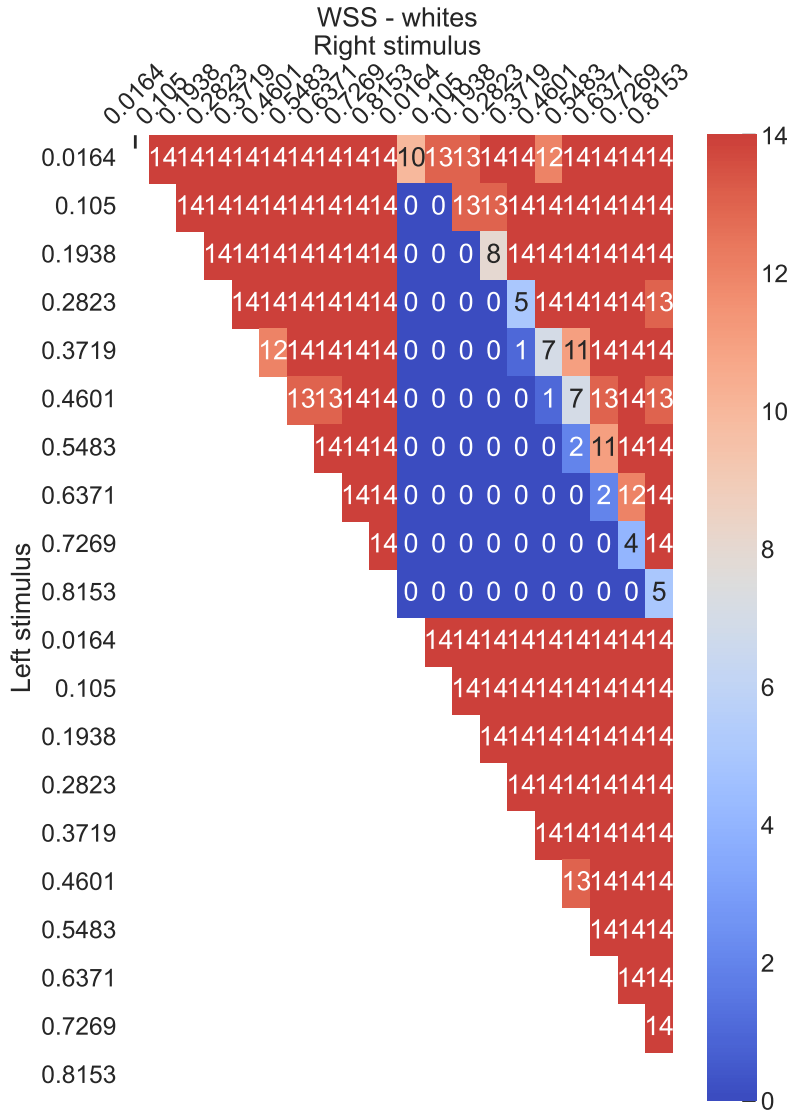

Figure 38: Participant WSS: Heatmap of absolute frequencies of choice in control condition.

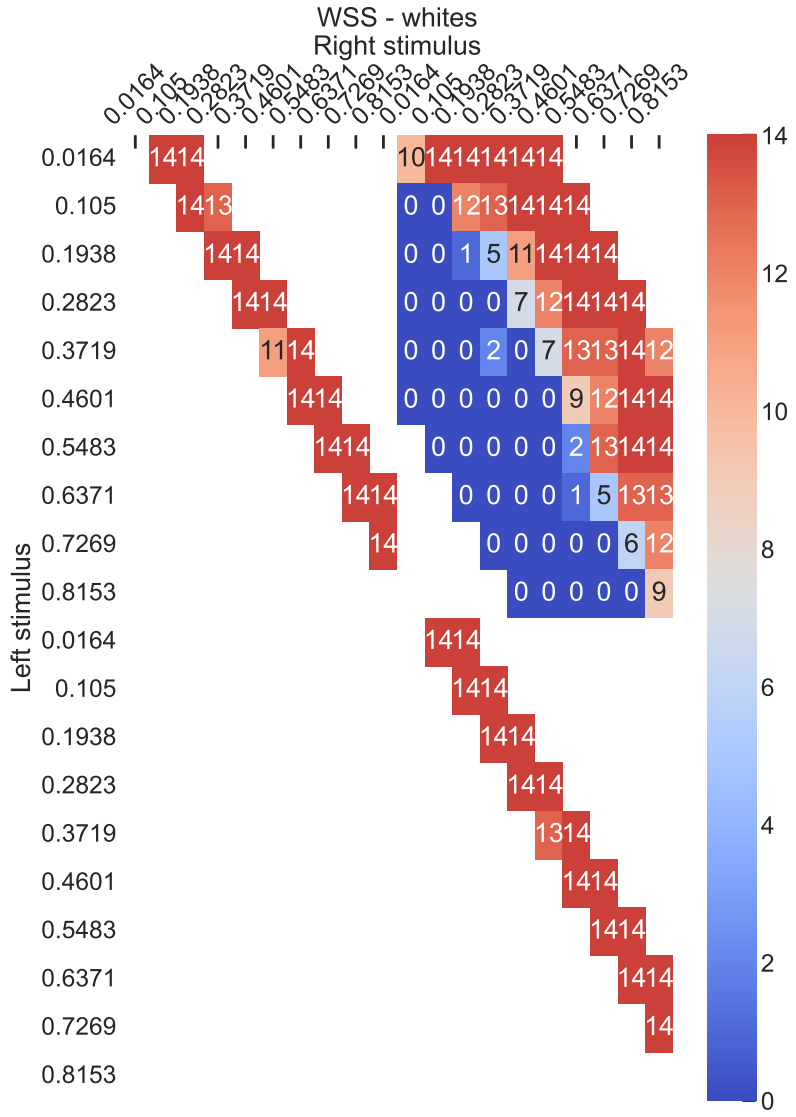

Figure 39: Participant WSS: Heatmap of absolute frequencies of choice in *a priori* (Static) Sampling condition.