

2023

Investigating Letter Identification for Visual Acuity Measurements in the Paracentral Visual Field

Barhoom, Hatem

<https://pearl.plymouth.ac.uk/handle/10026.1/21271>

<http://dx.doi.org/10.24382/5092>

University of Plymouth

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

Copyright Statement

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.



**UNIVERSITY OF
PLYMOUTH**

**Investigating Letter Identification for Visual Acuity Measurements
in the Paracentral Visual Field**

by

Hatem Barhoom

A thesis submitted to the University of Plymouth in partial fulfilment for
the degree of

Doctor of Philosophy

School of Health Professions

June 2023

Author's Declaration

I hereby declare that:

- At no time during the registration for the degree of Doctor of Philosophy have I been registered for any other university award without the prior agreement of the Doctoral College Quality Sub-Committee.
- The work submitted for this research degree at the University of Plymouth has not formed part of any other degree, either at the University of Plymouth or at another institution.
- This degree was financed with the aid of a scholarship from Qatar University and Islamic University in Gaza.
- The work presented in this thesis is my own, except for quotations and summaries, which have been duly acknowledged.
- Referred Conferences and Publications:
 1. Conferences talks and poster presentations.

Barhoom, H., Joshi, M. R., & Schmidtman, G. (2020). The effect of response biases on resolution thresholds of Sloan letters in central and paracentral vision. The British Congress of Optometry and Vision Science (BCOVS) 2020 (Talk).

Barhoom, H., Georgeson, M. A., Joshi, M. R., Artes, P. H., Schmidtman, G. The role of bias, sensitivity and similarity in letter identification task: a noisy template model. Applied Vision Association (AVA) 2023 (poster).

Barhoom, H., Schmidtman, G., Joshi, M. R., Artes, P. H., & Georgeson, M. A. (2022). The role of similarity and bias in letter acuity measurements: a noisy template model. European Conference on Visual Perception (ECVP) 2022 (poster),

Barhoom, H., Schmidtman, G., Joshi, M. R., Artes, P. H., & Georgeson, M. A. (2021). The role of bias in a letter acuity identification task: a noisy template model. ECVF 2021 (poster).

2. Journal Publications.

Barhoom, H., Joshi, M. R., & Schmidtman, G. (2021). The effect of response biases on resolution thresholds of Sloan letters in central and paracentral vision. Vision Research. Vol. 187, pp. 110 - 119.

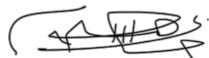
DOI: <https://doi.org/10.1016/j.visres.2021.06.002>

PEARL (OA): <http://hdl.handle.net/10026.1/17433>

Georgeson, M. A., **Barhoom, H.,** Joshi, M. R., Artes, P. H., & Schmidtman, G. (2023). Revealing the influence of bias in a letter acuity identification task: A noisy template model. Vision Research. Vol 208, pp. 108 – 233.

DOI: <https://doi.org/10.1016/j.visres.2023.108233>

Hatem Barhoom



June / 2023

Word count: 4518

Acknowledgements

First and foremost, I would like to express my gratitude to Almighty Allah for all his blessings and for granting me patience and good health throughout the duration of this research.

I am fortunate to have had Prof Paul H. Artes (previous director of studies), Drs Gunnar Schmidtmann (current director of studies and first supervisor), and Mahesh R. Joshi (second supervisor) as my mentors.

I would also like to extend my sincere appreciation to our collaborator, Prof Mark A. Georgeson, for his invaluable contributions and support during my study.

Furthermore, I would like to thank all my colleagues and the staff of the Optometry department at the School of Health Professions for their assistance, and camaraderie, and for fostering a pleasant working environment throughout my study period at the University of Plymouth.

Finally, I am immensely grateful to Qatar University and the Islamic University in Gaza for their sponsorship of my studies at the University of Plymouth.

Abstract

Investigating letter identification for visual acuity measurements in the paracentral
visual field

by

Hatem Barhoom

Sloan letters are commonly used optotypes in clinical practice, but they exhibit different relative legibility, which may be attributed to response bias, sensitivity differences, and letter similarity. In this thesis, we employed Luce's choice model and developed a new noisy template model to investigate the role of response bias, sensitivity differences, and letter similarity in letter identification of Sloan letters at central and paracentral locations. Results show that the best model was the one that accounted for the effects of bias, sensitivity, and similarity, with bias contributing more than sensitivity and similarity. However, when estimating the letter acuity from the pooled data across all letters, no significant effects of bias, sensitivity, or similarity were observed. The models incorporating similarity demonstrated a substantial increase in the spread of the underlying psychometric function (the percent correct as a function of letter size), particularly in the periphery and upper portion of the function. Given that most letter stimuli in clinical vision tests are presented at supra-threshold sizes, it is plausible to attribute any increase in test-retest variability, particularly in peripheral vision, to similarity alone.

Furthermore, this thesis explored the use of letters as stimuli and speech as a response method to assess macular visual sensitivity in healthy observers and individuals with

Glaucoma. Dissimilar letters following Sloan's design were used to estimate peripheral letter acuity within 4 degrees of fixation. A speech recognition algorithm was employed to enable participants to perform the task without supervision. The participants' perceived task difficulty was assessed through a questionnaire.

Results from this experiment show that in observers with Glaucoma, letter acuity exhibited a close correlation with conventional perimetry, and most observers found the task easy to perform. These results demonstrate that letter acuity perimetry with speech input is a viable method for capturing macular damage in Glaucoma. These approaches have the potential to facilitate more intuitive and patient-friendly tests for macular visual field assessments.

Table of Contents

Copyright Statement	1
Thesis Title	2
Author's Declaration	3
Acknowledgements	5
Abstract	6
Table of Contents	8
List of Tables	14
List of Figures	15
List of Abbreviations	18
Chapter 1 . Research Background	19
1.1 Motivation	19
1.2 Visual acuity definition	20
1.3 Vision charts and optotypes evolution	20
1.4 Optotypes standardisation	22
1.5 Letters as optotypes	27
1.5.1 The superiority of letters as optotypes	27
1.5.2 Letter identification task	28
1.5.2.1 Letter features in letter identification task	29
1.6 Visual acuity at the centre and periphery	31
1.7 Letter acuity at the centre and periphery	33

1.8 Source of observed differences in letter acuity	-----36
1.9 Methods to analyse the source of errors in letter identification tasks	----36
1.9.1 The effect of sensitivity, response bias and similarity on the estimated letter acuity	-----38
1.10 Letter acuity measurements in the paracentral visual field	-----39
1.10.1 Glaucoma definition	-----40
1.10.2 Pathophysiology and macular involvement in Glaucoma	-----40
1.10.2.1 Retinal ganglion cells death in Glaucoma	-----40
1.10.2.2 Macular damage in Glaucoma	-----41
1.10.3 Macular damage assessment in Glaucoma	-----41
1.10.3.1 Clinical macular visual field tests	-----43
1.10.3.2 Monitoring of macular damage progression in Glaucoma	-----45
1.10.3.3 Methods used to monitor the damage progression at the macula	----45
1.10.4 Factors that determine an efficient test for macular damage monitoring	46
1.10.4.1 The visual task with the strongest structure-function relationship	---47
1.10.4.2 Easy to access test for patients (home-based vs clinic-based)	-----48
1.10.4.3 Available home-based visual field tests	-----49
1.10.4.4 Home-based tests that employ resolution task	-----51
1.10.5 Suggestions for a new approach of home-based test for macular damage monitoring	-----52
1.10.5.1 Visual task to be employed	-----52
1.10.5.2 Response method to be employed	-----53
1.11 Problem statement and objectives	-----53
1.11.1 Problem statement	-----53

1.11.2 Objectives-----	54
Chapter 2 . The Effect of Response Bias and Similarity on Letter Acuity	
Measurements of Sloan Letters in Central and Paracentral Vision -----	55
2.1 Motivation-----	55
2.2 Methods-----	55
2.2.1 Observers-----	55
2.2.2 Stimuli -----	56
2.2.3 Apparatus-----	57
2.2.4 Software -----	57
2.2.5 Procedure-----	57
2.2.5.1 Data collection-----	57
2.2.5.2 Letter acuity measurement-----	59
2.2.5.3 Estimation of letters response bias and similarity between letters ----	60
2.3 Results-----	62
2.4 Discussion -----	74
Chapter 3 . The Role of Bias, Sensitivity and Similarity in Letter Acuity	
Measurements: A Noisy Template Model-----	80
3.1 Motivation-----	80
3.2 Methods-----	81
3.2.1 The noisy template model (NTM) -----	81
3.2.1.1 Letter usage -----	81
3.2.1.2 Outline of the model -----	83
3.2.1.3 Model structure and equations -----	86

3.2.1.4 The MAX operator-----	89
3.2.1.5 Letter identification-----	90
3.2.1.6 Illustration of the effect of bias, sensitivity and similarity on letter identification task -----	91
3.2.1.7 Model fitting-----	95
3.3 Results-----	96
3.3.1 Initial observations-----	96
3.3.2 Formal model comparison -----	100
3.3.3 The relative role of bias, sensitivity, and similarity in the letter identification task -----	105
3.3.3.1 The effect of bias, sensitivity, and similarity on letter acuity measurement-----	112
3.3.3.2 The effect of bias, sensitivity, and similarity on the spread of the underlying PMF -----	113
3.3.3.3 Simulations to examine the effect of bias, sensitivity, and similarity on letter acuity measurements and variabilities-----	115
3.4 Discussion -----	118
3.4.1 Best model -----	118
3.4.2 Validating the NTM using Luce’s choice model -----	119
3.4.2.1 Response bias-----	120
3.4.2.2 Similarity between letters -----	120
3.4.3 Source of error in similarity estimation using Luce’s choice model -----	121
3.4.4 Does the random error mimic the effect captured in experimental data? -----	126
3.4.4.1 Similarity between letters -----	126

3.4.4.2 Letter usage differences -----	128
3.4.5 The effect of bias, sensitivity and similarity on letter acuity measurement and test-retest variability-----	131
3.4.6 Conclusion -----	133
Chapter 4 . Acuity Perimetry with Speech Input for Mapping Macular Visual Field in Glaucoma -----	134 -
4.1 Motivation-----	134 -
4.2 Methods-----	135
4.2.1 Observers-----	135
4.2.2 Stimuli -----	135
4.2.3 Apparatus-----	136
4.2.4 Software -----	136
4.2.5 Procedure-----	137
4.2.5.1 Data collection-----	137
4.2.5.2 Observers' experience of the experiment -----	140
4.2.5.3 Letter acuity measurements-----	141
4.3 Results:-----	142
4.3.1 Letter acuity measurements-----	142
4.3.2 Observers' experience of acuity perimetry -----	147
4.3.3 The reliability of the speech recognition algorithm -----	150
4.3.4 Reliability of FSS staircase -----	151
4.3.4.1 Performance of FSS staircase -----	151
4.3.4.2 Test–retest variability-----	152
4.3.5 The effect of the lapse rate-----	155

4.3.5.1 Lapse rate in the experimental data-----	155
4.3.5.2 Simulating the effect of lapsing on letter acuity measurements -----	157
4.4 Discussion -----	169
4.4.1 Letter acuity measurements-----	169
4.4.2 Observers' experience of the acuity perimetry-----	170
4.4.3 Performance of the speech recognition algorithm -----	170
4.4.4 The rationale for employing the FSS staircase in the current study -----	171
4.4.5 Sources of test-retest variability in the current study-----	172
4.4.5.1 Lapse rate as a primary cause of increased test-retest variability----	172
4.4.5.2 Other potential sources of increased test-retest variability -----	175
4.4.6 Conclusion -----	175
Chapter 5 . Summary -----	177
5.1 General discussion-----	177
5.2 General conclusion and future work-----	179
Appendix A-----	180 -
Appendix B-----	191
Supporting Documents-----	201
References -----	230

List of Tables

Table 1.1 The percent correct of Sloan letters at the resolution threshold (Sloan, 1959).	35
Table 2.1 Normality test results for individual letter data.	63
Table 2.2 shows the central tendency and distribution of individual letter acuities (i.e., letter size in min or arc) across observers at the three locations.	65
Table 3.1 Comparison of the eight models via AIC analysis, for the 10 observers	101
Table 3.2 Group analysis of deviance, 10 observers	102
Table 3.3. Comparison of the seven models via AIC analysis, for the 10 observers	103
Table 3.4 Comparison of the four models via AIC analysis, for the 10 observers	104
Table 3.5 Comparison of the three models via AIC analysis, for the 10 observers	104
Table 3.6 Letter acuity (letter stroke width in min of arc (')) calculated from the eight fitted models and averaged across observers \pm SE.	112
Table 3.7 Spread of the model psychometric functions (σ) averaged across observers (\pm SE), for each of the eight models.	114
Table 3.8 Average of the upper spread of model psychometric functions (from 0.56 to 0.99 proportion correct) across observers (\pm SE).	114
Table 3.9 Average of the lower spread of model psychometric functions (from 0.11 to 0.54 proportion correct) across observers (\pm SE)	115
Table 4.1 shows the median, 25th & 75th percentile quartiles, and the statistical test results of comparing scores of responses to 10 items of the questionnaire in healthy and Glaucoma observers.	149
Table 4.2 The calculated lapse at different letter sizes that are larger than letter acuity.	156

List of Figures

Figure 1.1 The minimum angle of resolution.	20
Figure 1.2 A) Tobias Mayer's (1754) patterns and B) Stampfer's (1834) bars.	21
Figure 1.3 The world's first vision chart.	22
Figure 1.4 The first Snellen chart.	23
Figure 1.5 The early Landolt C optotypes.	24
Figure 1.6 A) John Green's (1868) and B) Bailey and Lovie's (1976) vision charts.	25
Figure 1.7 A) the design of the 10 Sloan letters and B) the ETDRS chart.	27
Figure 1.8 The resolution limit of different eye components.	32
Figure 1.9 The change in letter size as a function of eccentricity.	34
Figure 1.10 Differences in measured letters acuities at fovea and periphery.	35
Figure 2.1 The test locations and the stimulus design.	56
Figure 2.2 The sequence of the events in the trial in the experiment.	59
Figure 2.3 The average proportion correct and PMF across observers.	64
Figure 2.4 Measured acuities and responses for individual letters.	66
Figure 2.5 Correlation between letter usage and measured acuity.	68
Figure 2.6 The correlation in the proportion of correct vs incorrect responses.	69
Figure 2.7 The confusion matrices of presented vs responded letters.	70
Figure 2.8 Results of Luce's model fitting.	70
Figure 2.9 Luce's similarity parameter (η) for each letter pair at each location.	71
Figure 2.10 The letter confusability at the three locations.	72
Figure 2.11 Correlation between Luce's parameters and letter acuity.	73
Figure 2.12 Measured letter acuities of the current and previous studies.	74
Figure 2.13 The similarity of letter pairs of the current and previous studies.	76
Figure 2.14 The Luce bias of the current and Hamm et al.'s (2018) studies.	78
Figure 3.1 Data overview: letter usage.	82
	15

Figure 3.2 A) a schematic representation of the NTM and (B) template responses.	84
Figure 3.3 Demonstration of NTM in case of bias, sensitivity, similarity or none.	85
Figure 3.4 The optotype correlation matrix (OC) for Sloan letters.	87
Figure 3.5 Simulation of bias effect on letter identification performance.	92
Figure 3.6 Simulation of sensitivity effect on letter identification performance.	93
Figure 3.7 Simulation of similarity effect on letter identification performance	94
Figure 3.8 The experimental data (presented vs responses of letters).	97
Figure 3.9 The variants of the NTM model estimated from experimental data.	99
Figure 3.10 The best model recomputed with different activated factors.	108
Figure 3.11 Luce's parameters of the best model vs experimental data.	111
Figure 3.12 Bias and similarity of Luce's model vs the best model.	111
Figure 3.13 The effect of bias, sensitivity, and similarity on letter acuity.	118
Figure 3.14 The influence of bias on Luce's similarity calculation.	123
Figure 3.15 The effect of bias and sensitivity on the Luce's similarity (0 deg).	124
Figure 3.16 The effect of bias and sensitivity on the Luce's similarity (± 3 deg).	125
Figure 3.17 Frequency of estimated Cs from simulated and experimental data.	127
Figure 3.18 Simulated differences in letter usage (total) with different conditions.	129
Figure 4.1 The fixation cross and the test locations (grey circles).	138
Figure 4.2 The sequence of trials in the experiment.	139
Figure 4.3 Results of acuity perimetry for healthy eyes.	142
Figure 4.4 Results of acuity perimetry and M-pattern test for glaucomatous eyes.	143
Figure 4.5 Corresponding test points of the acuity and the M-pattern perimeters.	144
Figure 4.6 The pointwise agreement of the letter acuity vs contrast sensitivity.	145
Figure 4.7 The correlation of letter acuity vs contrast sensitivity at each quadrant.	146
Figure 4.8 The correlation of letter acuity vs contrast sensitivity at all locations.	146
Figure 4.9 Observers' responses to the questionnaire.	148
Figure 4.10 The number of letters presented (i.e., validated) vs responded to.	151

Figure 4.11 The number of reversals in all staircases.	152
Figure 4.12 The agreement in letter acuity between the two sessions.	153
Figure 4.13 The Bland-Altman plots at each test location.	154
Figure 4.14 The effect of lapsing on letter acuity for inner eccentricities.	159
Figure 4.15 The effect of lapsing on letter acuity for outer eccentricities.	160
Figure 4.16 The effect of step-up size on the SDC for different lapses.	162
Figure 4.17 The effect of lapsing on SDC for different ground truths.	163
Figure 4.18 Effect of lapse on letter acuity (by FSS) for different ground truths.	164
Figure 4.19 Effect of lapse on letter acuity (by FSS) for different step sizes and slopes.	164
Figure 4.20 Effect of lapse on letter acuity (by ZEST) for different ground truths.	166
Figure 4.21 Effect of lapse on the SDC (by ZEST) for different ground truths.	167
Figure 4.22 Effect of lapse on letter acuity and SDC (by ZEST) for different conditions.	168
Figure 4.23 Measured letter acuity of the current and two previous studies.	169
Figure 4.24 Effect of lapsing on letter acuity and SDC (by ZEST) for different λ_{ZEST} .	174

List of Abbreviations

logMAR	Logarithm of the Minimal Angle of Resolution
ETDRS	Early Treatment of Diabetic Retinopathy Study
AFC	Alternative Forced Choice
NTM	Noisy Template Model
PMF	Psychometric Function
OC	Optotype Correlation
LL	Log-Likelihood
AIC	Akaike Information Criterion
NS	Nasal Superior
NI	Nasal Inferior
TS	Temporal Superior
TI	Temporal Inferior
FSS	Fixed Step Size
ZEST	Zippy Estimation by Sequential Testing
SDC	Smallest Detectable Change

Chapter 1 . Research Background

1.1 Motivation

This chapter will discuss the rationale for using letters as optotypes to measure visual acuity in the paracentral visual field and the significance of using letter acuity measurements to map paracentral (i.e., macular) damage in a clinical population, focusing specifically on Glaucoma. Firstly, the evolution of letters as optotypes in clinical vision charts and a brief history of the adoption of different optotypes will be presented, with letters being the primary focus. Additionally, the visual perception of letters and the sources of errors in letter identification tasks will be discussed. The measurement of visual acuity using letters as optotypes and the sources of differences in letter acuities between individual letters at central and peripheral visual fields will also be addressed. Finally, the potential superiority of using letter acuity measurements to map macular damage in Glaucoma will be illustrated. At the end of this chapter, the objectives of the experiments conducted in the current thesis will be determined.

1.2 Visual acuity definition

Visual acuity is a clinical measurement that quantifies the ability of the visual system to discriminate the finest details, typically measured as the minimum angle of resolution of an object or a structurally standardised target (i.e., optotype) from a certain distance (e.g., six meter) (Figure 1.1).

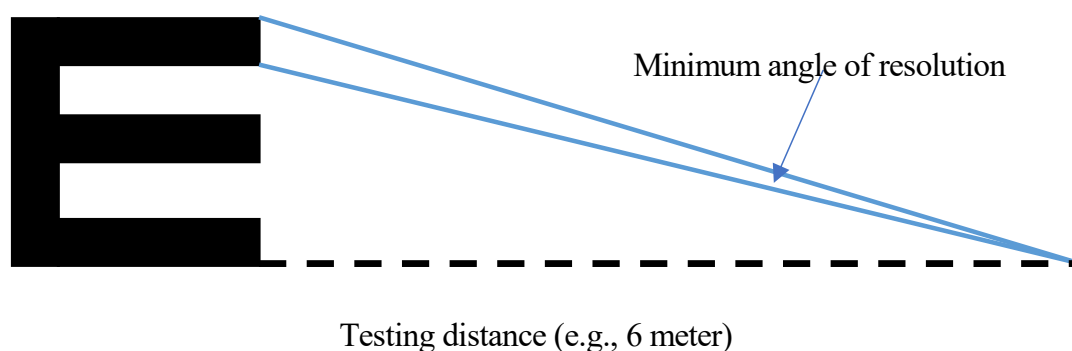


Figure 1.1 The minimum angle of resolution.

1.3 Vision charts and optotypes evolution

One of the first objects in ancient history that were used as optotypes were stars within the constellations of the night sky. For example, ancient Persians used the double star of the Big Dipper in the Big Bear to assess the vision of warriors and was called the "riddle". Another example is the "Arab Eye Test", which was the name of the test used by the desert Arabs 1000 years ago to assess the ability to perceive the separation of the double stars Alcor and Mizar (Colenbrander, 2008; Bohigian, 2008). It was used to test children's eyesight. Both measures were equivalent to current normal standard for visual acuity measurement (6/6) (Colenbrander, 2008; Bohigian, 2008). Similarly, in 1623, Daza de Valdes used the ability to see and resolve mustard seeds thrown on a

table to assess visual acuity (Colenbrander, 2008; De Jong, 2022). In 300 BCE, the concept of the minimum angle of resolution was first introduced by Euclid (*fl.* 300BC), an ancient Greek mathematician who lived in Alexandria. His theory stated that the rays travel in straight lines and diverge from the eye forming a cone and that we can only see objects encompassed by the cone base (emission theory). In a Latin translation (AD 1350) of Euclid's Greek book, the term "minimum angle" first appeared. However, around AD 1000 the Arab scientists Ibn-Sina and Alhazen disproved the emission theory (De Jong, 2022).

In his nice review, De Jong (2022) outlined the first attempts to introduce structured and standardised optotypes to measure visual acuity. For instance, in 1754, Tobias Mayer used various test objects such as dots, lines, and grids drawn on white paper to assess visual acuity at a certain distance (Figure 1.2 A). Stampfer in 1834 used plates with lines that showed different thicknesses and distances from each other to assess the smallest visual angle at a given distance (Figure 1.2 B).

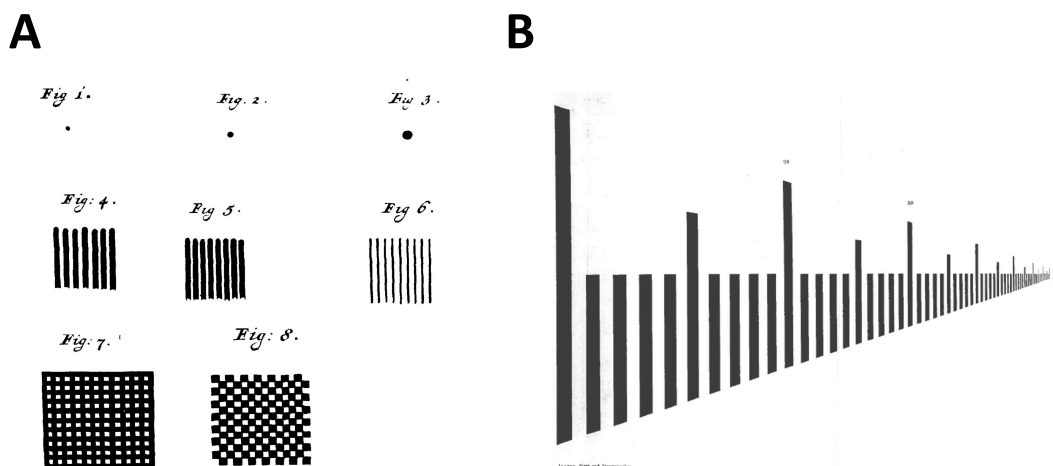


Figure 1.2 A) Tobias Mayer's (1754) patterns and B) Stampfer's (1834) bars. They assess the smallest visual angle (De Jong, 2022).

Küchler in 1835 was the first to use pattern optotypes (i.e., figures cut from almanacs and posted on cardboard). He realised the importance to standardise the visual acuity measurement and that inspired him to introduce the world's first optotypes (a chart with one word per line) with variable font sizes to test distance vision in 1843 (Figure 1.3) (Colenbrander, 2008; De Jong, 2022).



Figure 1.3 The world's first vision chart. It was introduced by Küchler in 1843. The chart has optotypes (one word per line) of variable font sizes to assess distance vision (De Jong, 2022).

1.4 Optotypes standardisation

Many charts and optotypes were introduced in an attempt to standardise the optotypes used in vision charts. The significant breakthrough occurred in 1861 when Donder introduced a formula and a reference standard to express visual acuity in terms of the ratio of letter size to viewing distance. Further work on Donder's findings was conducted by his Ph.D. student Herman Snellen. In 1862, Snellen introduced his first

vision chart (Figure 1.4). He used single letters as optotypes. In his optotype design, each letter subtends a visual angle of 5 minutes of arc where the visual angle = d/D in which d is the distance in Paris feet at which the letter is recognised, and D is the optotype size (Colenbrander, 2008; De Jong, 2022). The Snellen vision chart has been used in clinical measurements since then. Following Snellen's optotype design, in 1876, Monoyer introduced a decimal letters chart using distances in meters.

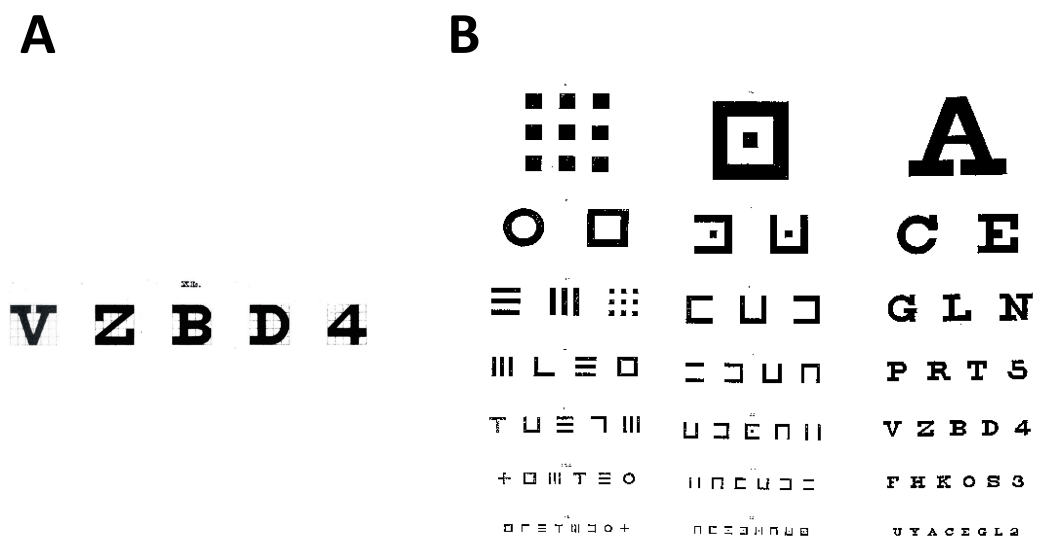


Figure 1.4 The first Snellen chart.

A) one line of the first Snellen chart (1862). B) Original Snellen charts (Colenbrander, 2008; De Jong, 2022).

However, all the attempts to standardise the optotype's design adopted optotypes that were not universally accepted nor equally legible. Küchler (1844) set three conditions for optotype's design. Firstly, test optotypes should be universally understandable. Furthermore, the optotypes should be equally legible and they should only vary in size. Lastly, the difference in sizes of subsequent optotypes should be as equal as possible. In 1899, Landolt introduced the "C" chart, and it was the first to fulfil all three

conditions (Figure 1.5) (Landolt, 1899; De Jong, 2022). The task in Landolt’s C chart is to name the direction of the gap in the C instead of recognising and naming the optotype. Currently, the Landolt C optotype is internationally regarded as the reference optotype for laboratory testing (Ferris & Bailey, 1996; Treacy et al., 2015).

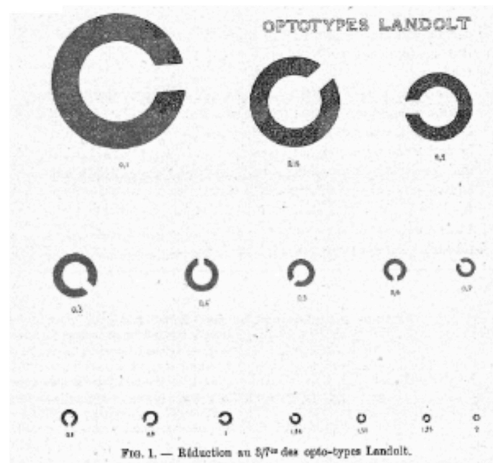


Figure 1.5 The early Landolt C optotypes.
The opening in the C has eight possible directions (Landolt, 1899; De Jong, 2022).

Bennett (1965), the chairman of the British Standards Institution sub-committee on ophthalmic test types pointed out that the vision charts (mainly based on Snellen’s design) had no comparable basis for estimating visual acuity. Despite the efforts made to standardise the chart design, there was a difference in approach on three fundamental aspects, which are (i) the adopted pattern and number of optotypes per line, (ii) the progression of sizes between them, and (iii) the visual acuity recording format.

However, in 1868 John Green introduced the first chart that addressed Bennett’s concerns (Figure 1.6). It was very similar in many aspects to today’s well-known Bailey and Lovie vision chart (Bailey & Lovie, 1976). For instance, they both have an equal number of letters per line (i.e., 10 letters per line in Green’s chart as opposed to

five in Bailey and Lovie’s chart) and proportional spacing between the letters of the same line, and geometrical progression of the subsequent lines. The reason for not accepting John Green’s vision chart at that time as a standardised vision chart is that it was too early for the need for such a standardised test (Colenbrander, 2008). The need for an accurate and more standardised vision test emerged with the increased interest in low-vision rehabilitation, which started in the second half of the 20th century. Consequently, the aim of vision tests shifted from being a tool to grossly differentiate between normal and abnormal vision to a tool for a precise measurement of the visual acuity for categorising, assessing, diagnosing, and monitoring different eye conditions (Colenbrander, 2008).

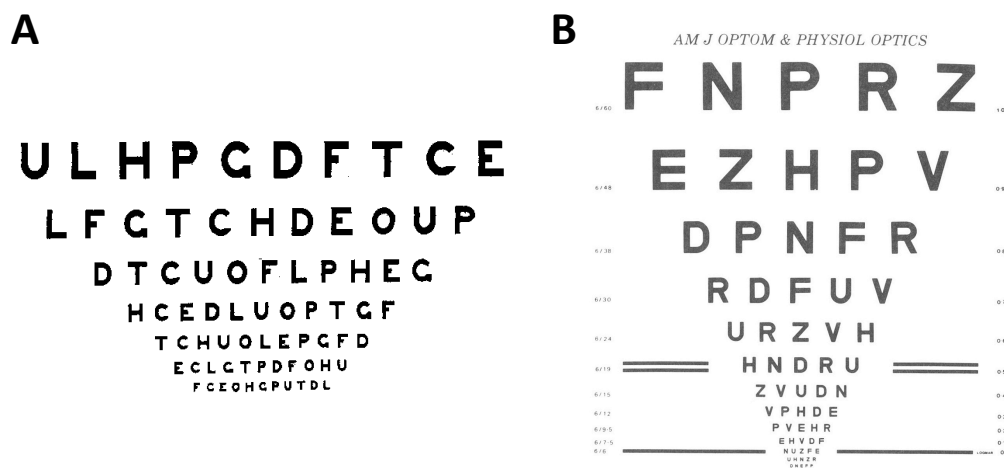


FIG. 1. New visual acuity letter chart.

Figure 1.6 A) John Green’s (1868) and B) Bailey and Lovie’s (1976) vision charts. (Bailey & Lovie, 1976; Colenbrander, 2008).

The progression of the stroke width of the subsequent optotypes can be geometrical (logarithmic) or arithmetical. Snellen was the first to approximate geometric progression for subsequent optotypes (i.e., each line is 1.25 - 1.50 times larger than the previous one). However, John Green used the geometric progression in his chart so

that each line increment represents a fixed ratio of $10^{\sqrt{10}}$ (i.e., 0.1 log unit). The logarithmic progression became popular in industrial standards and is known as the "preferred numbers series" (Colenbrander, 2008).

The visual acuity notation format was also not standardised. It was expressed in either an absolute (visual angle in minutes of arc, e.g., 2 minutes of arc) or a relative value (Snellen notation, e.g., 6/6 or 20/20). Like John Green's work, Bailey and Lovie addressed Bennett's criticisms and they were the first to use the Logarithm of the Minimal Angle of Resolution (logMAR) in visual acuity recording. Consequently, the logMAR chart was introduced (Bailey & Lovie, 1976; Colenbrander, 2008; De Jong, 2022).

Louise Sloan (1959) introduced new set of letter optotypes with a Snellen design (the height and width of a letter is five times its stroke width) and chose a 0.1 log unit as the geometrical magnification factor, resulting in the Sloan chart (Sloan, 1959; Colenbrander, 2008; De Jong, 2022). Sloan letters were designed and chosen so that their average measured visual acuity at the fovea differs by no more than 0.05 log units from the visual acuity measured with the Landolt C chart (i.e., the international reference optotype) (Figure 1.7) (Sloan, 1959; Treacy et al., 2015). In 1982, the National Eye Institute employed the 10 Sloan letters set in the Bailey and Lovie layout to produce charts for use in the Early Treatment of Diabetic Retinopathy Study (ETDRS) and it has been employed in the majority of subsequent clinical studies. The ETDRS format is the international gold standard today (Ferris et al., 1982).

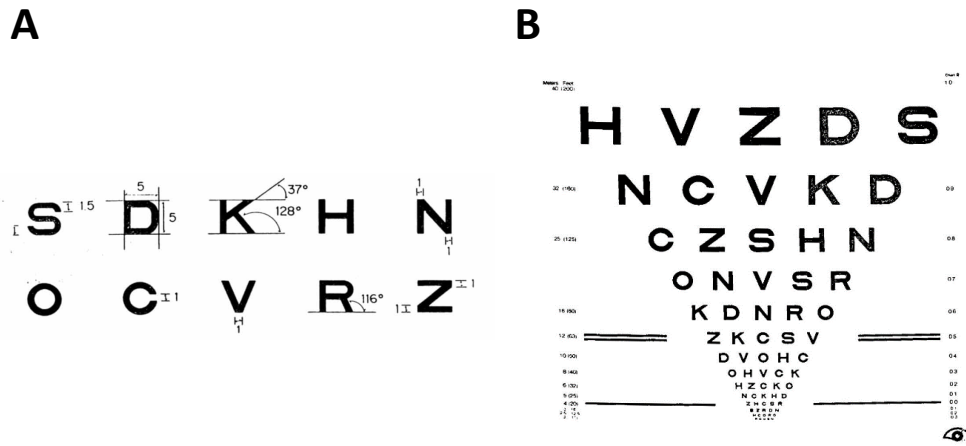


Figure 1.7 A) the design of the 10 Sloan letters and B) the ETDRS chart. (Sloan, 1959; Ferris et al., 1982).

1.5 Letters as optotypes

1.5.1 The superiority of letters as optotypes

Different types of optotypes can be used to estimate visual acuity, for example, symbols, shapes, numbers, or letters. The Tumbling E, Landolt C, and letter optotypes have been widely accepted as standard optotypes in clinical practice (Kniestedt & Stamper, 2003). However, letters are characterised by their familiarity and ease of use in clinical settings. Furthermore, unlike Tumbling E and Landolt C optotypes, letters are frequently encountered in daily life, making them intuitive and familiar for literate patients. Another useful feature is that most letter optotypes contain a variety of letters. For instance, the Sloan letter set contains 10 letters, which translates to a one interval 10 Alternative Forced Choice (AFC) test and reduces the guessing rate to 10% compared with other optotypes, such as the Tumbling E (4AFC), or the Landolt C (4AFC or 8AFC), where the guessing rate is 25% or 12.5%, respectively (Pelli & Robson, 1991). For these reasons, letters have been adopted in many visual acuity chart designs (i.e., ETDRS chart).

1.5.2 Letter identification task

Resolving letters is fundamentally a visual resolution function where the task is letter identification. Letter identification tasks have been extensively investigated in the previous studies. Pelli et al. (2006) investigated the letter identification efficiency for a wide range of observers' ages and under various conditions. It has been found that the letter identification task is inversely correlated with the complexity of the letter¹. Furthermore, the letter identification efficiency was unaffected by contrast, duration of presentation of the letter, and eccentricity from fixation, and only weakly dependent on the letters' size. Interestingly, letter identification efficiency increases quickly with experience and only three thousand trials were enough for a novice observer of any age to become fluent in identifying a given alphabet (Pelli et al., 2006). Pelli et al. (2006) concluded that consistent letter identification efficiency across a wide range of conditions suggests that the letter identification task is a fundamental visual process (i.e., not a language skill) (Pelli et al., 2006). In other words, to identify a letter, the process starts with identifying the *features* of the letter as opposed to whole-letter templates (Pelli et al., 2006).

It has been proposed that the letter features are detected by a set of feature detectors (i.e., channels) in the cortex (Graham, 1989). Majaj et al. (2002) investigated the effect of letter size on the selection of channels used to identify a letter. They found that channel frequency depends solely on the stroke frequency of the letter² where it is identified through a single channel (i.e., 1.6 ± 0.7 octaves wide). Identifying the letter

¹ The complexity of the letter is defined as inside-and-outside perimeter, squared, then divided by the ink area of the letter (Attneave & Arnoult, 1956).

² Defined as the average number of lines crossed by a cross-section through a letter, divided by the letter width.

through that particular channel leads to a high letter identification efficiency for the small letters (giving that it is not limited by visual acuity) when the channel frequency matches stroke frequency. In the case of large letters, however, the mismatch between the letter and channel frequencies leads to a drop in letter identification efficiency (Majaj et al., 2002; Pelli et al., 2006).

1.5.2.1 Letter features in letter identification task

Many previous studies proposed letters' features sets that are responsible for letter identification (Geyer, 1971; Gibson, 1969; Laughery, 1970). However, the basis for defining the features sets by these studies was hypothetical and contained substantial degrees of subjectivity, without empirical support (Geyer & DeWald, 1973; Holbrook, 1975). The empirical approach to determine the features sets was introduced by some investigators via analysing the confusion matrix of letter identification task (Townsend, 1971 a & b; Gilmore et al., 1979). The procedure starts by collecting data on letter identification tasks at the threshold level of a given visual task (typically at 50% correct performance of the underlying psychometric function (PMF) of the visual task, e.g., the contrast threshold for identifying a letter from noise). The data is collated as presented vs responded confusion matrix where each cell counts the number of times a given letter was chosen in response to the letter presented. Then, Luce's choice model is typically used to decompose the confusion matrix into a bias vector³ and letter similarity matrix (Luce, 1963). The letter similarity matrix undergoes further analysis

³ It has been claimed that response bias has a small role in the confusion matrix structure, therefore usually ignored (Loomis, 1982). In Chapter 3, I showed that this is not the case and in fact, response bias is the key factor that determines the differences in individual letter responses in letter identification tasks.

(e.g., multidimensional scaling) to extract the features that are common between similar letters. These features are considered the crucial factors that distinguish letters from each other and therefore are the key factors in letter identification (Townsend, 1971 a & b; Gilmore et al., 1979). Using this method, Gilmore et al. (1979) and Townsend (1971 b) identified five and four features respectively that are crucial for letter identification. However, this method would only define the perceptual structure and not the physical dimension (i.e., the features cannot be named) (Townsend, 1971 a & b; Gilmore et al., 1979). Using a different approach, Pelli et al. (2006) extended the model of detection: probability summation (Robson & Graham, 1981; Watson & Ahumada, 2005) to identify the number of features used in letter identification. They found that identifying one of 26 letters is based on 7 ± 2 features detections, independent of the complexity of the alphabet (Pelli et al., 2006).

Nevertheless, all previous studies either subjectively proposed the features (Geyer, 1971; Gibson, 1969; Laughery, 1970) or the features were estimated *indirectly* from the empirical data (Townsend, 1971 a & b; Gilmore et al., 1979). Fiset et al. (2008) derived the letter features *directly* from empirical data by manipulating the presented letters in the letter identification task using the "Bubbles" technique (a classification image technique) (Gosselin & Schyns, 2001; Fiset et al., 2008). Using this technique, it has been found that font terminations, i.e., clear discontinuities in bars and curves, are the key features for letter identification (Fiset et al., 2008). Using the same logic, the similarity between two given letters can be estimated by determining the common features of the two letters. For the work presented in this thesis, I assumed that the similarity between two letters is determined by the correlation in total physical measures (i.e., features) between the two letters (Pelli et al., 2006; Fülep et al., 2017). The correlation of total physical measures between two letters is calculated by

Pearson's normalised cross-correlation between undistorted letters (Neto et al., 2013). Therefore, letters that share common features (e.g., C and O) would show a higher correlation (i.e., similar letters) compared to the case of letters with less common features (e.g., V and D; dissimilar letters).

1.6 Visual acuity at the centre and periphery

It has long been known that visual acuity shows a progressive deterioration from the fovea toward the periphery of the retina (Ludvigh, 1941; Millodot et al., 1975). Subsequent studies showed that peripheral acuity is dependent on the meridian and eccentricity (Anderson, 2006). In his review, Anderson (2006) discussed the basis of why peripheral acuity differs from foveal acuity. The optical quality at the fovea is found to be better than that at the periphery. Moreover, the cone density decreases with increasing eccentricity (Curcio et al., 1987; Himmelberg et al., 2023). Therefore, the decrease in visual acuity could be attributed to the poor quality of the retinal image and/or to the retinal under-sampling (i.e., beyond the resolution limit of the retina) of the image at the periphery (Anderson, 2006; Ludvigh, 1941; Millodot et al., 1975). However, it has been shown that the deterioration in optical quality is slower than that of the retinal resolution limit as we move toward the periphery (Figure 1.8) (Navarro, 2009).

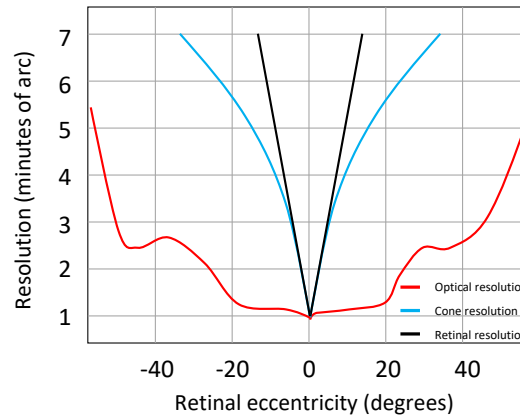


Figure 1.8 The resolution limit of different eye components.

The Figure shows the difference between the optical, cone sampling, and the effective retinal resolution in deterioration rate (from the fovea to the periphery) (Navarro, 2009).

In other words, at the periphery, the retinal image can be received through the optics of the eye in good quality, but if it is beyond the resolution limit of the retina (i.e., higher than the Nyquist frequency⁴) it would be under-sampled causing a phenomenon called aliasing (detected but not resolved), hence resolution at the periphery is limited by the retinal sampling capacity. The optics at the fovea works as an anti-aliasing filter because it will filter out the spatial frequencies that are beyond the resolution limit of the retina (Williams et al., 1996; Anderson, 2006). The sampling-limited nature of the resolution task at the periphery was further investigated in subsequent studies. It has been shown that the resolution acuity at the periphery is robust to the loss of contrast (up to a contrast of 10%) and optical defocus (up to 3 - 4 dioptres and some claim only up to 1 diopter) (Millodot et al., 1975; Thibos et al., 1996; Anderson, 1996b; Wang et al., 1997, Rosén et al., 2011). These lines of evidence suggest that the resolution threshold at the fovea is optically-limited whereas the resolution at the periphery is sampling-limited.

⁴ Nyquist frequency is the highest frequency that can be correctly represented by a given sampling array.

Further studies investigated the anatomical and physiological basis of the visual resolution reduction towards the periphery. For instance, Rossi and Roorda (2010) studied the relationship between the visual resolution of the Tumbling E and cone spacing at multiple locations at 0 to 2.5 degrees (deg) eccentricity from the fovea. At these relatively small eccentricities, it has been found that the visual resolution is well explained by the cone spacing at the fovea and up to 0.5 deg eccentricity only. Beyond that eccentricity, the midget retinal ganglion cell mosaic explains the reduction in visual resolution better than the cone spacing (Anderson, 2006; Rossi & Roorda, 2010).

1.7 Letter acuity at the centre and periphery

Similar to other optotypes used in previous studies (i.e., gratings or Tumbling E), letter acuities for individual letters decrease with increasing eccentricity. Figure 1.9 shows the classical result presented by Antis (1974) that showed the gradient in letter size (at the resolution threshold) as a function of eccentricity.

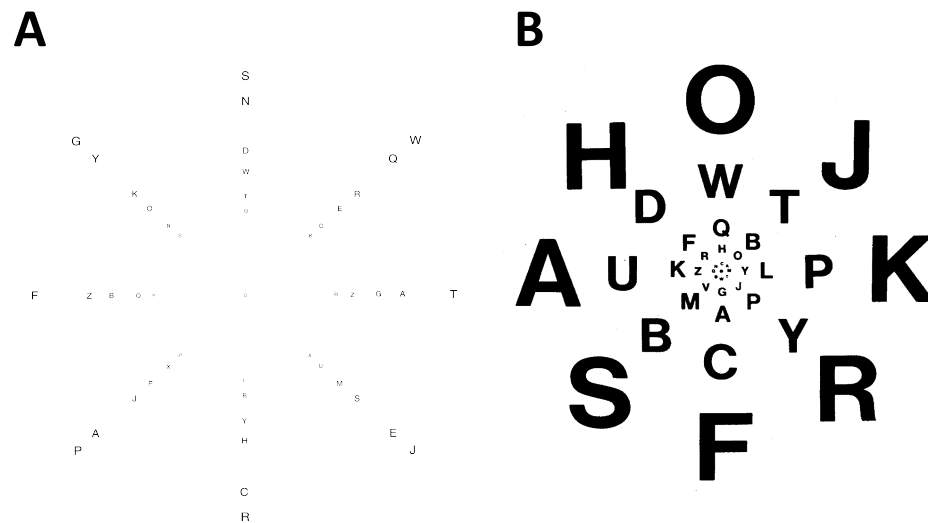


Figure 1.9 The change in letter size as a function of eccentricity.

To obtain the chart in its original size, enlarge it such that the centre of the lower "R" is 66 mm from the fixation point. A) at threshold level and B) at the letter 10 times its threshold size (Antis, 1974; Strasburger et al., 2011).

However, further previous studies showed that different letters showed different legibility at the fovea (Grimm et al., 1994; Alexander et al., 1997; Reich & Bedell, 2000; Shah et al., 2012; Hamm et al., 2018; Ludvigh, 1941; Strasburger et al., 2011; Hairol et al., 2015) and peripheral visual field locations (Ludvigh, 1941; Strasburger et al., 2011; Hairol et al., 2015; Anderson & Thibos, 2004; Shah et al., 2011; Shah et al., 2012). Furthermore, Sloan (1959) found that the percent correct (as a measure of difficulty in recognising the letter, hence legibility) differs among Sloan letters at resolution threshold in central vision (Sloan, 1959; Ferris et al., 1982; Alexander et al., 1997) (Table 1.1). The differences in legibility among letters would cause differences in individual letter acuities. Using Sheridan-Gardiner letters (A, H, O, T, U, V, X), Hairol et al. (2015) found substantial differences in individual letter acuities and these differences increased with the increasing in eccentricity (Figure 1.10 A) (Hairol et al.,

2015). Differences in individual letter acuities were also observed when employing Sloan letters (Figure 1.10 B) (Reich & Bedell, 2000; Alexander et al., 1997; Shah et al., 2012; Barhoom et al., 2021).

Table 1.1 The percent correct of Sloan letters at the resolution threshold (Sloan, 1959).

Sloan letter	Percent correct at resolution threshold
S	70.60
O	71.00
C	71.40
D	79.50
K	82.10
V	84.60
R	86.30
H	89.30
N	91.60
Z	94.00

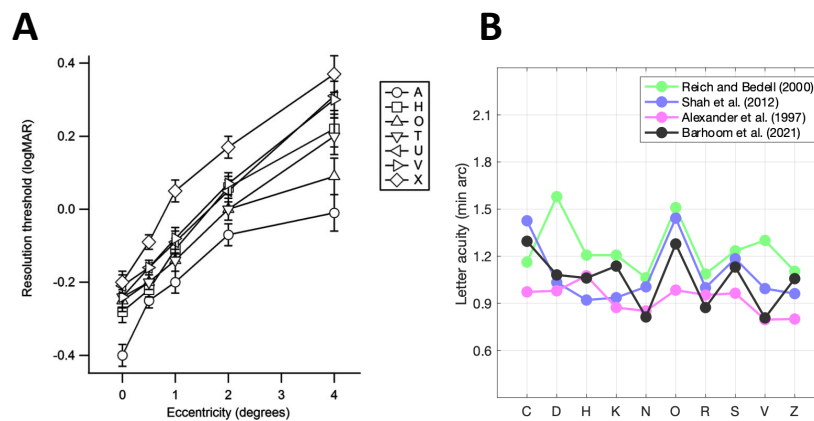


Figure 1.10 Differences in measured letters acuities at fovea and periphery.

A) the letter acuities for individual Sheridan-Gardiner letters as a function of eccentricity (Hairiol et al., 2015). B) Individual Sloan letters acuities at the fovea for several previous studies.

1.8 Source of observed differences in letter acuity

The letter acuity is determined by the letter's identification accuracy compared to the alternative letters. The accuracy of letter's identification in turn is influenced by its perceivability, similarity with other letters, and the bias towards or against it (Mueller & Weidemann, 2012). Perceivability is a measure of how legible the letter is depending solely on the characteristics of the letters, such as the letter size, contrast, or shape. To avoid ambiguity in using the term perceivability, I will adhere to the term *sensitivity* instead, which is more relevant to the Signal Detection Theory (SDT) framework (Green & Swets, 1966) which will be employed in this thesis. Response bias is defined as the tendency to favour one response over the other alternatives (Macmillan & Creelman, 1990), and similarity is defined as the confusion in letter perception which arises among certain letters (e.g., C and O or N and H). In other words, letter recognition, i.e., the letter resolution threshold, could be affected by changing the amount or the type of the sensory input, e.g., size and contrast (sensitivity), the bias towards certain letters in case of uncertainty (response biases), and the confusion between similar letters such for instance C and O (similarity). Note that from these definitions it is well understood that response biases, unlike sensitivity and letter similarities, are independent of the sensory inputs of the stimulus.

1.9 Methods to analyse the source of errors in letter identification tasks

Mueller and Weidemann (2012) reviewed more than 70 studies that investigated the source of error in letter identification accuracy (Mueller & Weidemann, 2012). Many of the reviewed studies attributed the differences in letter identification to the response

bias and similarity (Townsend, 1971 a & b; Gilmore et al., 1979; Mueller & Weidemann, 2012). In the majority of these studies, versions of Luce's choice model (1963) were used to decompose the presented *vs* responded confusion matrix into a bias vector and similarity matrix. The bias vector contains the bias parameter for each letter used as a response alternative in the experiment where all bias parameter values should sum to the guessing rate value of the task. In other words, the response bias of a given letter is the amount of the increase or decrease in the *usage* of that letter from the guessing rate (Luce, 1963; Mueller & Weidemann, 2012). The similarity matrix shows the similarity parameters for each pair assuming that the similarity is symmetrical (i.e., if H is similar to K, then K is similar to H by the same amount). Luce's choice model does not account for the sensitivity of the letters, and it is assumed that the sensitivity of a given letter is the average of Luce's similarity parameters of the letter with the other letters (Mueller & Weidemann, 2012). In our previous work (Georgeson et al., 2023), we investigated the role of response bias and sensitivity in letter identification tasks by devising a model (Noisy Template Model (NTM)) using the framework of SDT. The main finding suggests that both the response bias towards or against letters and sensitivity differences between letters were influential on individual letter identification. Additionally, the response bias was found to be the key factor in determining how often the letters were used in the experiment (either correct or incorrect). However, we did not estimate the effect of similarity between letters on letter identification, and it is assumed that the sensitivity is a measure of a letter's similarity with the other letters. Mueller and Weidemann (2012), on the other hand, introduced a model to estimate the role of similarity in addition to sensitivity (they call it *perceivability*) and response bias simultaneously. They found that the three factors were important in the letter identification task and better than the two-factor models

incorporating only bias & sensitivity or bias & similarity (Mueller & Weidemann, 2012).

1.9.1 The effect of sensitivity, response bias and similarity on the estimated letter acuity

Most of the previous studies investigated the sensitivity, response bias, and similarity in letter identification tasks to establish a better understanding of how we perceive letters or to introduce new models that explain the letter identification task (Townsend, 1971 a & b; Gilmore et al., 1979; Mueller & Weidemann, 2012). However, these studies did not investigate the role of these factors on the estimated letter acuities. Further, they did not interpret or investigate the role of bias in letter acuity measurements and often assumed that bias had little effect on the measured letter acuity (Candy et al., 2011; Coates, 2015; Hamm et al., 2018; Barhoom et al., 2021; Georgeson et al., 2023). Moreover, unlike bias, it has been shown that similarity was a major cause of error in letter acuity estimations (Candy et al., 2011; Coates, 2015; Hamm et al., 2018; Barhoom et al., 2021; Barhoom et al., 2022) and that visual acuity estimation using letters can be improved by weighting the responses to the letters according to their similarity (Grimm et al., 1994; McMonnies & Ho, 1996; Fülep et al., 2017; McMonnies & Ho, 2000). Similar to these previous studies, we found that the observed biases had negligible effect on letter acuity estimated from the data pooled across letters (Georgeson et al., 2023). These results can be understood from the fact that bias towards some letters necessarily entails bias against other letters. Hence, pooling responses across letters cancels the biases of individual letters and leaves little or no effect on letter acuity estimated from the pooled data (Georgeson et

al., 2023). On the other hand, we expect that the effect of similarity in a letter acuity estimation would be different from that of the bias because the similarity between two given letters will result in both letters having fewer correct responses compared to the case of dissimilar letters. In other words, the similarity between letters has only the effect of decreasing correct responses for similar letters. Since similarity between letters causes fewer correct responses. Barhoom et al. (2022) showed that the spread of the underlying PMF will become wider when data are pooled across similar letters compared with data pooled across dissimilar letters. Hence the test-retest variability of letter acuity estimation using an adaptive method (the method frequently adopted in clinical tests, e.g., staircase) will be higher (less precise) in the presence of similar letters. Many previous studies showed that compensating for the effect of similarity between letters reduces the test-retest variability significantly and enhances the precision of the estimated letter acuity (Grimm et al., 1994; McMonnies & Ho, 1996; Fülep et al., 2017; McMonnies & Ho, 2000). This is very important when letter acuity tests are used to detect clinically significant deteriorations while monitoring the progression of certain eye conditions over time. By taking into consideration the aforementioned factors, letter acuity can be efficiently used to estimate the sensitivity of the paracentral visual field in healthy and clinical populations compared to the methods used in conventional visual fields (i.e., light/contrast sensitivity).

1.10 Letter acuity measurements in the paracentral visual field

For the remainder of this chapter, I will outline the rationale and advantages of using letter acuity measurements in the mapping of the paracentral visual field (i.e., macula) in a clinical population, namely Glaucoma.

1.10.1 Glaucoma definition

Glaucoma is an optic neuropathy associated with characteristic structural damage to the optic nerve, and visual dysfunction (Foster et al., 2002). There are eight subtypes of Glaucoma, the most common is the primary open-angle Glaucoma (Casson et al., 2012). All subtypes of Glaucoma have in common a potentially progressive and characteristic optic neuropathy which is associated with visual field loss as damage progresses, and in which the intraocular pressure is usually a key modifying factor (Kanski & Bowling, 2011). In 2020, Glaucoma was classified as the most common cause of irreversible blindness worldwide and one of the major causes of moderate to severe visual impairment, especially in adults aged 50 years and older (Quigley & Broman, 2006; Allison et al., 2020).

1.10.2 Pathophysiology and macular involvement in Glaucoma

1.10.2.1 Retinal ganglion cells death in Glaucoma

The neurodegenerative process of Glaucoma affects all parts of retinal ganglion cells. In vivo studies showed that the degenerative process is typically a progressive dendritic shrinkage followed by the loss of the axons and the cell bodies of retinal ganglion cells (Leung et al., 2011; Almasieh et al., 2012; Kalesnykas et al., 2012; Chong & Martin, 2014).

1.10.2.2 Macular damage in Glaucoma

The diameter of the macula is approximately 6 mm (central 20 deg). 50% of retinal ganglion cells in the retina are located within 4.5 mm (central 16 deg) of the macula. This area (i.e., central 16 deg) represents less than 5% of the total area of the retina (Curcio & Allen, 1990; Provis et al., 2005). Despite its small area, the macula has a vital role in central vision. Central vision is essential for most daily life activities. Consequently, macular damage significantly deteriorates the vision-related quality of life (Garg et al., 2018). Many studies showed that the macula is involved in the early stages of the retinal damage caused by Glaucoma (Zhang et al., 2014). Hou et al. (2013) found that the retinal ganglion cell complex thickness and macular vessel density in pre-perimetric and primary open-angle Glaucoma eyes are significantly lower compared to healthy eyes (Hou et al., 2019). Moreover, Hood et al. (2013) investigated macular damage in Glaucoma and described its characteristic pattern. They found that macular damage is typically arcuate and often associated with local retinal nerve fibre layer thinning in a narrow region of the disc (Hood et al., 2013).

1.10.3 Macular damage assessment in Glaucoma

The macular damage in Glaucoma can be assessed structurally and functionally. Structurally, macular damage can be quantified using advanced imaging techniques such as ocular coherence tomography. Ocular coherence tomography is a high-resolution imaging device that applies the principle of interferometry to interpret reflectance data from a series of multiple side-by-side A-scans (i.e., amplitude scans using ultrasound waves) combined to form a measurable cross-sectional image. It has

been found that macular damage (quantified as the reduction in thickness) was evident in early Glaucoma stages (Sung et al., 2012; Michelessi et al., 2015; Hood et al., 2019; Tsamis et al., 2022).

The early macular damage in Glaucoma was noticed functionally in earlier studies and was described as a visual field defect in the upper visual field within 10 deg in the macular region (Aulhorn & Harms, 1967; Anctil & Anderson, 1984). Furthermore, it has been found that deep arcuate defects occur in and near the foveal region of the visual field (Hood et al., 2011). But clinically, this damage is usually missed when the common 24-2 visual field test is used in Glaucoma management mainly because of the poor sampling of the central visual field (Hangai et al., 2014; Grillo et al., 2016; Alluwimi et al., 2018). Even though macular damage in Glaucoma is evident, unfortunately, the current classification of Glaucoma severity is based on 24-2 or 30-2 visual field tests which underestimate the disease severity by missing the macular damage (DeMoraes et al., 2019). The current classification might be one of the major causes why Glaucoma specialists are reluctant to employ macular visual field tests widely in Glaucoma management. For instance, a study found that most Glaucoma specialists in the United Kingdom use standard automated perimetry for Glaucoma management, and only a few use more sensitive tests to damage, such as frequency doubling technology and short wavelength automated perimetry (Gordon-Bennett et al., 2008). Hence, I speculate that only very few will consider macular visual field tests in Glaucoma management to either detect or monitor the Glaucoma damage at the macula.

1.10.3.1 Clinical macular visual field tests

The macular damage can be assessed functionally using many available clinical visual field tests. Standard automated perimetry 10-2 and M-pattern strategies are designed to test the central 20 deg retina. 10-2 test contains 76 test locations that are equidistant with 2 deg separations. M-pattern design is less common and has 81 test locations with higher resolution test points (i.e., more test points that are closer to each other) at the fovea (Racette et al., 2016). Both tests use light/contrast sensitivity detection tasks. The 10-2 strategy is commonly used in the assessment of macular damage. It showed a superior ability to detect macular damage compared to the 24-2 strategy in all stages of Glaucoma. A 10-2 test can detect abnormal central visual field that is classified as normal in a 24-2 test in early Glaucoma, Glaucoma suspects, and ocular hypertension (Grillo et al., 2016; DeMoraes et al., 2017; Tomairek et al., 2020).

The frequency doubling technique 10-2 and macular tests employ a special target design to preferentially stimulate magnocellular ganglion cells which are believed to show early damage in the early stages of Glaucoma (Khanna et al., 2022). The target size is 2×2 deg and it is designed so that it consists of 0.5 cycle/deg spatial frequency with a temporal counter phase flickering rate of 12 Hz. When presented, the spatial frequency will double as an effect of the flickering. The task is to detect the presence of the target. Frequency doubling technique 10-2 has 44 test locations to cover the central 20 deg whereas frequency doubling technique macular has 16 locations to cover the central 8 deg. It has been found that the detection ability of early macular damage in Glaucoma is higher when using the frequency doubling technique compared to standard automated perimetry (Jung et al., 2017; Park et al., 2018).

Short wavelength automated 10-2 utilises a blue stimulus on a yellow background to

preferentially stimulate the blue cones whilst adapting the green and red cones and inhibiting the activity of the rods simultaneously (Wild, 2001; Khanna et al., 2022). Like standard automated perimetry 10-2 and frequency doubling technique 10-2, short wavelength automated 10-2 test can detect the macular functional loss in pre-perimetric Glaucoma patients which is usually missed in 24-2 standard automated perimetry (Leeprechanon et al., 2007; Jung et al., 2015). However, it has been found that short-wavelength automated 10-2 test is more influenced by media opacities and shows higher test-retest variability compared to standard automated perimetry (Sharma et al., 2008; Khanna et al., 2022).

Hood et al. (2014) examined adding two additional central points to the 24-2 strategy. They found that adding two test points in a particular location in the macular area increased the efficacy of this test to detect macular damage (Hood et al., 2014; Chen et al., 2015). The test has been recently introduced for clinical use.

Microperimetry is a technology that allows the study of retinal sensitivity at different foveal and macular areas as well as eye fixation (Molina et al., 2018). One of the main uses of microperimetry is to monitor and quantify Glaucoma residual visual functions and functional vision in the macula. However, this application in Glaucoma practice is rare today (Markowitz & Reyes, 2013). Matsuura et al. (2018) evaluated the test-retest reproducibility and structure-function relationship of the MP-3 (one of the advanced microperimetry types) in Glaucoma and compared it with the standard automated perimetry. They found that the MP-3 has similar test-retest reproducibility to the standard automated perimetry, but a better structure-function relationship (Matsuura et al., 2018). There are many advantages of microperimetry over standard automated perimetry such as the precise fundus tracking in case of poor fixation, the co-registration of the results to the fundus image (allows for structural to functional

comparisons), and the possibility to change the stimuli type (Markowitz & Reyes, 2013; Acton & Greenstein, 2013).

1.10.3.2 Monitoring of macular damage progression in Glaucoma

The most challenging factor that affects the detection of visual field damage progression is the test-retest variability. The test-retest variability is very important to differentiate between a true progression and random variability in visual field test interpretation. For this reason, two methods have been employed to help in the assessment and detection of the presence and degree of a true progression, namely event-based and trend-based methods. In the event-based method, progression is detected once a predetermined degree of deterioration has occurred (e.g., the guided progression analysis software). On the other hand, the trend-based method detects progression once there is a statistically significant downward trend in some visual field indices such as the mean deviation (O’Leary et al., 2012; Rao et al., 2013; Tanna & Desai, 2014).

1.10.3.3 Methods used to monitor the damage progression at the macula

Macular damage monitoring in Glaucoma has been shown to be efficient in monitoring Glaucoma progression at all stages. Moreover, in the advanced stages of Glaucoma, progression assessment by macular monitoring becomes essential (Hood et al., 2013; DeMoraes et al., 2018). Unfortunately, commercially available perimetry platforms do not include progression analyses to aid in the detection of progression of the macular damage using macular visual field tests (Tanna & Desai, 2014).

However, in a study that applied the same indices of the 24-2 test to monitor Glaucoma progression using the 10-2 test, it has been found that this test has an accurate representation of the extent and rate of visual field progression in the macula than any other currently used parameters of visual field progression (DeMoraes et al., 2014). Furthermore, Park et al. (2013) found that the progression detected by 10-2 is higher when compared to 24-2 tests for patients with Glaucoma and initial parafoveal scotoma (Park et al., 2013).

1.10.4 Factors that determine an efficient test for macular damage monitoring

The loss of retinal ganglion cells in Glaucoma is irreversible (Evangelho et al., 2019). Therefore, Glaucoma monitoring for further damage to the retina is the most important part of Glaucoma management. By timely detection of retinal damage, clinicians can amend the course of treatment to slow down and prevent further visual function loss, and as a consequence maintain good vision-related quality of life for the maximum possible period for Glaucoma patients. Ideally, the criteria of the test that should be used to monitor Glaucoma are (i) sensitivity to the progression of the damage over a short follow-up period (could be weeks to several months), (ii) having low test-retest variability (hence, needing few sessions to detect true progression), (iii) being an easy to do task for the patients (i.e. intuitive) and (iv) being an easy to access test (home-based as opposed to clinic-based test). In other words, the test should be able to determine a significant amount of retinal damage progression in a few sessions over a short period with an easy-to-access and easy-to-perform test for patients.

In the following, I will discuss the structure-function relationship of the visual task of a given visual field test as an indicator of the degree of test-retest variability and as an

indicator of the ability of the test to detect the presence and progression rate of macular damage. I will subsequently discuss how to create a test with a given visual task easily accessible and efficiently performed by Glaucoma patients for macular damage monitoring.

1.10.4.1 The visual task with the strongest structure-function relationship

In any visual task, the stronger the structure-function relationship is, the more sensitive it is to the change in the number of retinal ganglion cells. Although many macular tests that employ light sensitivity tasks have been found to produce good structure-function relationships (Hirooka et al., 2016; Jung et al., 2017; Hood et al., 2019), resolution perimetry has the strongest and the most direct correlation to retinal ganglion cells density because it is sampling-limited (Phelps, 1984; Rankin et al., 1999; Anderson, 2006; Wen et al., 2021). Nevertheless, Anderson (2006) pointed out that resolution tasks (e.g., letter recognition) could be demanding because the observer needs to identify or recognise the stimuli (e.g., identifying the direction of the gap in Landolt C or naming the letter) rather than just merely detect its presence. Vanishing optotypes employed in high pass resolution perimetry convert the visual task from a recognition task into an easier detection task whilst keeping the task sampling-limited (Frisen, 1986; Frisen, 1987; Frisen, 1993; Numata et al., 2016; Sabeti et al., 2021). However, it has been shown that this is not always the case, and a difference between the detection threshold and resolution threshold was found when employing the vanishing optotypes detection task at the periphery and it would be no longer a sampling-limited task (Thibos et al., 1987; Anderson & Ennis, 1999; Ennis et al., 2002). Thus, employing high-contrast optotypes in a recognition task (e.g., letter recognition)

maintains the advantage of being sampling-limited at the periphery. Consequently, the task performance closely represents the retinal ganglion cell loss associated with Glaucoma damage (Anderson, 2006).

1.10.4.2 Easy to access test for patients (home-based vs clinic-based)

Guidelines proposed by the European Glaucoma Society recommend that the frequency of visual field tests should be increased for newly diagnosed patients to better determine the rate of visual field progression (Glen et al., 2014). However, this will increase the number of follow-up visits to the clinics, adding pressure to the health sector and posing challenges to the delivery of Glaucoma services. (Vesti et al., 2003; Hamzah et al., 2020). Furthermore, the increase in the demand for Glaucoma services besides other chronic eye conditions is predicted to rise by 25% over the next 10 years because of many factors including the reduction in clinical services during the COVID-19 pandemic (Lakhani et al., 2021). This increase will add additional pressure on the clinics. The increased number of clinic visits also adds a burden on patients because frequent follow-up visits mean frequent appointments. In a clinic-based follow-up study, Glen et al. (2014) showed that the effectiveness of follow-up care can be affected by several factors such as distraction due to testing environments (patients felt they performed better in the morning when they were more alert), excessive waiting times (usually unpredictable but estimated to be two hours) and appointment booking. Moreover, traveling to the clinic has its own challenges such as the long distances patients need to travel, the rush hours they need to avoid, the travel costs, and concerns about traveling alone (Glen et al., 2014).

Home-based monitoring can solve many problems that arise from clinic-based follow-

up (Rodriguez & Azuara, 2018; Hamzah et al., 2020). For instance, Anderson et al. (2017) investigated the efficacy of home-based monitoring in the early detection of rapid visual field progression (by simulation). They found that weekly home-based monitoring is more sensitive to the progression compared to six months follow up in the clinic even with more variable test results and low patient compliance (which is expected in any home-based monitoring strategy due to, for example, IT logistics or lack of motivation) (Anderson et al., 2017; Prea et al., 2021). Consequently, self-assessment of the macular visual field by a home-based monitoring device would be of great value in closely monitoring macular damage progression.

1.10.4.3 Available home-based visual field tests

The efficient home-based test should employ a visual task that is easy to perform without supervision. It should also be a visual task with minimum requirements for ambient or display illumination calibration (Liu et al., 2014; Rodriguez & Azuara, 2018). This could be in the form of a simple hand-held chart or a smart device. A classic example of a hand-held chart is the Amsler grid test. This test is a very simple and cost-effective tool. Clinicians usually give an Amsler chart to patients with a risk of central vision damage and ask them to assess their central visual field daily or weekly at home. Nassar et al. (2015) found that the Amsler grid is an accurate test in the detection of macular disease with a sensitivity of 88%. A digital version of the Amsler grid allows the creation of a 3D map of the central visual field which shows the location, length, slope, depth, and shape of existing defects in Glaucoma (Fink & Sadun, 2004; Nguyen et al., 2009; Kovalevskaya et al., 2016). However, these results were observed under strict test conditions (Nguyen et al., 2009; Nassar et al., 2015).

Consequently, they had control over many factors that might make the Amsler grid less sensitive to macular damage. When used frequently at home, the Amsler test was found to have low sensitivity (less than 50%) to macular damage, and the outcome was found to be highly subjective. Apart from the lack of compliance of patients to the test procedure, the reason for this drop in sensitivity with frequent testing of the macular damage is the "filling in" or the "perceptual completion" phenomenon (Crossland & Rubin, 2007).

Using smart devices for home-based monitoring of macular damage is preferable because it reduces the subjective judgment of progression found with simple charts and it quantifies the visual loss. Many tests were reported in the literature to produce good results in home-based monitoring of macular damage progression. Most of the home-based tests utilise the light sensitivity task to detect macular damage using mainly tablet-based or laptop software (e.g., Eyecatcher, Melbourne Rapid Fields, Visual Fields Easy, and Viewi) (Rodriguez & Azuara, 2018; Hamzah et al., 2020; Jones et al., 2021).

Additionally, it has been found that Vernier acuity, and similar hyperacuity tasks that assess spatial sampling, may be also useful in the detection of early glaucomatous damage (Mckendrick et al., 2002). For example, the ForeseeHome AMD Monitor is a portable version of the preferential hyperacuity perimetry. It has been developed for self-assessment of maculopathy at home, but an important disadvantage of this home preferential hyperacuity perimetry device is the cost barrier (the patient needs to acquire the device and it is not a software that can be installed) (Trevino, 2008). The Shape Discrimination Test is an application that could be installed on smartphones to be used as a self-assessment tool that also utilises hyperacuity tasks (Chhetri et al., 2010). This application is intuitive to use, readily accessible, and sensitive to the

severity of maculopathy (Wang et al., 2013). However, hyperacuity tasks are independent of lesion size and location (Trevino, 2008), which is a limitation in monitoring the characteristic Glaucoma macular visual field damage.

1.10.4.4 Home-based tests that employ resolution task

Nevertheless, none of the aforementioned home-based visual field tests employs a recognition task (i.e., resolution task) using either Landolt C, Tumbling E, alphabets, or numerical stimuli. The reasons are the additional demanding nature of the recognition task compared to the detection task. Additionally, the recognition task cannot be performed independently (i.e., supervision is needed to type in the responses using a keyboard for instance). Although not designed for home-based monitoring, the Macular Mapping Test is a software that uses recognition tasks to screen for macular damage (MacKeben, 2008). The optotypes can be uppercase letters, Landolt C, or Tumbling E. The 32 test locations are positioned around foveal and parafoveal regions within 8 deg eccentricity. The task is to identify the presented target verbally. The examiner types in the response. The scores are recorded as two points for seen and recognised correctly, one point for seen and not recognised, and zero points for not seen. The basic test of 36 trials takes only three minutes to complete. The disadvantages include the high test-retest variability, no resolution estimate at each test location (lack of pointwise tracking of progression), and the need for a person to type in the responses. (Bartlett et al., 2005; MacKeben, 2008; Hahn et al., 2009).

1.10.5 Suggestions for a new approach of home-based test for macular damage monitoring

1.10.5.1 Visual task to be employed

As mentioned previously, the resolution task is a sampling-limited task that has a stronger structural-functional relationship than the light/contrast sensitivity detection task employed in most home-based tests. Furthermore, it has been found that resolution acuity at the periphery is tolerant of blur up to 1 - 4 dioptres before performance declines (Anderson, 1996b; Wang et al., 1997). Generally, employing light/contrast sensitivity detection tasks in home-based tests may be a task that is difficult to control because it is demanding in terms of the calibration of parameters such as background illumination and stimulus intensity. Hence, resolution tasks, which employ black-on-white stimuli will make it easier to calibrate the test. Even a loss of up to 90% in contrast will not affect the resolution performance at the periphery (Thibos et al., 1996). Therefore, a letter recognition task that is easy to perform (i.e., intuitive), sampling-limited, tolerant to defocus, and robust to the effects of contrast loss, whether through lens opacity or oblique axis astigmatism, allows more detection of neural loss in diseases like Glaucoma (Millodot et al., 1975; Thibos et al., 1996; Anderson, 1996b; Wang et al., 1997; Anderson, 2006; Rosén et al., 2011). Furthermore, using many letters (Sloan letters for instance) will reduce the guessing rate associated with forced choice tests (Pelli & Robson 1991), hence, will reduce the test-retest variability. Additionally, adopting high contrast letter recognition task will ensure that the task remains sampling-limited (Thibos et al., 1987; Anderson & Ennis, 1999; Ennis & Johnson, 2002; Anderson, 2006).

1.10.5.2 Response method to be employed

Adopting many alternatives in a vision test (e.g., 10 Sloan letters) will make the visual task demanding in terms of entering the responses if performed independently. Usually, an experimenter will be needed to help with typing in the responses using a keyboard if the responses are to be given verbally. However, by employing speech recognition technology, the task can be performed without supervision. Many studies showed that using speech recognition technology is efficient in entering verbal responses in vision tests (Khan & Ullah, 2017; Nisar et al., 2022).

In Chapter 4, I employed a simple subject-dependent algorithm of speech recognition (Selvaraju, 2022) to enter responses in a visual field experiment using letters as stimuli in healthy and Glaucoma observers. Due to the cost barrier, it was impossible to employ a more advanced speech recognition method in the current project.

1.11 Problem statement and objectives

1.11.1 Problem statement

A resolution task using letters as optotypes, combined with speech recognition technology as an input method, can be a promising and efficient approach for developing a home-based flagging tool to monitor macular damage. To the best of my knowledge, there is currently no visual field test that takes the aforementioned factors into consideration.

1.11.2 Objectives

The general aim of this thesis was to explore the possibility of mapping macular damage in glaucomatous eyes with a novel test that employs letters as stimuli and speech recognition technology. To achieve this, I conducted three studies. In the first study (Chapter 2), I investigated the effect of response bias and similarity calculated by the Luce's choice model (1963) on the estimated letter acuity of Sloan letters in the letter identification task. In the second study (Chapter 3), I introduced a novel and significant extension to our original model (Georgeson et al., 2023) by incorporating the similarity factor into the NTM. The new model allows us to estimate the joint effect of bias, sensitivity differences, and similarity between letters simultaneously, instead of bias & sensitivity or bias & similarity only. Furthermore, through simulations using the NTM, I examined the effect of bias, sensitivity differences, and similarity between letters on the systematic error and the test-retest variability of letter acuity estimated by an adaptive method. These two studies aim to establish a better understanding of the key factors responsible for the variability in letter acuity measurements. In Chapter 4, I explored the feasibility of acuity perimetry (i.e., visual field test with letters as stimuli and speech recognition technology as an input method) in healthy and Glaucoma observers.

Chapter 2 . The Effect of Response Bias and Similarity on Letter Acuity Measurements of Sloan Letters in Central and Paracentral Vision

2.1 Motivation

In this chapter, I will investigate a Sloan letter identification task. This experiment aims to assess the role of response bias and similarity between letters (calculated by Luce's choice model) on the letter acuity measurements of individual Sloan letters in healthy observers.

2.2 Methods

2.2.1 Observers

Data were collected from 10 naïve observers (seven females, mean age 23.8 ± 4.4 (SD), age range: 19 - 32 years) with healthy eyes with no history of a systematic disorder or learning disability. The mean best-corrected visual acuity and the mean refractive error (spherical equivalent) were 0.05 ± 0.06 logMAR and -2.5 ± 2.3 DS, respectively. Written informed consent was obtained from all observers. The study was approved by the University of Plymouth Ethics Committee. The study was conducted in accordance with the Declaration of Helsinki.

2.2.2 Stimuli

High-contrast letters were used in the study. They were black letters (luminance = 2.2 cd/m²) on a white background (luminance = 215 cd/m²), resulting in 99% Weber contrast according to the following equation:

$$\text{Weber contrast} = \frac{L_{\max} - L_{\min}}{L_{\text{background}}} \quad (\text{Eq. 2.1})$$

where L_{\max} , L_{\min} , and $L_{\text{background}}$ are luminance maximum, minimum, and background, respectively.

A total of 10 standard Sloan letters (C, D, H, K, N, O, R, S, V, Z) were employed. The letters were presented centrally and at paracentral locations at vertical eccentricities of ± 3 deg from fixation (Figure 2.1 A). The structure of the letters follows the Sloan letter design so that their height is equal to their width and five times the stroke width (Figure 2.1 B).

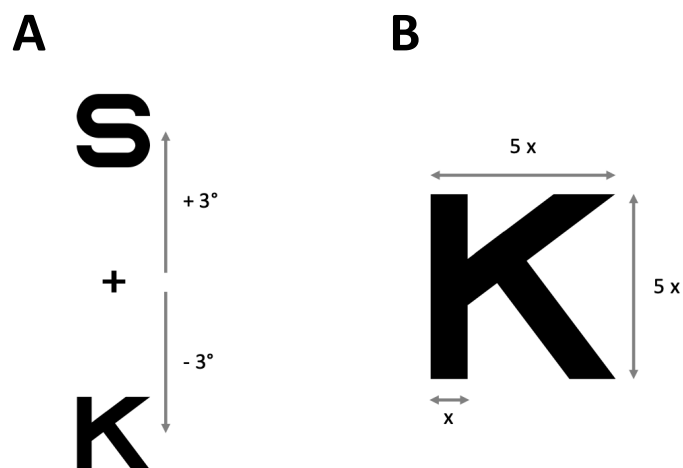


Figure 2.1 The test locations and the stimulus design.

Figure shows A) the central and paracentral test locations. B) the dimensions of the Sloan letters, exemplified by the letter K. The stroke width $\times = 1/5$ of the letter's height. The height and the width of the letter are equal.

2.2.3 Apparatus

The stimuli were presented on a *Dell* P2317H LCD monitor (1920 × 1080 pixels) with a frame rate of 60 Hz. Observers viewed the targets at a viewing distance of 350 cm while sitting on a chair without using a chin or forehead rest. At this distance, one pixel subtended 0.258 minutes of arc ('). The examiner monitored a constant viewing distance through regular checks. The experiment was carried out under a room illumination of 160 lux. The observer responded by calling out the presented letters. The responses were entered by the experimenter via a standard computer keyboard to minimise errors caused by mistyping and to improve fixation compliance.

2.2.4 Software

Stimuli were generated using *Matlab* 2016b (MathWorks, Natick, Massachusetts, USA). Routines from the Psychtoolbox-3 were used to present the stimuli (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). All model fitting routines and statistical analyses were performed using *Matlab* 2019a.

2.2.5 Procedure

2.2.5.1 Data collection

The experiment was conducted monocularly; either the left or right eye was chosen at random, and the fellow eye was occluded using an opaque eye patch. For this experiment, the method of constant stimuli was used to collect the data. The responses were collected to briefly presented (250 milliseconds and accompanied by an auditory

signal) individual Sloan letters. Observers were asked to recognise the presented letters in a single-interval 10AFC task and to report it verbally (Figure 2.2). The experimenter entered the responses via the computer keyboard. The Sloan letters were presented randomly in an interleaved fashion across the three locations. During the experiment, the observers were asked to fixate on a fixation cross (dimensions: length and width = 1.55', stroke width = 0.516') presented at the centre of the screen. The fixation cross disappeared for the duration of the central presentations and reappeared for the paracentral presentations. At each location, six letter sizes were presented. Multiple pilot experiments were conducted to establish appropriate letter sizes to cover the whole range of responses (ranging from guessing (10%) to certain decisions (100%)). Six different letter sizes (always defined by their stroke width in minutes of arc, and spaced logarithmically) were tested; 0.3', 0.44', 0.64', 0.94', 1.37', and 2' for central presentations and 0.5', 0.79', 1.26', 1.99', 3.15' and 5' for peripheral presentations. At each size, each letter was presented 10 times. Each observer completed 1800 trials (10 Sloan letters \times six letter sizes \times 10 presentations per letter \times three test locations). The data were collected in the form of the frequency of responses of each letter to the presented letters (either correct or incorrect). Consequently, for each observer and for each location, a confusion matrix of the presented *vs* responded letters was obtained. Only choices of the 10 Sloan letters were accepted. To familiarise the observers with the Sloan letters, the experimenter demonstrated the Sloan letters at the beginning of the session. In the rare cases where observers responded with other letters, the experimenter prompted a second response. If the observer failed the second attempt, a reminder of the Sloan letter set was provided (this occurred very rarely, on average not more than once per observer). Observers showed excellent compliance in responding from the Sloan letter set (< 30 errors per observer in 1800 trials).

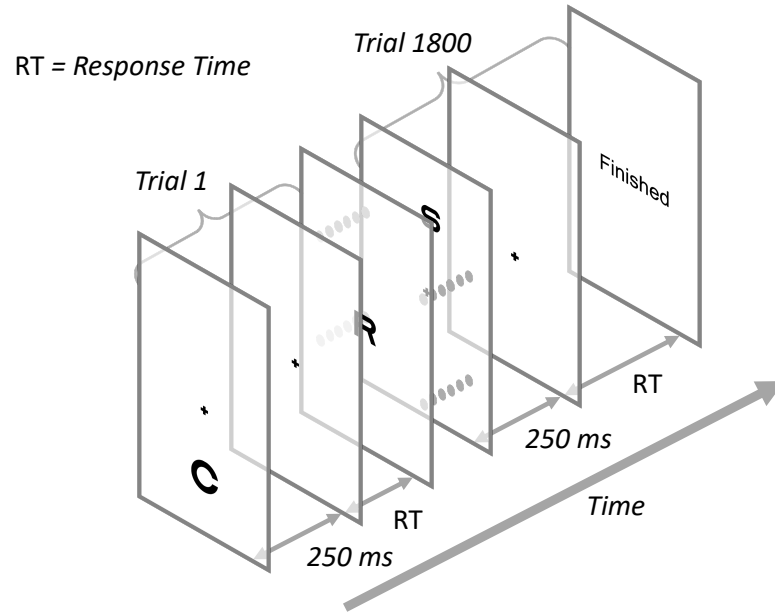


Figure 2.2 The sequence of the events in the trial in the experiment.

2.2.5.2 Letter acuity measurement

To estimate the letter acuity for a given letter, a psychometric function (PMF) (Log-Quick function was chosen) (Eq. 2.2) was fit to the experimental data (i.e., proportion corrects for six letter sizes of that letter identity) separately for each observer and each tested eccentricity according to the equation:

$$P_{\text{correct}} = (\gamma + (1 - \gamma - \lambda)) \times \left(1 - 2^{-10^{\beta(x-\alpha)}}\right) \quad (\text{Eq. 2.2})$$

where P_{correct} is the percent correct of the Log-Quick function when $\gamma = 0.10$ (the guessing rate to respond to the letter by chance), $\lambda = 0.02$ (the lapse rate for naïve observers), x is the letter size, α is the threshold (i.e., letter acuity) and β is the slope of the function. The letter acuity α was defined as x yielding 54% percent correct, according to the following equation.

$$P_{\text{correct}} = (0.1 + (1 - 0.1 - 0.02)) \times \left(1 - 2^{\left(-10^{\beta(\alpha-\alpha)}\right)}\right) = 0.54 \quad (\text{Eq. 2.3})$$

The model was fitted to data using maximum likelihood - adjusting parameter values of letter acuity (α) and slope (β) to maximise the log-likelihood (LL) of the parameters given the data according to the equation.

$$\text{LL} = \sum \{n \times \log(P_{\text{correct}}) + (t - n) \times \log(1 - P_{\text{correct}})\} \quad (\text{Eq. 2.4})$$

where for each condition n is the observed number of correct responses in t trials, P_{correct} is the Log-Quick function probability of being correct, the variables n , t , and P_{correct} range over the six letter sizes, and the summation takes place over those six conditions. A finely sampled grid search was used to find the best values of the letter acuity (α) and slope (β) for each observer and test location that maximized LL.

2.2.5.3 Estimation of letters response bias and similarity between letters

Luce's choice model (1963) was employed to compute the response biases and similarity between letters. The model was fitted to the obtained presented vs responded confusion matrix for each observer and for each location. Luce's choice model attempts to disentangle the response factor that is sensory-independent (i.e., response biases towards or against some letters) from the sensory-dependent response factor (i.e., similarities between letters). The fitting algorithm was originally presented by Smith (1982) and consists of two steps (Smith, 1982; Coates, 2015). The first step is to find the maximum likelihood estimate of the model. The iterative proportional

fitting is used to converge the raw data (i.e., confusion matrix) to the maximum likelihood estimate of the model. The starting matrix values are all ones. To perform the first iteration, adjustments for rows, columns, and similarities are carried out successively. The adjustment for rows is performed by dividing the value of each cell by the sum of the corresponding row values of the starting matrix (ones for the first iteration) then multiplied by the marginal sum of the corresponding row values of the raw (confusion) matrix, followed by the adjustment for columns then for similarities. The resulting matrix will be the starting matrix for the second iteration. These iterations are repeated until there is no significant change in the estimated values. The resulting matrix is the maximum likelihood estimate of the model. The second step is to compute the parameters of the response bias vector and the similarity matrix from the maximum likelihood estimate of the model, according to Eq. 2.5 and 2.6 respectively.

$$\beta_j = \frac{1}{\sum_{k=1}^N \sqrt{\frac{\hat{P}_{jk} \times \hat{P}_{kk}}{\hat{P}_{kj} \times \hat{P}_{jj}}}} \quad (\text{Eq. 2.5})$$

and

$$\eta_{ij} = \sqrt{\frac{\hat{P}_{ij} \times \hat{P}_{ji}}{\hat{P}_{ii} \times \hat{P}_{jj}}} \quad (\text{Eq. 2.6})$$

where the variable β_j in Eq. 2.5 denotes the response bias parameter for the letter j . N is the number of letters (10 letters). η_{ij} is the similarity parameter of each cell between the letter i and the letter j . \hat{P} is the expected value in each cell obtained from the maximum likelihood estimates matrix.

A *Matlab* function was developed by the author to compute the response biases and similarities parameters using Luce's choice model which can be downloaded from the

link: <https://github.com/HBarhoom/Codes->.

The model parameters are presented as a response bias value for each letter and a similarity matrix capturing the similarity between each pair of letters.

2.3 Results

Figure 2.3 shows the average proportion correct and PMF across observers (see Appendix A for the proportion correct and PMF of individual observers). The average letter acuity (\pm SE) across observers was $1.044' \pm 0.06$, $2.60' \pm 0.29$, and $2.59' \pm 0.24$ at the central, upper (+3 deg), and lower (-3 deg) visual field locations respectively. Paired sample t-test revealed no statistically significant difference between the average letter acuity at upper and lower visual field locations ($t(9) = -0.05$, $p = .70$). However, the average letter acuity at the upper visual field ($t(9) = 5.40$, $p < .001$) and lower visual field ($t(9) = 6.70$, $p < .001$) were significantly poorer (i.e., higher when expressed in (')) compared to the central location. The average proportion correct and PMF across observers for individual letters are also shown in the small panels in Figure 2.3.

The average (\pm SE) of the individual letter acuities across observers at the three locations are shown in Table 2.2. Since some of the data showed significantly skewed distributions at $p = .05$ (see Table 2.1 for the significant p-values (i.e., not normally distributed data) highlighted in bold font), the median, inter-quartile range (IQR), 25th quartile (25th Q) and 75th quartile (75th Q) values for each letter at the three locations were also included (Table 2.2)

Table 2.1 Normality test results for individual letter data.

Table shows the results of Shapiro-Wilk normality test for the distribution of observers' data for individual letter.

	The test statistic (W)			p-value		
	ECC = -3 deg	ECC = 0 deg	ECC = +3 deg	ECC = -3 deg	ECC = 0 deg	ECC = +3 deg
C	0.82	0.84	0.79	0.03	0.04	0.01
D	0.96	0.93	0.90	0.75	0.43	0.25
H	0.87	0.88	0.90	0.10	0.12	0.20
K	0.97	0.92	0.90	0.90	0.36	0.17
N	0.92	0.94	0.92	0.40	0.55	0.40
O	0.90	0.90	0.84	0.24	0.17	0.04
R	0.94	0.93	0.91	0.56	0.49	0.30
S	0.77	0.88	0.85	0.01	0.12	0.06
V	0.93	0.87	0.78	0.44	0.11	0.01
Z	0.95	0.76	0.87	0.65	0.01	0.10

Friedman tests showed statistically significant differences between the medians (across observers) of individual letter acuities at the central location ($\chi^2(9) = 42.72$, $p < .001$) (with best and worst acuities for V = 0.74' (0.26) and O = 1.28' (0.34) respectively), at the upper visual field ($\chi^2(9) = 39.79$, $p < .001$) (with best and worst acuities for H = 1.81' (1.07) and O = 3.10' (2.54) respectively) and at the lower visual field ($\chi^2(9) = 32.17$, $p < .001$) (with best and worst acuities for Z = 1.82' (1.14) and O = 3.70' (2.87) respectively).

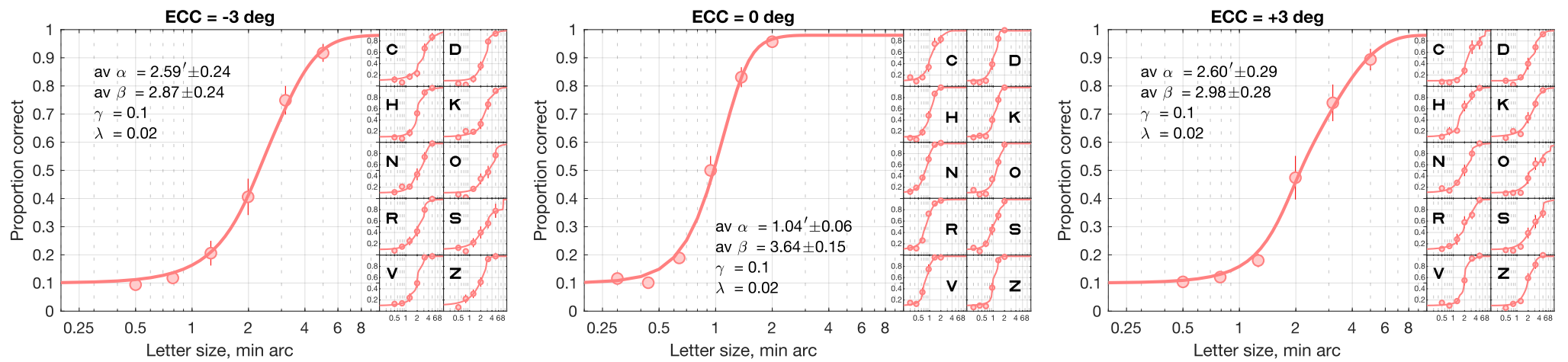


Figure 2.3 The average proportion correct and PMF across observers.

Large panels show the results for the pooled data across letters at the three locations. α , β , γ and λ denote the threshold (i.e., letter acuity defined as the letter size in min or arc that gives a performance of 54% proportion correct), the slope of the PMF, the guessing rate, and the lapse rate respectively. The error bars represent the SE of the average. The small marginal panels show the proportion correct and PMF for individual letters averaged across observers.

Table 2.2 shows the central tendency and distribution of individual letter acuities (i.e., letter size in min or arc) across observers at the three locations.

	ECC = -3 deg				ECC = 0 deg				ECC = +3 deg			
	Mean \pm SE	Median (IQR)	25th Q	75th Q	Mean \pm SE	Median (IQR)	25th Q	75th Q	Mean \pm SE	Median (IQR)	25th Q	75th Q
C	3.10' \pm 0.48	2.70' (1.77)	2.02'	3.79'	1.26' \pm 0.18	1.07' (0.60)	0.92'	1.52'	3.29' \pm 0.56	2.46' (2.67)	2.02'	4.69'
D	2.66' \pm 0.20	2.66' (0.89)	2.23'	3.12'	1.08' \pm 0.08	1.07' (0.45)	0.89'	1.34'	2.77' \pm 0.32	2.73' (1.88)	1.90'	3.78'
H	2.20' \pm 0.19	2.01' (0.77)	1.80'	2.57'	1.06' \pm 0.07	1.00' (0.37)	0.93'	1.30'	2.04' \pm 0.28	1.81' (1.07)	1.37'	2.44'
K	2.81' \pm 0.32	2.54' (1.22)	2.26'	3.48'	1.14' \pm 0.08	1.09' (0.40)	0.95'	1.35'	2.60' \pm 0.31	2.43' (1.34)	1.78'	3.12'
N	2.46' \pm 0.26	2.56' (1.62)	1.52'	3.14'	0.80' \pm 0.04	0.82' (0.19)	0.73'	0.92'	2.19' \pm 0.26	2.14' (1.65)	1.47'	3.12'
O	3.60' \pm 0.47	3.70' (2.87)	2.11'	4.98'	1.29' \pm 0.09	1.28' (0.34)	1.05'	1.39'	3.86' \pm 0.76	3.10' (2.54)	2.02'	4.56'
R	2.33' \pm 0.21	2.38' (0.85)	2.00'	2.85'	0.87' \pm 0.05	0.82' (0.20)	0.78'	0.98'	2.28' \pm 0.29	2.01' (1.26)	1.71'	2.97'
S	3.49' \pm 0.79	2.73' (1.91)	1.63'	3.54'	1.13' \pm 0.12	1.04' (0.48)	0.86'	1.34'	3.22' \pm 0.46	2.67' (3.22)	1.84'	5.06'
V	1.98' \pm 0.18	2.00' (0.29)	1.78'	2.07'	0.81' \pm 0.06	0.74' (0.26)	0.67'	0.93'	2.03' \pm 0.19	1.95' (0.28)	1.74'	2.02'
Z	1.92' \pm 0.21	1.82' (1.14)	1.36'	2.50'	1.06' \pm 0.06	0.94' (0.24)	0.93'	1.17'	2.13' \pm 0.19	1.85' (1.00)	1.69'	2.69'

The analysis revealed statistically significant Pearson correlations of letter acuities between the lower visual field and the central location ($r = 0.71$, $n = 10$, $p < .05$), between the upper visual field and central location ($r = 0.80$, $n = 10$, $p < .05$) and between the lower visual field and upper visual field ($r = 0.95$, $n = 10$, $p < .001$). This suggests that the pattern of differences between individual letter acuities is consistent at the central, upper, and lower visual fields (Figure 2.4 A).

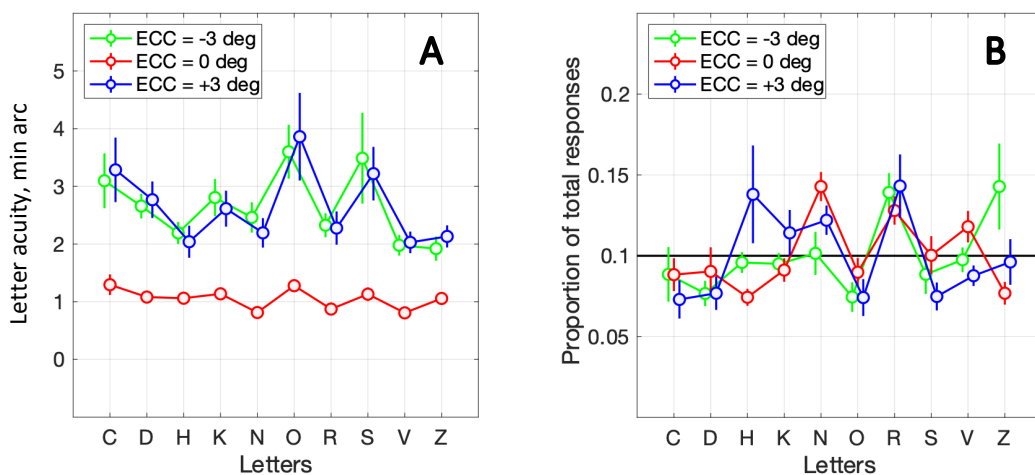


Figure 2.4 Measured acuities and responses for individual letters.

A) the averages of individual letter acuities and B) The letter usage across observers at the three test locations. The letter acuity is expressed as letter size in min of arc. The letter usage is expressed as the proportion of total responses (i.e., either correct or not). The solid black horizontal line shows the expected proportion of letter usage. The error bars are the SE of the average.

Each of the 10 Sloan letters was presented 60 times per location (10 presentations \times six stimulus levels (i.e., letter sizes) per location). Therefore, there were 600 presentations for the 10 letters per location. If observers were equally sensitive to all letters in the 10-letter Sloan set, exhibited no biases to favour some letters more than others, and were not influenced by the similarity between the letters, then all 10 letters would be chosen as a response (correctly or not) with equal expected frequency (i.e., 60 responses out of 600 presentations), namely 10% or 0.1 (Figure 2.4 B). I shall refer

to the relative frequency of letter choice as letter usage (total), and we expect the letter usage to deviate from 0.1 (60 responses per location) when differences mainly in sensitivity, bias and/or similarity exist across the set of letters.

The Chi-squared (χ^2) test showed a significant difference in letter usage (total) among the letters set (i.e., they deviate significantly from 60 responses out of 600 presentations) for each observer at each location (see appendix A for results of χ^2 and p-values). Additionally, each observer showed a unique pattern of letter usage (total) (see Appendix A for plots of letter usage (total) with the binomial confidence intervals for each letter at the three locations). For further analysis, the letter usage for *correct* and letter usage for *incorrect* responses were also calculated for each observer and at each location. The letter usage for correct responses was calculated as the relative proportion of the correct responses given by the letter out of the sum of correct responses given by all the letters at each location. The same method was used to calculate the letter usage for incorrect responses (i.e., the number of incorrect responses given by the letter divided by the sum of incorrect responses given by all the letters).

The analysis showed that the high letter usage (total; i.e., either correct or incorrect) is associated with better letter acuity (i.e., smaller letter size in min of arc) and vice versa as revealed by the negative and strong Pearson correlation between the averages of letter acuities and the averages of the corresponding letter usage (total) across observers at the central ($r = -0.94$, $n = 10$, $p < .001$), upper ($r = -0.87$, $n = 10$, $p = .001$) and lower ($r = -0.80$, $n = 10$, $p = .006$) visual field locations (Figure 2.5 A). Similarly, a strong correlation was found between letter usage for correct responses and letter acuity at the central ($r = -0.98$, $n = 10$, $p < .000$), upper ($r = -0.97$, $n = 10$, $p < .001$) and lower ($r = -0.96$, $n = 10$, $p < .001$) visual field locations (Figure 2.5 B).

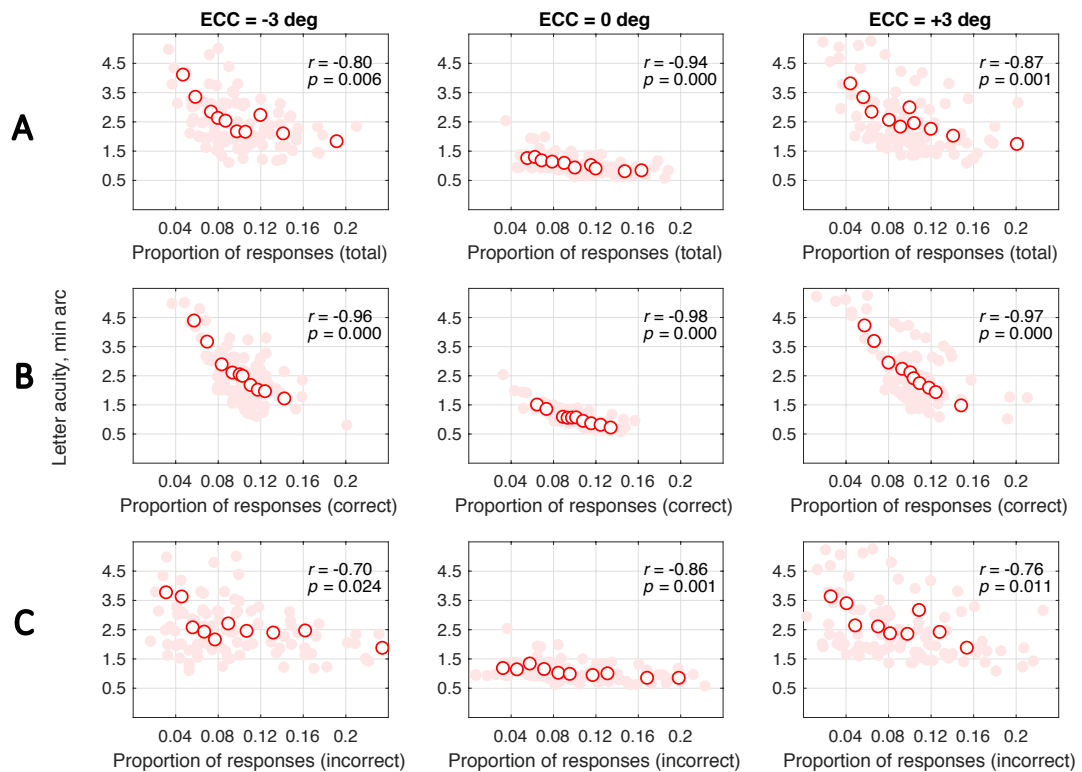


Figure 2.5 Correlation between letter usage and measured acuity.

It shows the correlation between the letter usage of A) total responses vs the letter acuities B) *correct* responses vs the letter acuities and C) *incorrect* responses vs the letter acuities at the three visual field locations. Letter usage is expressed as a proportion of responses. The letter acuity is expressed as letter size in min of arc. Here and throughout, the pink data points are the observers' individual data whereas the red circles are the average across observers.

This was not unexpected; however, the letter usage for *correct* responses was found to be highly associated with the letter usage for *incorrect* responses (Figure 2.6). This was shown by the strong Pearson correlation between the averages of letter usage (*correct*) and the averages of letter usage (*incorrect*) across observers at the central ($r = 0.92$, $n = 10$, $p < .001$), upper ($r = 0.90$, $n = 10$, $p < .001$) and lower ($r = 0.84$, $n = 10$, $p = .003$) visual field locations.

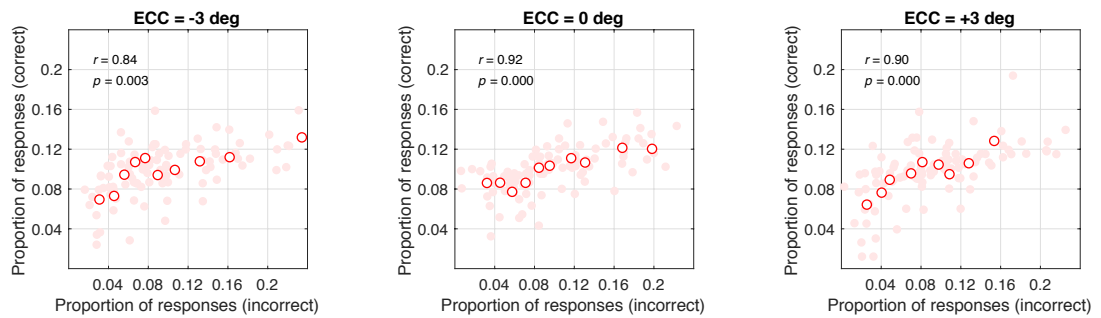


Figure 2.6 The correlation in the proportion of correct vs incorrect responses.

This analysis suggested that; in addition to being influenced by the letter usage (correct) (Figure 2.5 B); the individual letter acuities are also influenced by the letter usage (incorrect). This was demonstrated by the strong Pearson correlation between the averages of letter acuities and the averages of letter usage (incorrect) across observers at the central ($r = -0.86$, $n = 10$, $p = .001$), upper ($r = -0.76$, $n = 10$, $p = .011$) and lower ($r = -0.70$, $n = 10$, $p = .024$) visual field locations (Figure 2.5 C).

The differences in letter usage for incorrect responses are determined by either response bias (towards or against some letters) or similarity between letters. Luce's choice model (1963) was employed to compute the response biases and similarity between letters for each observer to further investigate their relative contribution to the differences of individual letter acuities. For the following analyses, confusion matrices (presented vs responses of letters) were created for individual observers (pooled across the six letter sizes) at each location (see Appendix A). Figure 2.7 shows the confusion matrices averaged across observers at each location (pooled across the six letter sizes for each observer and then averaged across observers).

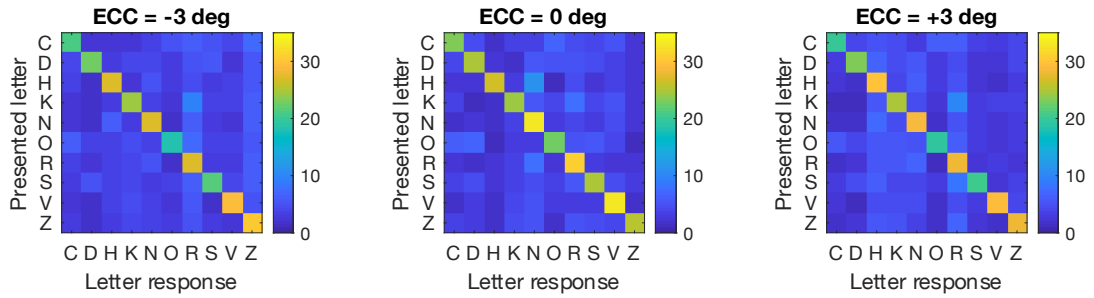


Figure 2.7 The confusion matrices of presented *vs* responded letters. Data is averaged across observers for each location. The colour scale illustrates the frequency of letter responses, where warmer colours show higher frequencies. The diagonal cells represent correct responses, whereas the non-diagonal cells represent incorrect responses.

Luce’s choice model was fitted to letter responses for each observer at each location. Figure 2.8 summarises the results of fitting Luce’s choice model (i.e., response biases (β) and similarity (η) parameters).

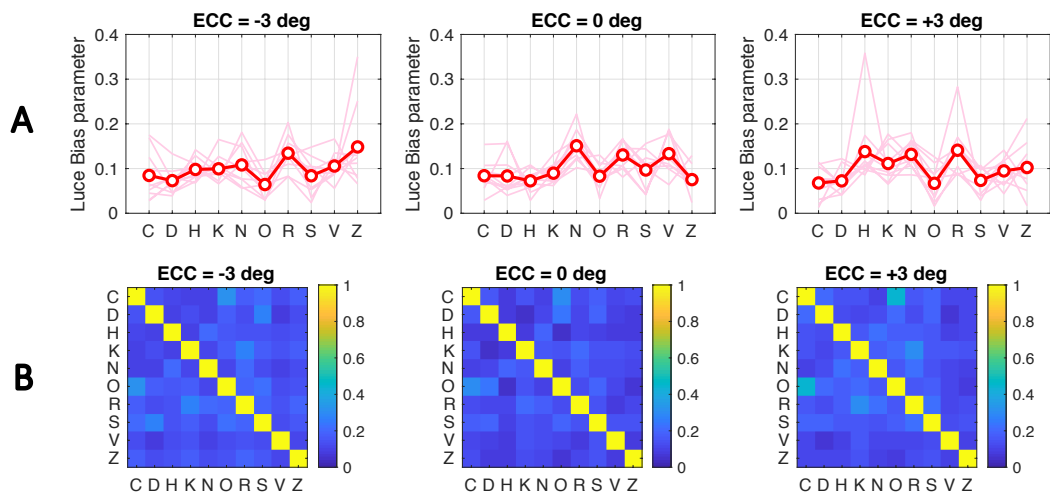


Figure 2.8 Results of Luce’s model fitting. A) the response biases parameters (β) for individual letters and B) the letter similarities parameters (η) for letter pairs calculated by the Luce choice model at each location. Here and throughout, the thin pink curves represent the individual observers’ data (i.e., 10 curves for 10 observers). The red curve represents the average across observers. The colour scale illustrates η parameters, where warmer colour shows higher η parameters (i.e., higher similarity between the corresponding letters of the pair). η values of 1 (i.e., diagonal yellow cells) represent completely identical letters.

For each letter, the average across the similarity parameters of each letter with the other letters was calculated and referred to as letter *confusability* (Figure 2.10). For the next analysis, the letter confusability was used as a measure of the similarity of the letter to the other letters.

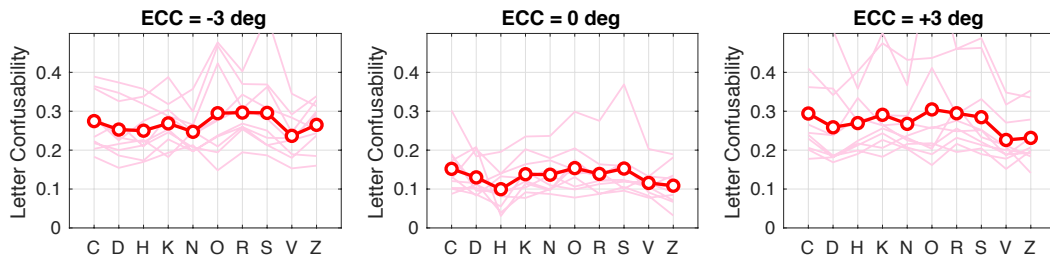


Figure 2.10 The letter confusability at the three locations.

Figure 2.11 A shows that the response bias was strongly correlated (Pearson correlation) to letter acuity at central ($r = -0.96$, $n = 10$, $p < .001$), upper ($r = -0.90$, $n = 10$, $p < .001$) and lower ($r = -0.86$, $n = 10$, $p = .001$) visual field locations. The correlation was weaker between letter confusability and letter acuity at central ($r = 0.50$, $n = 10$, $p = .141$), upper ($r = 0.64$, $n = 10$, $p < .05$) and lower ($r = 0.64$, $n = 10$, $p < .05$) visual field locations (Figure 2.11 B). These results suggest that response bias might highly influence the estimation of individual letter acuities compared to letter similarity.

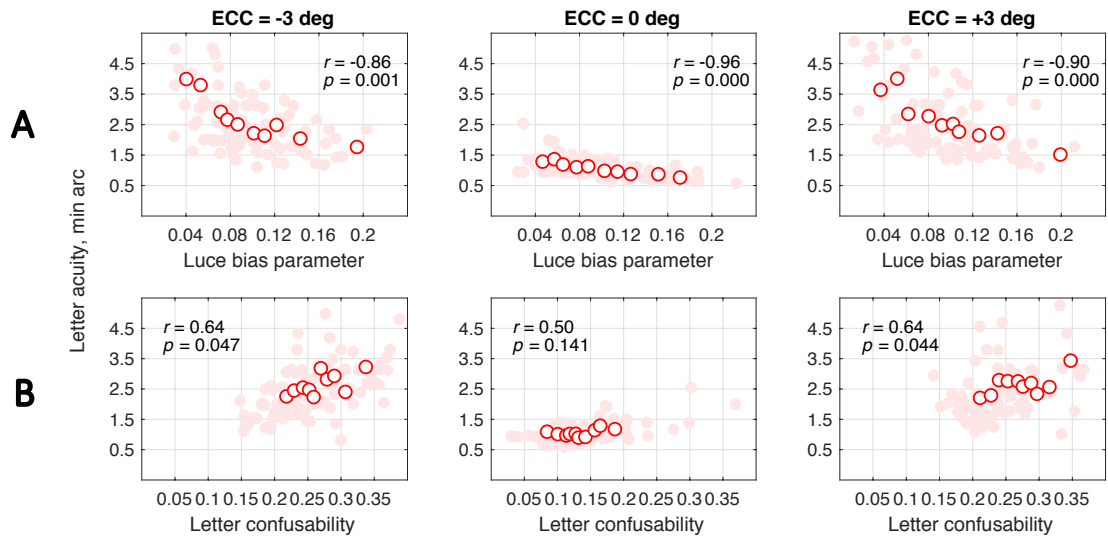


Figure 2.11 Correlation between Luce's parameters and letter acuity.

A) the correlation between the averages of Luce bias parameters and the averages of letter acuities across observes. B) The correlation between the averages of letter confusability and the averages of letter acuities across observes.

2.4 Discussion

This study aimed to investigate the role of response bias and similarity in the estimation of individual Sloan letter acuities. Results showed that letter acuity was significantly different between the central and the paracentral visual field locations with no significant difference between the upper (+3 deg) and lower (-3 deg) locations. Figure 2.12 A shows the results of the current experiment and two previous studies (Ludvigh, 1941; Hairol et al., 2015). The poorer letter acuities in the current study could be the result of differences in study design. Hairol et al. (2015) used Sheridan Gardiner letters and unlimited viewing time, and Ludvigh (1941) used F, E, C, L, and T letters and did not specify the viewing time.

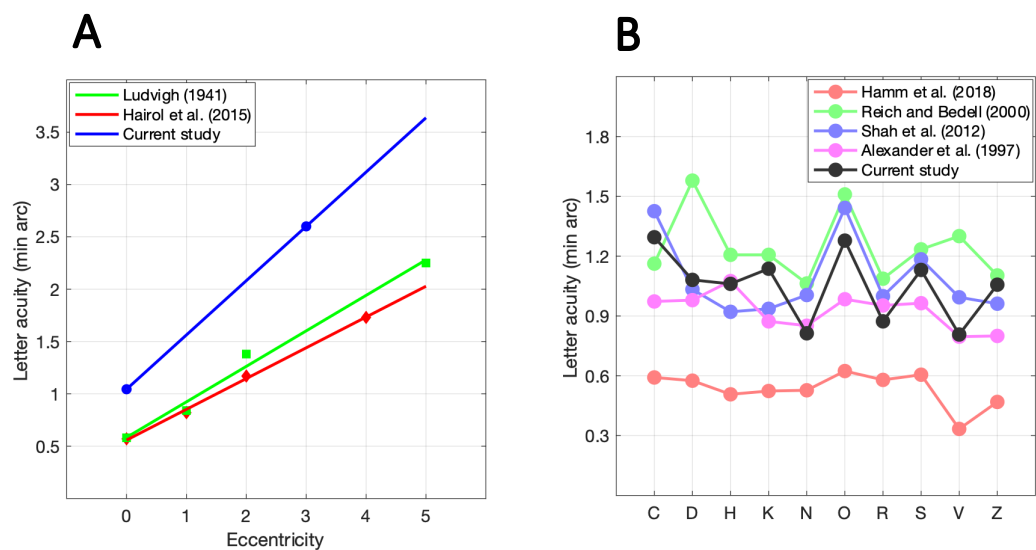


Figure 2.12 Measured letter acuities of the current and previous studies.

A) the average letter acuity across observers as a function of eccentricity for the current and two previous studies (Ludvigh, 1941; Hairol et al., 2015). B) The average of individual letter acuities for the current and four previous studies (Alexander et al., 1997; Reich & Bedell, 2000; Shah et al., 2012; Hamm et al., 2018).

Similar to previous studies, our results show substantial differences in individual letter acuities (Alexander et al., 1997; Reich & Bedell, 2000; Shah et al., 2012; Hamm et al., 2018). Figure 2.12 B shows the individual letter acuities (central location only)

compared to four previous studies (Alexander et al., 1997; Reich & Bedell, 2000; Shah et al., 2012; Hamm et al., 2018). Our results were generally similar to Alexander et al.'s (1997), Reich & Bedell's (2000), and Shah et al.'s (2012) results, but different from Hamm et al.'s (2018) results, who measured higher acuities for all letters. The better acuities in Hamm et al.'s (2018) could be the consequence of using white optotypes on a black background, using an interleaved QUEST staircase procedure, and a potential learning effect that could arise from the long and multiple experimental sessions (10 hours per observer) (Westheimer, 2003; Hamm et al., 2018). On the other hand, Alexander et al. (1997), Reich & Bedell (2000), and the current study employed the method of constant stimuli, and Shah et al. (2012) employed the method of limits. The current study showed similar patterns of the differences in individual letter acuities at the fovea, upper, and lower visual field locations, but with higher variability in individual letter acuities in the upper and lower visual field locations (Figure 2.4 A). This was consistent with what has been reported by Hairol et al. (2015) (see Figure 1.10 A).

Letter usage (total) was found to be influential on the estimation of the individual letter acuities (Figure 2.5 A). As expected, letter usage (correct) was found to be significantly and strongly correlated to the individual letter acuities (Figure 2.5 B). The letter usage (correct) is determined mainly by the sensitivity towards the letters. Interestingly, there was a significant correlation between the letter usage (correct) and (incorrect) (Figure 2.6). Furthermore, it has been found that the individual letter acuities are influenced by the letter usage (incorrect) (Figure 2.5 C). This suggests that; besides being determined by letter sensitivity, the letter usage (correct) might also be influenced by the letter usage (incorrect). The latter is determined by response bias and similarity between letters.

Here I used Luce's choice model (Luce, 1963) to estimate letter similarity and response bias parameters to investigate their relationship with the individual letter acuities of Sloan letters. Most confusion pairs (i.e., similar letters) at the central location were found to be similar to those at paracentral locations (but with different values and ranks of similarity parameters (η)) (Figure 2.8 B & 2.9). The results of fitting the model showed that the pattern of similarity between letters in the current study is consistent with the findings of previous studies (Figure 2.13).

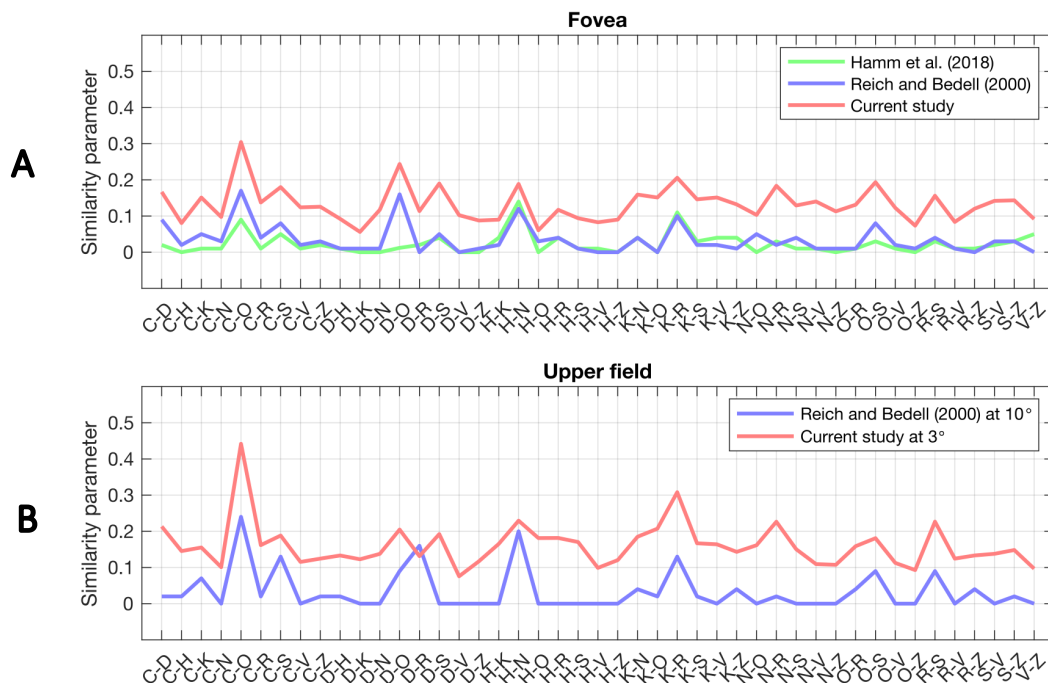


Figure 2.13 The similarity of letter pairs of the current and previous studies. A) at the fovea and B) at the upper field. The similarity is expressed as η parameter for the current and Hamm et al.'s (2018) results. However, the similarity is expressed as the probability of confusion in Reich and Bedell's (2000) results.

The similarity parameters of the current study were significantly correlated (Pearson correlation) with the similarity parameters of Hamm et al. (2018) ($r = 0.61$, $n = 45$, $p < .001$) and Reich & Bedell (2000) ($r = 0.82$, $n = 45$, $p < .001$) at the fovea. A significant correlation (Pearson correlation) was also found at the upper visual field

location ($r = 0.71$, $n = 45$, $p < .001$) between the similarity parameters of the current (+3 deg) and Reich & Bedell (2000) (+10 deg) studies.

Nevertheless, the similarity parameters in the current study were found to be substantially and consistently higher than the similarity parameters in the other two studies. This could be due to differences in study design. In contrast to the current study, where observers were forced to respond only from the Sloan letters set, Reich and Bedell (2000) used 25 letters which would result in different and more combinations of confusion pairs. In this case, observers can exhibit similar combinations to the current study but with weaker similarity strengths (Carkeet, 2001). Hamm et al. (2018) also showed lower similarity parameters (η). As mentioned earlier, this could be the result of using reversed contrast optotypes (white optotypes on a black background) and/or employing different methods to determine the letter acuity. They employed the QUEST method where they present letters mainly at threshold size, hence they did not equally include letter sizes above and below the threshold where higher similarity might occur. The next study (Chapter 3) will investigate this aspect in detail.

I further aimed to investigate the relationship between the letter similarities and the differences in individual letter acuities. The similarity (expressed as confusability) was found to be influential on the individual letter acuities especially at the upper and lower visual field locations (Figure 2.11 B). Hence, the similarity between letters might be a source of the differences in the estimation of individual letter acuities. These findings are consistent with the previous studies (Erdei et al., 2019; Grimm et al., 1994; McMonnies & Ho, 1996; McMonnies & Ho, 2000; Hamm et al., 2018).

The response bias computed by Luce's model in the current study is similar to the pattern of bias parameters presented in Hamm et al. (2018) (fovea) (Figure 2.14).

However, the Pearson correlation was not significant ($r = 0.55$, $n = 10$, $p = .10$).

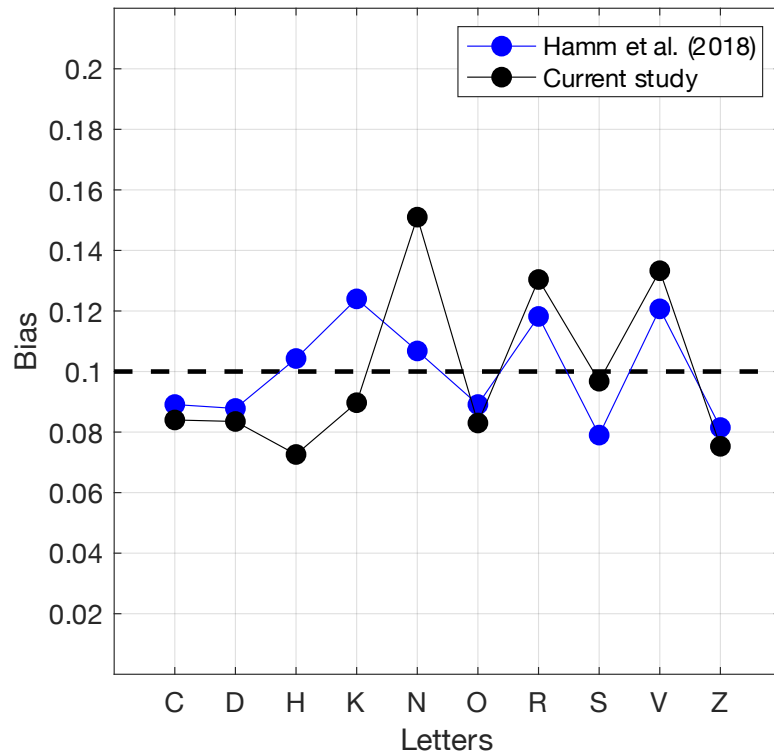


Figure 2.14 The Luce's bias of the current and Hamm et al.'s (2018) studies.

Our analysis demonstrates that the individual letter acuities are substantially influenced by the response bias (Figure 2.11 A). Consequently, compared to the similarity between letters, response bias might have a major role in the differences in individual letter acuities.

In summary, response bias and similarity between letters play a role in the differences of the individual Sloan letter acuities where response bias is the most influential factor. However, the sensitivity towards the letters has not been quantified or investigated in the current study acknowledging that it also might contribute to the differences in individual letter acuities. Furthermore, the confusability in the current study might be regarded as a measure of the sensitivity of the letter making it even more difficult to differentiate between the relative role of sensitivity and similarity on individual letter

acuties (Mueller & Weidemann, 2012). Luce's choice model is not a *generative* model. Therefore, it cannot be used as a simulation tool to investigate the extent of the effect of response bias and similarity on the letter acuity measurements.

In the next chapter, I will introduce a novel generative model to reveal the role of sensitivity, response bias, and similarity in the letter identification task simultaneously and the relative contribution of each one of the three factors on the estimation of the individual letter acuties.

Chapter 3 . The Role of Bias, Sensitivity and Similarity in Letter Acuity Measurements: A Noisy Template Model

3.1 Motivation

In the following chapter, I will investigate the relative role of response bias, sensitivity differences, and similarity between letters simultaneously in letter identification task using the same data (Chapter 2). Here, I introduce a novel extension (by incorporating the similarity factor) to our previously derived model (Georgeson et al., 2023). In our previous model (i.e., Noisy Template Model (NTM)), only two factors, (i) response bias and (ii) sensitivity differences between letters were investigated in the letter identification task using the same data set from the previous experiment (Chapter 2). Since the NTM is a generative model, it will be used as a tool for further simulations to demonstrate the relative role of each one of the three factors in estimating the letter acuity (in particular, the systematic error and test-retest variability of the estimated letter acuity).

3.2 Methods

In the current study, the data from the previous experiment (Chapter 2) were re-analysed. In the previous study, the data were collected in the form of a confusion matrix of the presented *vs* responded letters for each observer and for each location. Here I will focus on elaborating on the details of the model (NTM) employed to explore the data.

3.2.1 The noisy template model (NTM)

3.2.1.1 Letter usage

In Chapter 2, I showed that the total letter usage (the frequency with which each letter is used as a response, regardless of correctness) was significantly different between letters for each observer and test location (see Figure 3.1 A, which is similar to Figure 2.4 B, for the results across observers and Appendix A for the results of individual observers). We expect the letter usages to deviate from 0.1 when differences mainly in bias or sensitivity exist across the set of letters. However, several serious issues need to be resolved before any inferences about bias (or sensitivity) can be made, as Figure 3.1 illustrates.

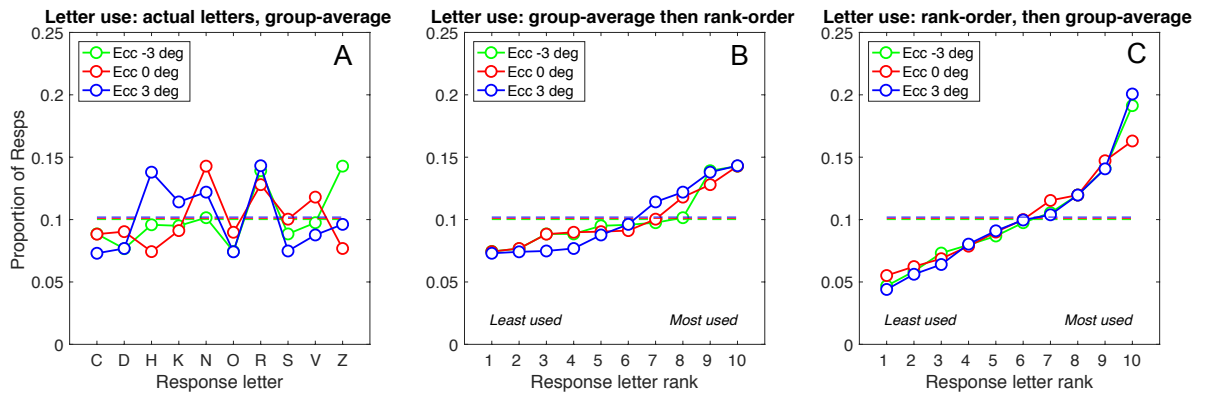


Figure 3.1 Data overview: letter usage.

A) letter usage (total), expressed as proportion of responses, in the experiment on which each of the 10 letters was reported (correctly or not), averaged over the six letter sizes and 10 observers. Each point is based on 600 trials. B) Same data as A, but rank-ordered by group-average proportion of responses, separately for the three locations. Because the ordering of letters was different for the three locations, we must label the x-axis in terms of letter rank (1-10) rather than letter identity. C) Similar to B, but the proportion of responses were rank-ordered separately for each observer, then averaged over observers.

Figure 3.1 A shows the pattern of letter usage (expressed as proportion of responses) averaged across observers, for each of the three test locations in the experiment. In central vision (red) N, R, V appear to be over-called, while at -3 deg eccentricity (green) it was R and Z, and at +3 deg eccentricity (blue) it was H, N, R. Figure 3.1 B shows the same data, but rank-ordered in terms of group-average usage, from least to most. These variations in letter usages might reflect different biases (or sensitivities) for different letters. But if observers have different *patterns* of bias (or sensitivity) across letters, then averaging over observers *before* rank-ordering is likely to underestimate the range of individual biases (i.e., it washes out the effect we are aiming to capture). A potential answer to this problem is to rank-order the letter usage of individual observers first, and then average the ranked letter usages across observers, as shown in Figure 3.1 C. We used this form of averaging in all our analyses. Note how the range of variation around the unbiased mean (0.1) has approximately doubled

from Figure 3.1 B to 3.1 C, and how similar the trends are for the three test locations.

3.2.1.2 Outline of the model

The model presented here is an extended version of our previously devised model (i.e., NTM) (Georgeson et al., 2023) (Figure 3.2). In our original model, we made two simplifying assumptions about the templates. The first is that there are as many letter templates as there are test letters in the experiment (i.e., 10 templates), and secondly, the templates are assumed to give a response only to their own preferred letter, with no response to other letters (Figure 3.3 A); that is, the templates are orthogonal. Macmillan and Creelman (2005) noted that the assumption of orthogonality between detectors is *optimistic* but appealing because of the simplicity it confers on the modelling (Macmillan & Creelman, 2005). However, in the current version of the model, the second assumption of the NTM (i.e., the orthogonality assumption in the original model) has been changed as follows: the templates respond not only to their preferred letter but also to other, similar letters. That is, the templates are correlated (i.e., non-orthogonal).

When a given letter is presented, the output of each letter's template (i.e., Template responses in Figure 3.2) is subject to bias, sensitivity, and/or similarity to the presented letter. Furthermore, the net output of each template is perturbed by additive Gaussian noise. The bias affects the mean output level of the template by a constant amount whether or not the template's preferred letter is presented (Figure 3.3 B). The sensitivity affects the mean output level of the template when the preferred letter is presented (Figure 3.3 C). The similarity influences the mean output of the non-preferred letter templates according to their correlation to the preferred letter template

(i.e., similar letters) (Figure 3.3 D).

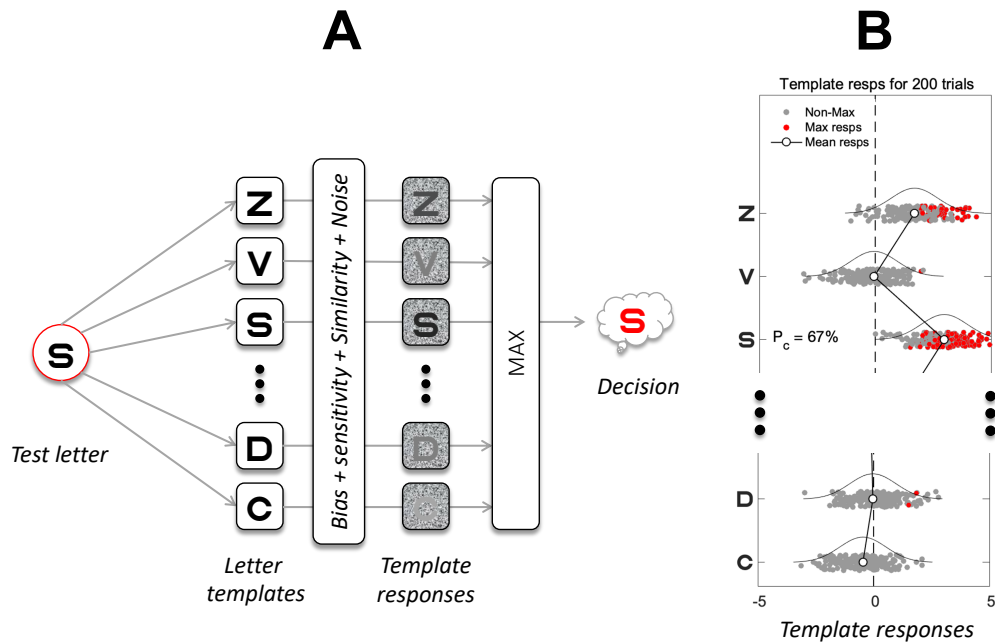


Figure 3.2 A) a schematic representation of the NTM and (B) template responses.

After adding the bias, sensitivity, similarity, and noise on the corresponding templates, the most active (i.e., MAX response) template on a given trial determines the letter choice. Because of noise, the most active template over trials (red dots in B) may be the correct (e.g., S) or the incorrect letter (e.g., Z). The joint effect of the three factors, bias, sensitivity and similarity determines the mean response of each template. For example, positive bias (top three rows) gives a higher mean response compared to negative bias (bottom two rows), hence, increasing the chance of incorrect responses (e.g., red dots for Z) and correct responses (i.e., S). Note that because of the dissimilarity between letter V and S, the mean response of letter V was low (although it has positive bias), therefore, it is unlikely for the template V to be chosen (as an incorrect response in this case). Negative biases (bottom two rows) decrease the chance of these letters being called, correctly or not. However, the similarity between the letters S and D increases the likelihood of letter D (although it has a negative bias) being chosen (as an incorrect response in this case).

For simplicity, the model does not further specify the nature of the template-matching process, but a detailed model for that process, incorporating optical and neural filtering, along with the limitations imposed by spatial sampling and noise, was developed by Watson and Ahumada (2015). They noted that all image-based models of letter identification to date are based on the template concept (Watson & Ahumada, 2015).

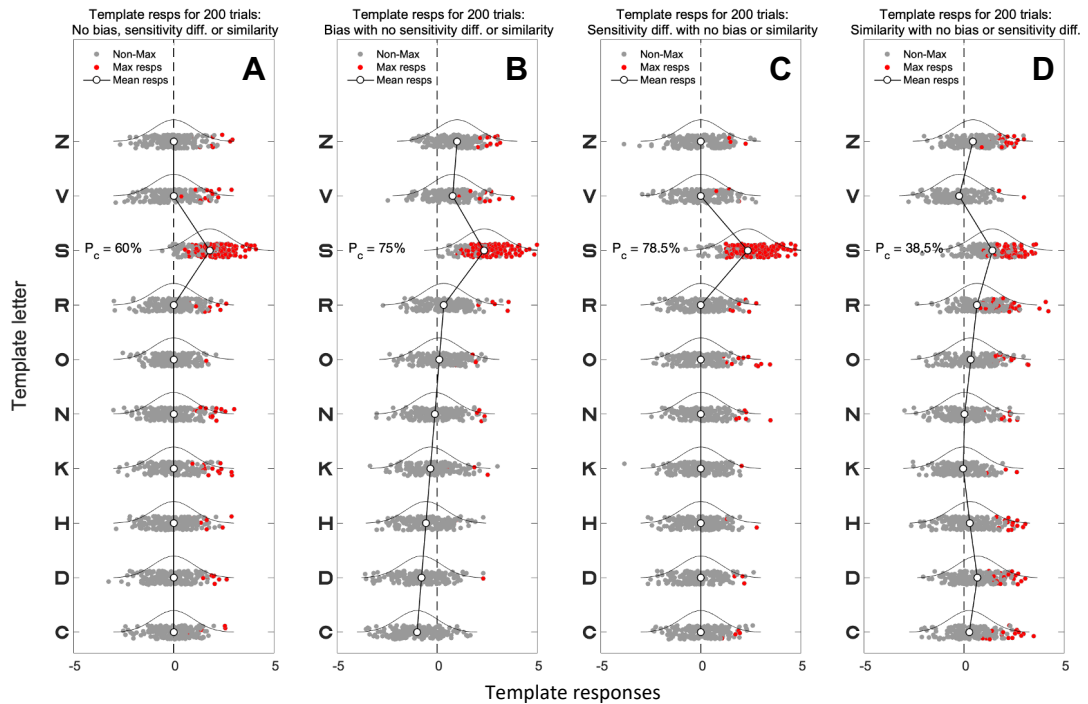


Figure 3.3 Demonstration of NTM in case of bias, sensitivity, similarity or none.

Figure illustrates the (A) NTM without bias, sensitivity differences, or similarity between letters. (B) NTM with different biases on each template. (C) NTM with different sensitivity on each template. (D) NTM with a similarity between templates. The bias and sensitivity in panels B & C are ordered from negative to positive with a mean of zero. I call these the bias gradient B' and the sensitivity gradient S' (see text). The letter identity for each template here is arbitrary, for illustration only. Presenting a test letter (e.g., S) increases the mean response for the S-template but leaves others unchanged (A). The letter decision on a given trial is made by choosing the template with the largest response on that trial (the MAX operator, Figure 3.2 A). Choices vary from trial to trial because of noise in each template channel (indicated by the Gaussian distribution). For illustration purposes, the red points represent those trials on which a given template gave the max response, whereas the grey points are activations that were lower, hence not chosen. Notice how the positive bias (e.g., Z), and presentation of the preferred letter (e.g., S) increase the likelihood of choosing certain letters, sometimes correctly, sometimes not (B). In case of sensitivity differences (C), larger sensitivity increases the chance of the letter template to be chosen when its preferred letter is presented (i.e., increase the likelihood of correct responses as shown by the increase of the percent correct response (P_c) from 60% to 78.5%). Panel D shows the case with no bias or sensitivity differences but with the similarity between the presented letter S and the other templates. Note that the similarity between the presented letter S and the other templates (e.g., D and Z) decreases the likelihood of choosing the letter S (i.e., P_c decreased from 60% to 38.5%) and increases the likelihood of choosing the similar letters such as D and Z (i.e., incorrect responses). On the other hand, the dissimilarity between the letters S and V decreases the likelihood of choosing the letter V (i.e., incorrect responses).

3.2.1.3 Model structure and equations

For bias and sensitivity modelling, the structure and equations of the original model remained the same as follows: If differences in letter usages (total) are caused by biases, then the most-used letter would be associated with the highest positive bias and vice versa. Suppose that the templates can be rank-ordered from least- (most-negative) bias to most-positive bias, indexed by $i = 1$ to m (where $m = 10$). Furthermore, bias values B_i (assigned to templates i to m) are assumed to be a linear function of i , ranging from $B_1 = -B'$ to $B_m = B'$. It is a strong but a simplifying assumption. This linear gradient is a good place to start because it greatly reduces the number of free parameters per observer (i.e., one bias parameter (B') per location instead of nine).

Therefore, the bias B_i for the i^{th} template is

$$B_i = \frac{B'(2i - m - 1)}{m - 1}. \quad (\text{Eq. 3.1})$$

where B' is a free parameter. Subsequently, the bias values B_i are assigned to the corresponding templates according to their letter usages (total). Therefore, the highest bias value is assigned to the template of the most used letter and vice versa. The same procedure is followed to assign the rest of the bias values to the corresponding templates. Note that when $B' = 0$, the system is unbiased.

Similarly, we assumed that sensitivity differs (a linear variation) between templates around a baseline value (i.e., overall sensitivity S_0 which is a free parameter) ranging from $S_0(1 - S')$ to $S_0(1 + S')$ where S' is the sensitivity gradient. Therefore, the sensitivity of the i^{th} template to the j^{th} test letter is

$$S_{ij} = S_0 \left(1 + \frac{S'(2i - m - 1)}{m - 1} \right) \text{ if } i = j. \quad (\text{Eq. 3.2})$$

We shall refer to B' and S' as the *bias gradient* and *sensitivity gradient* respectively. When fitted to data, B' and/or S' were free parameters for individual observers, and if not fitted then B'=0 and/or S'=0. S₀ was also a free parameter for each observer and test location, controlling the overall level of performance.

For the similarity simulations and modelling, I assumed that the similarity between Sloan letters arises solely from the features of the letters. Hence, the similarity between pairs of Sloan letters can be quantified by the Optotype Correlation (OC) matrix (Figure 3.4) (Fülep et al., 2017). The calculation is based on Pearson's normalised cross-correlation of the original letters (Neto et al., 2013). To simplify the model, I propose that similarity is a symmetrical relation (e.g., if H is similar to K, then K is similar to H by the same amount). I also propose that the pattern in the OC matrix does not change for different letter sizes.

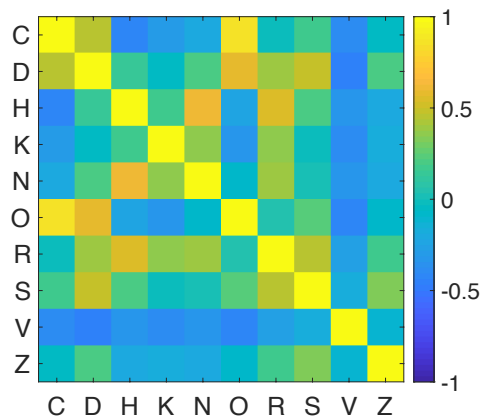


Figure 3.4 The optotype correlation matrix (OC) for Sloan letters. A correlation value of one indicates identical letters, zero means no correlation, and a negative correlation indicates dissimilar letters. A higher positive value means stronger similarity between the corresponding letters.

Another key assumption is that, while keeping the similarity pattern unchanged, the overall strength of the similarity between letters can vary from observer to observer. This is controlled by multiplying the off-diagonal correlation values by a factor that we called the Confusion Strength (C_s).

Using the OC values and the C_s factor to quantify the similarity between Sloan letters for a given observer, the sensitivity of the i^{th} template to the j^{th} test letter is

$$S_{ij} = (S_0 \times C_s \times OC_{ij}) \quad \text{if } i \neq j \quad (\text{Eq. 3.3})$$

where C_s is a free parameter. Thus, the sensitivity of the template i when $i \neq j$ is controlled by its correlation with the letter j (i.e., OC_{ij} , see Figure 3.4) and by the C_s for a given observer. For example, consider test letter C and suppose $S_0 = 1$ and $C_s = 1$. In this case, the sensitivity of the letter C template to test letter C is 1, while sensitivity of the letter O template is 0.866, but that of the (rather dissimilar) letter H template is -0.408. Similar considerations apply to the rest of the templates.

Letter identification improves with increasing letter size. I capture this essential behaviour by assuming that the template's mean response (μ_{ij}) increases as a function of sensitivity S_{ij} , bias B_i , and test letter size t (expressed as letter stroke width, in min arc)

$$\mu_{ij} = B_i + (S_{ij} \times t)^p \quad (\text{Eq. 3.4}),$$

where p is a constant exponent of a power function relation between mean response μ and letter size t . The exponent p controls the slope of the underlying psychometric

function (PMF) (proportion of correct responses *vs* test letter size), and from initial explorations of model and data, we set $p = 2.5$ for central vision, and $p = 2$ at ± 3 deg eccentricity. When $p = 2$ the signal (μ) underlying correct performance increases with the square of the letter size. We might interpret this as arising from physiological nonlinearities (e.g., the half-squaring model of V1 cell responses) (Heeger, 1992). Furthermore, I adopted the standard SDT assumption that the i^{th} template output is perturbed by additive zero-mean Gaussian noise with variance σ_i^2 , and that $\sigma_i = 1$ for all i . The decision rule for letter identification was the MAX operator. On a given trial, the template that has the highest activity is chosen as the response letter (Figure 3.2 A).

3.2.1.4 The MAX operator

We assume that the observer's decision about which letter was shown on a given trial is determined by whichever template is most active (Figure 3.2 A). This winner-take-all process, often known as the "MAX operator", has a long history in psychophysics and pattern recognition. Watson and Ahumada (2015) used the MAX over templates as the decision rule for letter identification in their Neural Image Classifier model but did not address possible effects of bias. When all the templates are equally sensitive ($S' = 0$), unbiased ($B' = 0$) with no similarity ($C_s = 0$), the MAX rule is the optimal decision rule (Kingdom & Prins, 2010; Wickens, 2002). We think it remains a reasonable assumption even in the face of some bias and unequal sensitivity or similarity across templates. This amounts to assuming that the decision-making apparatus has no information about variation in S_{ij} or B_i and so cannot devise a better decision rule than the MAX rule.

3.2.1.5 Letter identification

In the modelling of bias and sensitivity, a more positive μ_{ij} increases the likelihood of the corresponding letter to generate the MAX response (Figure 3.3 B & C). On the other hand, in the similarity modelling, when a test letter is presented, the mean response becomes more positive for the test letter template but also for similar letters. They will hence be more likely to generate the MAX response across templates, and so will tend to be *confused* with the test letter. For letters dissimilar to the test letter, their templates' mean responses shift in the opposite direction, making them less likely to be confused with the test letter. For example, similar letters like S and D are likely to be confused with each other, whereas dissimilar letters S and V are unlikely to be confused (Figure 3.3 D). Note that in Eq. 3.3, if C_s is zero, S_{ij} becomes zero when $i \neq j$. This means that our original NTM presented in Georgeson et al. (2023) is a special case of the present model, with zero C_s , hence orthogonal templates.

I used a *Matlab* function (`M_stats_maxN`) to compute (for all i,j) the probability P_{ij} that the i^{th} template delivers a response to the j^{th} test letter that is larger than that of all the other templates. `M_stats_maxN.m` is freely available from the Journal of Vision as supplementary material to a paper by Zhou et al. (2014) at <https://doi.org/10.1167/14.13.24>. (Zhou et al., 2014). For a given set of parameters (p , S_0 , B' , S' , C_s), I used this function to compute the matrix of presented *vs* response probabilities from which I calculated the proportions of responses of each letter (correct or incorrect) at different letter sizes. i.e., I obtained one 10×10 confusion matrix (test letters *vs* response letters) for each tested letter size and each test location.

3.2.1.6 Illustration of the effect of bias, sensitivity and similarity on letter identification task

To illustrate the expected effect of bias, sensitivity, and similarity between letters on the correct and incorrect response pattern for individual letters, I ran simulations using parameters that mimic the letter identification task at the fovea ($p = 2.5$, $S_0 = 1.15$), firstly with three magnitudes of bias ($B' = 0, 0.3$ or 0.6) but no sensitivity differences or similarity between letters, secondly with three values of sensitivity differences ($S' = 0, 0.15$ or 0.3) but no bias or similarity, and finally with three values of similarity ($C_s = 0, 0.5$ or 1) but no bias or sensitivity. Figures 3.5, 3.6 & 3.7 show the modelled PMFs for individual letters in the three cases and the confusion matrices pooled across the six letter sizes used in the experiment.

Figure 3.5 shows the proportions of correct responses for individual letters and the confusion matrices for three different degrees of bias but no sensitivity differences or similarities between letters. Figure 3.5 A shows the case with no bias; the modelled PMFs for different letters are identical (complete overlap). Additionally, there is no over-calling or under-calling of any of the letters as indicated by the letter responses in the corresponding confusion matrix. However, in Figures 3.5 B & C, the biases added to the letter templates shifted the corresponding PMFs systematically. The simulation shows that when the largest bias is assigned to the letter Z (chosen arbitrarily for demonstration purposes) the letter Z will be over-called as shown by the corresponding confusion matrix, and the proportion correct will be higher than that of the other letters as indicated by the upwards shift of its PMF. The opposite is observed for the letter with the negative bias (i.e., V). The maximum effect is near the lower asymptote where the templates are not strongly stimulated such that the responses will

be determined mainly by bias and noise. The more extreme the bias, the larger the effect (Figure 3.5 B & C).

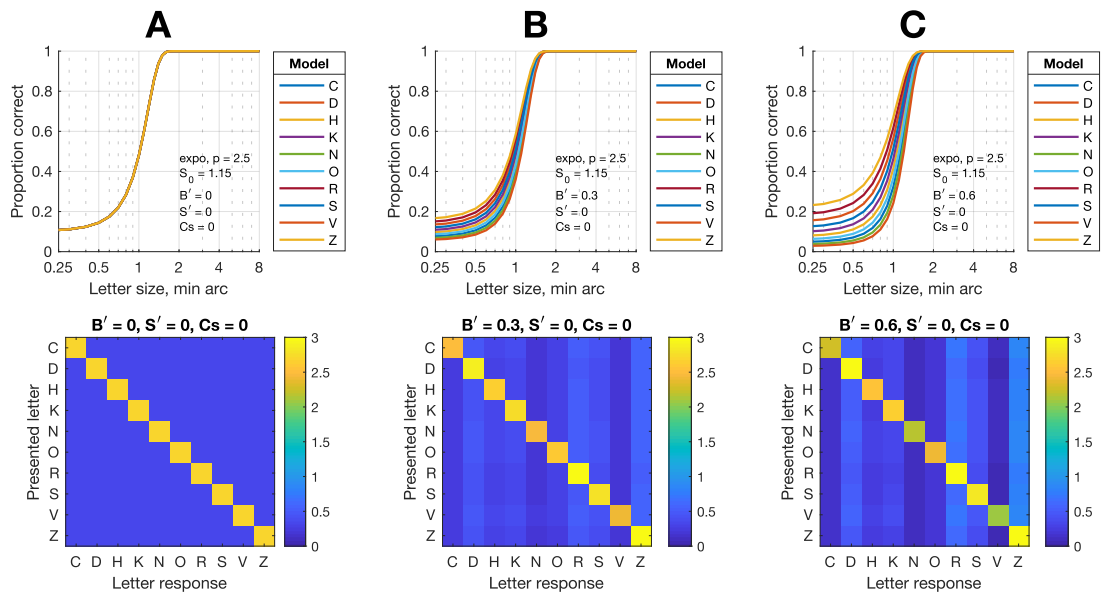


Figure 3.5 Simulation of bias effect on letter identification performance.

The top row shows three plots of simulated bias effect on psychometric functions of individual Sloan letters. B' values are 0, 0.3, and 0.6. The bottom row shows the corresponding confusion matrices pooled across the six letter sizes. The order of the letters' biases values is arbitrary for demonstration purposes only.

Figure 3.6 shows the effect in case of different letter sensitivity differences expressed as sensitivity gradient (i.e., S') of 0.15 and 0.3. Higher sensitivity towards a given letter increases the proportion correct at all the presented letter sizes compared to the other letters. Consequently, the resulting PMF shifts horizontally to the left (e.g., Z) and vice versa (e.g., V). The sensitivity differences between letters only affect the letters' correct responses (i.e., the diagonal cells in the corresponding confusion matrices; Figure 3.6 B & C). In this case, over-calling a given letter will be due to the increased sensitivity towards that letter when presented (i.e., high correct responses) compared to the other letters.

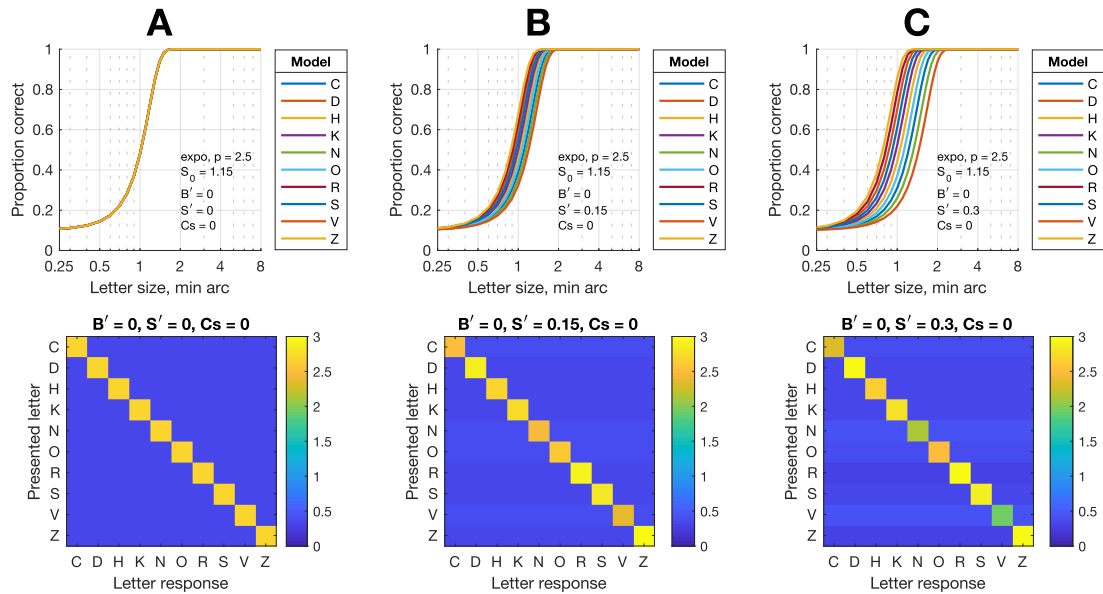


Figure 3.6 Simulation of sensitivity effect on letter identification performance.

The top row shows three plots of simulated sensitivity on psychometric functions of individual Sloan letters. S' values are 0, 0.15 and 0.3. The bottom row shows the corresponding confusion matrices.

Finally, Figure 3.7 shows the performance in a simulated letter identification task with three different degrees of similarity but no bias or sensitivity differences between letters. Figure 3.7 A (no similarity) shows the same result as Figure 3.5 A & 3.6 A. Figures 3.7 B & C show the cases with two different degrees of similarity. The stronger the similarity (C_s) of one letter to another (e.g., C vs O), the lower the proportion correct. Furthermore, the confusion matrices show patterns similar to the OC pattern used in the simulation (Figure 3.4), and this pattern is clearer with larger C_s (Figure 3.7 C).

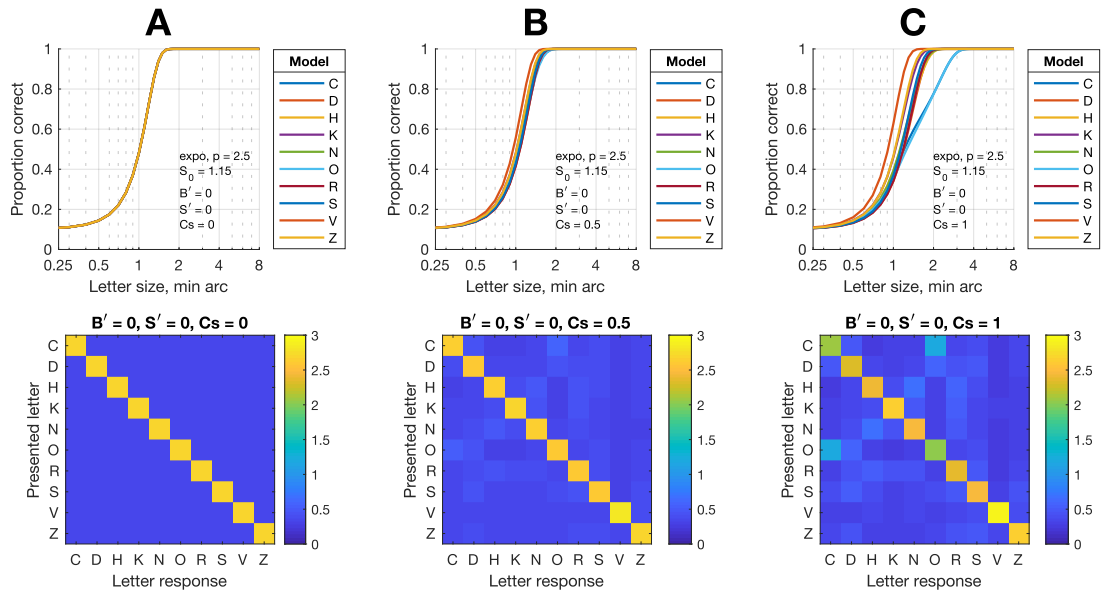


Figure 3.7 Simulation of similarity effect on letter identification performance
 The top row shows three plots of simulated similarity on psychometric functions of individual Sloan letters. C_s values are 0, 0.5 and 1. The bottom row shows the corresponding confusion matrices.

3.2.1.7 Model fitting

For each observer, I obtained a 10×10 (presented *vs* responded) confusion matrix for each of the six letter sizes and three test locations. The model was fitted to the observed number of responses (i.e., 600 responses at each location), using maximum likelihood—adjusting parameter values S_0 , B' , S' , and C_s to maximise the log-likelihood (LL) of the parameters given the data. To calculate the LL, I used the following equation:

$$LL = \sum \{n_{ij} \times \log(P_{ij})\} \quad (\text{Eq. 3.5})$$

where n_{ij} is the observed number of responses of the letter i when presenting the letter j . P_{ij} is the model's proportion of responses for the corresponding ij pair. The variables n and P range over 100 letter pair combinations for each letter size (There are six letter sizes per test location), and the summation takes place over those 600 pairs.

I fitted eight versions of the template model, in which the fitted parameters represented B' , S' & C_s , denoted as the model (B1C1S1), or B' & C_s (B1C1S0), B' & S' (B1C0S1), C_s & S' (B0C1S1), B' only (B1C0S0), C_s only (B0C1S0), S' only (B0C0S1) or none (B0C0S0). The fitting of the B1C1S1 model was performed in two steps. In the first step, I adjusted overall sensitivity S_0 to find the best-fitting value that maximised LL assuming that there is no bias ($B' = 0$), no sensitivity differences ($S' = 0$), and no similarity between letters ($C_s = 0$). In the second step, while using the best S_0 , I ran a 3D sampled grid search to find the best values of the B' , C_s , and S' . Then I readjusted for S_0 using the best B' , C_s , and S' . I repeated the procedure until there was no change in the estimated parameters.

For the remaining models, 3D sampled grid searches were performed to fit the B1C1S0, B1C0S1, and B0C1S1 models to find the best values of the (S_0 , B' & C_s), (S_0 , B' & S') or (S_0 , C_s & S') respectively. A 2D sampled grid search was performed to fit the B1C0S0, B0C1S0, and B0C0S1 models to find the best values of the (S_0 & B'), (S_0 & C_s) or (S_0 & S') respectively. A 1D sampled grid search was performed to fit the B0C0S0 model to find the best value of the S_0 . All the fittings were performed for each observer and each test location separately.

In the following results section, I present the results of fitting the eight models to the experimental data and comparison between them qualitatively and formally (i.e., via Akaike Information Criterion (AIC) (Akaike, 1974)) to determine the best fitting model.

3.3 Results

3.3.1 Initial observations

The results of experimental data are summarised in Figure 3.8 for the three test locations. Figure 3.9 A shows the resultant confusion matrices of the fitted models collapsed across the six letter sizes and then averaged across observers; each row shows the fit of one model (the basic model B0C0S0 is not included). To facilitate the identification of the model that accounts best for experimental data (Figure 3.8), the difference between each model and experimental data was calculated using the Chi-squared (χ^2) value. For each model and each location (Figure 3.9 A), χ^2 value was calculated as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (\text{Eq. 3.6})$$

where O_{ij} is the frequency of responses of the cell ij in the confusion matrix of experimental data (i.e., observed) and E_{ij} is the frequency of responses of the corresponding cell (i.e., ij) in the confusion matrix of the model data (i.e., expected).

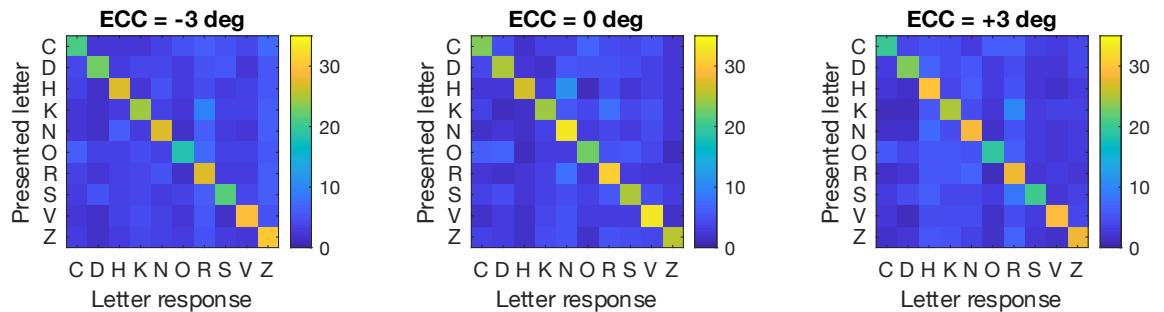
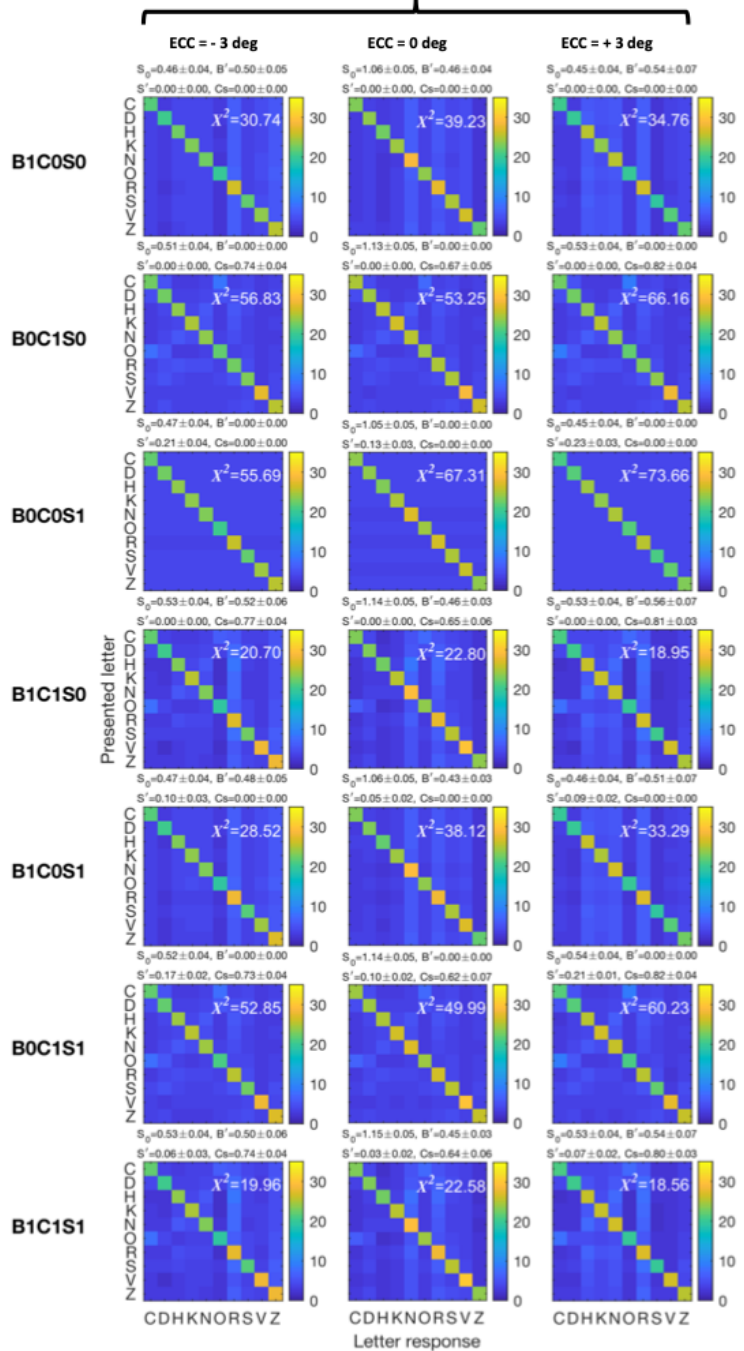


Figure 3.8 The experimental data (presented vs responses of letters).

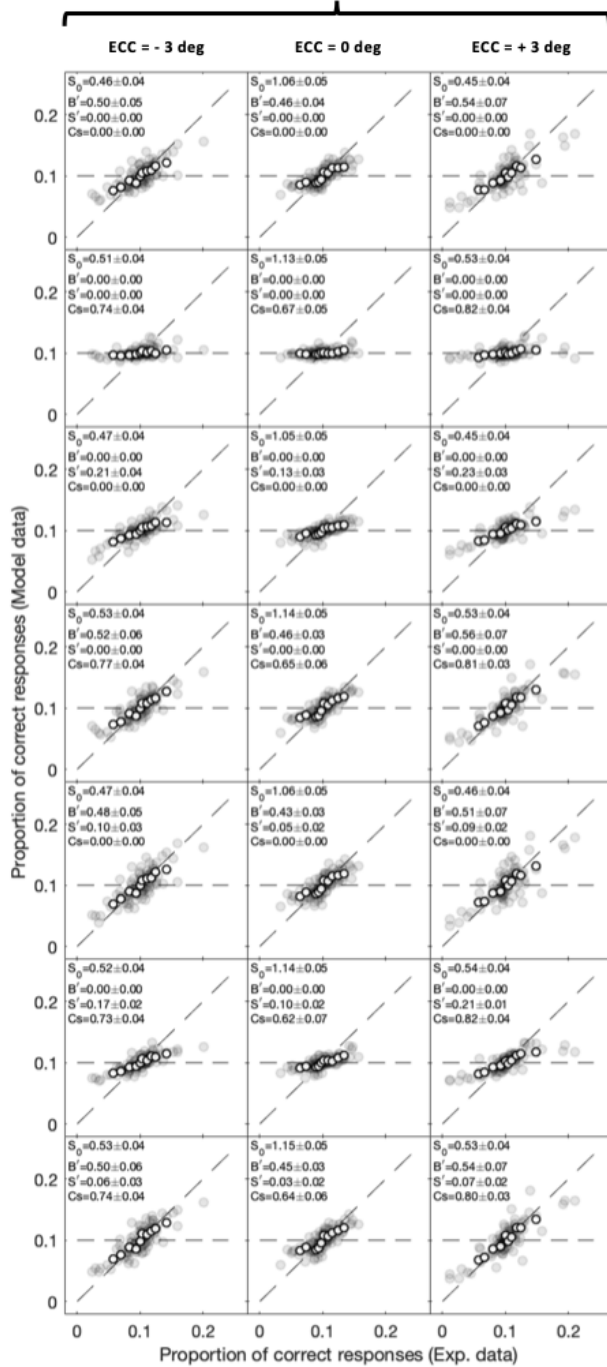
Data is collapsed across letter sizes first and then averaged across observers for each location. The colour scale illustrates the frequency of letter responses, where warmer colours show higher frequencies. The diagonal cells represent correct responses, whereas the non-diagonal cells represent incorrect responses.

Figures 3.9 B & C show the agreement between the experimental and model data in the letter usage correct and incorrect (expressed as proportion of correct and incorrect responses in the current chapter). For each location and each observer, the proportion of correct responses for a given letter is calculated as the probability of calling that letter correctly out of the total number of correct responses of all letters. Likewise, the proportion of incorrect responses for a given letter is calculated as the probability of calling that letter incorrectly out of the total number of incorrect responses of all letters. In both cases, the chance proportion became 0.1.

A



B



C

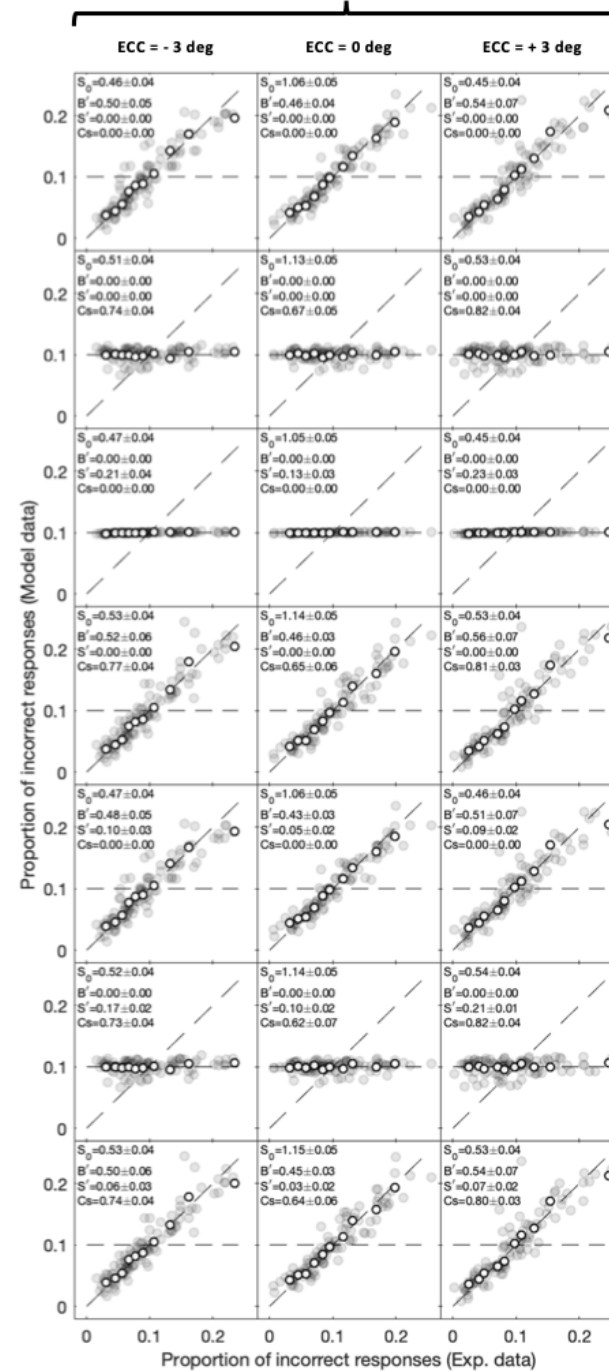


Figure 3.9 The variants of the NTM model estimated from experimental data.

I show here seven versions of the model after excluding the basic model B0C0S0. A) for each model, the confusion matrix is averaged across the six letter sizes for each observer and then averaged across observers. χ^2 value in each matrix shows the difference between a given model and the experimental data. The smaller the value, the better the model is in accounting for the experimental data. B) shows the agreement between the model and experimental data in the proportion of correct responses. The faint grey points are the observers' data for individual letters. The white data points are the average across observers. At each location and for each observer, the letters' proportions of correct responses in experimental data were rank-ordered first (from low to high proportions). Then the ranked proportions of the experimental data were averaged across observers. The model data average was calculated across observers' model data using the same ranking (i.e., based on ranking the proportion of correct responses of the experimental data). The white data points show the agreement between the two averages. The same procedure was applied to calculate the averages of the proportion of incorrect responses (C). Here and throughout (unless mentioned otherwise), the diagonal dashed line represents the perfect agreement whereas the horizontal dashed line represents the complete absence of agreement (i.e., chance proportion).

To assess the models' fitting in a way that might be more sensitive to the effects of bias, sensitivity, and similarity, I examined the confusion matrices. Figure 3.8 shows confusion matrices for the experimental data, pooled over the six letter sizes and averaged over observers. Inspection of Figure 3.9 A suggests that the fit of the bias, sensitivity, and similarity model B1C1S1 remarkably resembles the pattern of experimental data (Figure 3.8) at the three locations. Additionally, model B1C1S1 shows the smallest χ^2 values at the fovea ($\chi^2 = 22.58$), upper ($\chi^2 = 18.56$), and lower ($\chi^2 = 19.96$) visual field locations. Model B1C1S1 also shows a good account for the proportions of both correct and incorrect responses at individual and group levels at the three test locations (Figures 3.9 B & C).

Figure 3.9 also suggests that bias is the most influential factor in model B1C1S1 since it has the smallest χ^2 value at the fovea ($\chi^2 = 39.23$), upper ($\chi^2 = 34.76$), and lower ($\chi^2 = 30.74$) visual field locations when fitted individually (i.e., B1C0S0) compared to the sensitivity (i.e., B0C0S1) where value is higher at the fovea ($\chi^2 = 67.31$), upper ($\chi^2 = 73.66$) and lower ($\chi^2 = 55.69$) visual field locations and compared to similarity

(i.e., B0C1S0) where value is also higher at the fovea ($\chi^2 = 53.25$), upper ($\chi^2 = 66.16$) and lower ($\chi^2 = 56.83$) visual field locations (Figure 3.9 A). Furthermore, the bias is the key factor that accounts for the proportion correct and incorrect responses when fitted individually (i.e., B1C0S0) or in conjunction with the other factors (i.e., B1C1S0, B1C0S1, and B1C1S1) (Figure 3.9 B & C).

3.3.2 Formal model comparison

To test these observations, I conducted a formal model comparison using the Akaike Information Criterion (AIC) (Akaike, 1974) at the group level (Table 3.1). AIC weights were 0.9987 and 1 for the fuller model B1C1S1 at the upper and lower visual fields respectively. Therefore, at the periphery, the fuller model B1C1S1 is the best or most-favoured model among the eight candidates (Table 3.1). However, at the fovea, the AIC weight was found to be 0.84 for model B1C1S0 and 0.16 for the fuller model B1C1S1. Although model B1C1S1 shows a better account to experimental data than model B1C1S0 at the fovea in Figure 3.9, the AIC result suggests that adding the sensitivity factor in model B1C1S1 does not significantly improve the fit considering the price of adding it as an extra parameter compared to the lesser model (i.e., B1C1S0 with fewer parameters).

Table 3.1 Comparison of the eight models via AIC analysis, for the group of 10 observers

ECC	Model	LL	K	n	AIC	Δ AIC	Ak.Wt
-3 deg	B0C0S0	-24172.56	10	6000	48367.16	1985.31	0
	B1C0S0	-23532.44	20	6000	47107.03	725.18	0
	B0C1S0	-23851.95	20	6000	47746.06	1364.21	0
	B0C0S1	-24019.86	20	6000	48081.87	1700.02	0
	B1C1S0	-23173.47	30	6000	46409.27	27.42	0
	B1C0S1	-23476.34	30	6000	47015.02	633.16	0
	B0C1S1	-23752.18	30	6000	47566.69	1184.83	0
	B1C1S1	-23149.64	40	6000	46381.85	0	1
0 deg	B0C0S0	-23660.06	10	6000	47342.17	1486.10	0
	B1C0S0	-23155.49	20	6000	46353.14	497.07	0
	B0C1S0	-23401.94	20	6000	46846.03	989.96	0
	B0C0S1	-23578.04	20	6000	47198.24	1342.17	0
	B1C1S0	-22896.87	30	6000	45856.07	0	0.84
	B1C0S1	-23136.90	30	6000	46336.13	480.06	0
	B0C1S1	-23351.34	30	6000	46765.01	908.94	0
	B1C1S1	-22888.39	40	6000	45859.36	3.29	0.16
+3 deg	B0C0S0	-24197.24	10	6000	48416.53	2283.25	0
	B1C0S0	-23421.59	20	6000	46885.34	752.06	0
	B0C1S0	-23840.04	20	6000	47722.24	1588.96	0
	B0C0S1	-24069.58	20	6000	48181.32	2048.04	0
	B1C1S0	-23042.13	30	6000	46146.60	13.315	0.0013
	B1C0S1	-23398.74	30	6000	46859.81	726.53	0
	B0C1S1	-23722.46	30	6000	47507.25	1373.97	0
	B1C1S1	-23025.35	40	6000	46133.28	0	0.9987

Notation: LL, total log likelihood; K, total no. of free parameters across the group; n, total no. of data points; AIC is the Akaike scores; the Δ AIC is the relative difference of the scores; Ak.Wt is the Akaike weights

This finding is also confirmed by the analysis of deviance (Table 3.2) (Collett, 2003). The analysis of deviance showed that bias, sensitivity, and similarity factors have an important role in the fitting of the fuller model B1C1S1 as indicated by the statistically significant outcomes at the three test locations ($p < .001$) (Table 3.2).

Table 3.2 Group analysis of deviance, 10 observers

	ChiSq	df	p-value
Model B1C0S0 vs B1C1S1			
-3 deg	765.60	10	< .001
0 deg	534.20	10	< .001
+3 deg	792.48	10	< .001
Model B0C1S0 vs B1C1S1			
-3 deg	1404.63	10	< .001
0 deg	1027.09	10	< .001
+3 deg	1629.38	10	< .001
Model B0C0S1 vs B1C1S1			
-3 deg	1740.44	10	< .001
0 deg	1379.30	10	< .001
+3 deg	2088.46	10	< .001

Further model comparisons after excluding the fuller model (B1C1S1) reveal that the bias and similarity model (B1C1S0) is favoured over the other six models (Table 3.3). Additionally, comparing only the one-factor models reveals that the bias model (B1C0S0) is favoured over the other two models (B0C1S0 & B0C0S1) (Table 3.4). Comparing the similarity model to the sensitivity model after excluding the bias model shows that the similarity model is favoured over the sensitivity model (Table 3.5). These interesting findings suggest that bias has a higher contribution to the fuller model than the other two factors (sensitivity and similarity) and the joint effect of bias and similarity is the most influential in the fuller model than the joint effect of either bias and sensitivity or similarity and sensitivity. The results also suggest that the

sensitivity factor has the least effect on the fuller model especially at the fovea.

Table 3.3. Comparison of the seven models via AIC analysis, for the group of 10 observers

ECC	Model	LL	K	n	AIC	Δ AIC	Ak.Wt
-3 deg	B0C0S0	-24172.56	10	6000	48365.16	1957.91	0
	B1C0S0	-23532.44	20	6000	47105.10	697.77	0
	B0C1S0	-23851.95	20	6000	47744.05	1336.80	0
	B0C0S1	-24019.86	20	6000	48079.86	1672.61	0
	B1C1S0	-23173.47	30	6000	46407.25	0	1
	B1C0S1	-23476.34	30	6000	47013.00	605.75	0
	B0C1S1	-23752.18	30	6000	47564.67	1157.42	0
0 deg	B0C0S0	-23660.06	10	6000	47340.16	1486.11	0
	B1C0S0	-23155.49	20	6000	46351.13	497.08	0
	B0C1S0	-23401.94	20	6000	46844.02	989.96	0
	B0C0S1	-23578.04	20	6000	47196.23	1342.18	0
	B1C1S0	-22896.87	30	6000	45854.05	0	1
	B1C0S1	-23136.90	30	6000	46334.11	480.06	0
	B0C1S1	-23351.34	30	6000	46762.99	908.94	0
+3 deg	B0C0S0	-24197.24	10	6000	48414.52	2269.95	0
	B1C0S0	-23421.59	20	6000	46883.33	738.75	0
	B0C1S0	-23840.04	20	6000	47720.23	1575.65	0
	B0C0S1	-24069.58	20	6000	48179.31	2034.73	0
	B1C1S0	-23042.13	30	6000	46144.58	0	1
	B1C0S1	-23398.74	30	6000	46857.79	713.22	0
	B0C1S1	-23722.46	30	6000	47505.23	1360.66	0

Notation: LL, total log likelihood; K, total no. of free parameters across the group; n, total no. of data points; AIC is the Akaike scores; the Δ AIC is the relative difference of the scores; Ak.Wt is the Akaike weights

Table 3.4 Comparison of the four models via AIC analysis, for the group of 10 observers

ECC	Model	LL	K	n	AIC	Δ AIC	Ak.Wt
-3 deg	B0C0S0	-24172.56	10	6000	48365.16	1260.14	0
	B1C0S0	-23532.44	20	6000	47105.02	0	1
	B0C1S0	-23851.95	20	6000	47744.05	639.03	0
	B0C0S1	-24019.86	20	6000	48079.86	974.84	0
0 deg	B0C0S0	-23660.06	10	6000	47340.16	989.04	0
	B1C0S0	-23155.49	20	6000	46351.13	0	1
	B0C1S0	-23401.94	20	6000	46844.02	492.88	0
	B0C0S1	-23578.04	20	6000	47196.23	845.10	0
+3 deg	B0C0S0	-24197.24	10	6000	48414.52	1531.20	0
	B1C0S0	-23421.59	20	6000	46883.33	0	1
	B0C1S0	-23840.04	20	6000	47720.23	836.90	0
	B0C0S1	-24069.58	20	6000	48179.31	1295.98	0

Notation: LL, total log likelihood; K, total no. of free parameters across the group; n, total no. of data points; AIC is the Akaike scores; the Δ AIC is the relative difference of the scores; Ak.Wt is the Akaike weights

Table 3.5 Comparison of the three models via AIC analysis, for the group of 10 observers

ECC	Model	LL	K	n	AIC	Δ AIC	Ak.Wt
-3 deg	B0C0S0	-24172.56	10	6000	48365.16	621.11	0
	B0C1S0	-23851.95	20	6000	47744.05	0	1
	B0C0S1	-24019.86	20	6000	48079.87	335.81	0
0 deg	B0C0S0	-23660.06	10	6000	47340.16	496.15	0
	B0C1S0	-23401.94	20	6000	46844.02	0	1
	B0C0S1	-23578.04	20	6000	47196.23	352.21	0
+3 deg	B0C0S0	-24197.24	10	6000	48414.52	694.29	0
	B0C1S0	-23840.04	20	6000	47720.23	0	1
	B0C0S1	-24069.58	20	6000	48179.31	459.08	0

Notation: LL, total log likelihood; K, total no. of free parameters across the group; n, total no. of data points; AIC is the Akaike scores; the Δ AIC is the relative difference of the scores; Ak.Wt is the Akaike weights

3.3.3 The relative role of bias, sensitivity, and similarity in the letter identification task

The "best" model (i.e., B1C1S1) does not necessarily mean that the model is good at explaining the different aspects of the letter identification task in the current experiment. I showed that the best model accounts very well for the differences in the proportion of correct and incorrect responses (Figure 3.9 B & C). But it is crucial to investigate how well the best model captures the bias, sensitivity, and similarity separately in the experimental data, which I will describe in the following section. Figure 3.10 shows the best model confusion matrices and the agreement with letters proportion correct of experimental data in case of the model recomputed with three factors activated, sensitivity and similarity activated (i.e., B off), bias and sensitivity activated (i.e., C off), bias and similarity activated (i.e., S off), only sensitivity activated (i.e., B & C off), only similarity activated (i.e., B & S off) or only bias activated (i.e., C & S off).

Results show that the bias is the key factor in explaining the overall response pattern in the experimental data (i.e., has smaller χ^2 at the three test locations, Figure 3.10 A) and in explaining the differences in letter proportions of correct responses either individually (i.e., C & S off) or in conjunction with the other factors (i.e., B1C1S1, C off and S off) (Figure 3.10 B). On the other hand, one would assume that the sensitivity differences between letters would account for the differences in proportions of correct responses in the experimental data, but that was not the case when I examined the model recomputed for sensitivity individually (i.e., B & C off). The model with only the similarity factor activated (i.e., B & S off) also showed poor accounting for differences in proportions of correct responses in experimental data. Furthermore,

even the model with the joint effect of sensitivity differences and similarity between letters only (i.e., B off) does not account for the overall responses in experimental data (Figure 3.10 A) or for the differences in proportion corrects (Figure 3.10 B) as the model with the bias individually does (i.e., C & S off).

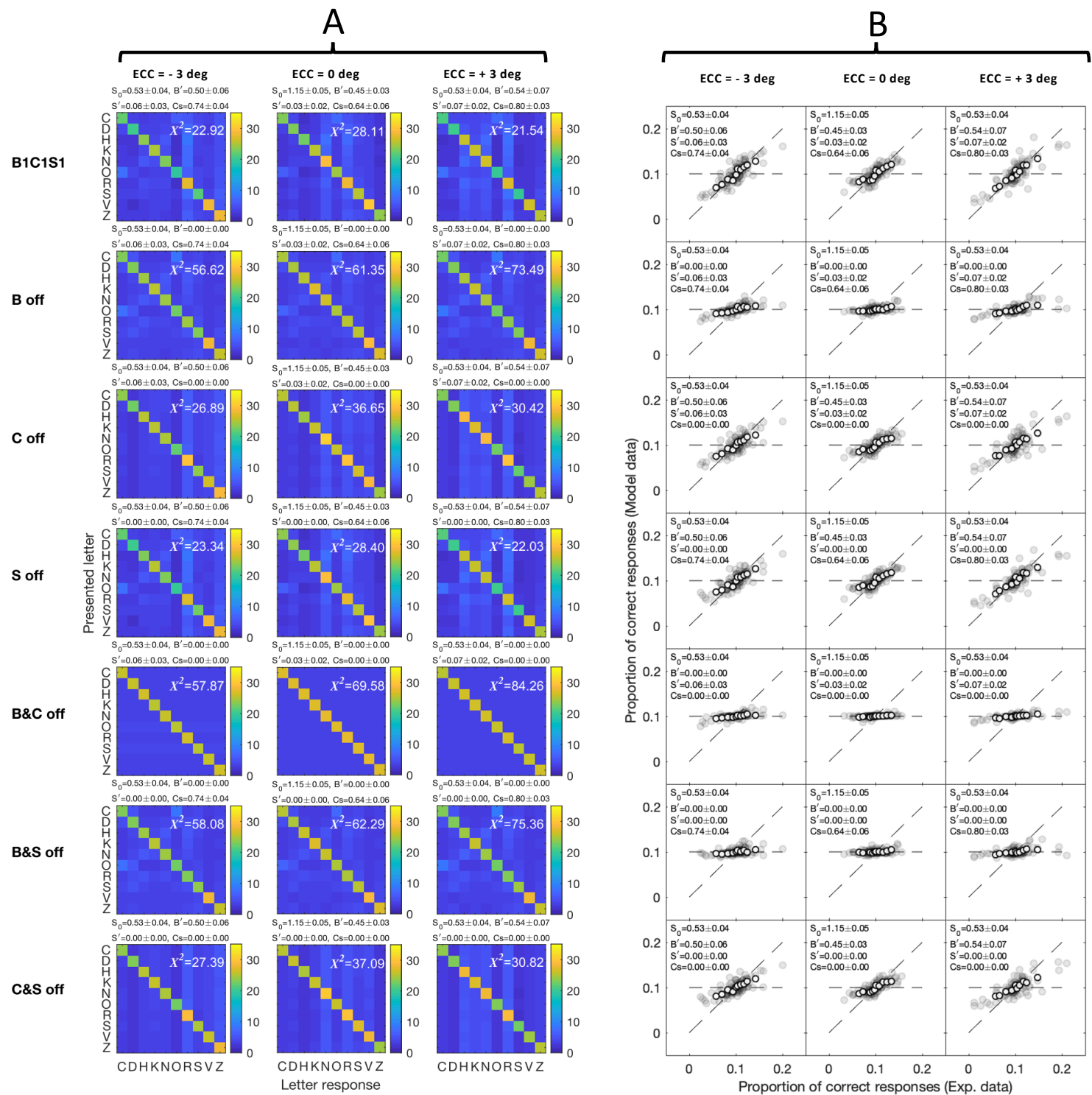


Figure 3.10 The best model recomputed with different activated factors.

(A) the confusion matrices of the best model (B1C1S1) recomputed with different activated factors (see text) and (B) The agreement between experimental and model data in the differences of letter proportions of correct responses at the three test locations. Bias is the key factor in defining the overall responses and the differences in letter proportions of correct responses in experimental data.

Results show that the best model (B1C1S1) accounts for the differences in proportion of *incorrect* responses (Figure 3.9 C). However, here I investigate more specifically how the best model accounts for the sources of the differences in letter *incorrect* responses which can be either bias, similarity between letters, or both. I employed Luce's choice model (Luce, 1963) to compute the bias and similarity from the best model B1C1S1 and the experimental data, and then I examined the agreement between the results of the best model and experimental data (Figure 3.11). The fitting algorithm of Luce's choice model was shown in Chapter 2 ("Methods" section). The fitting of Luce's choice model showed that model B1C1S1 described the bias and similarity in the experimental data at the three test locations very well.

Figure 3.11 A shows that Luce's bias in the best model in the cases with activated bias factor is strikingly similar to Luce's bias in the experimental data at the three test locations. This also indicates that our model estimates the same kind of bias that Luce's model does, and it supports our strong assumption about the *linearity* of the bias and its ranking based on the letter usage (total). To further illustrate this, Figure 3.12 A shows the agreement between the bias parameters estimated by the best model and the corresponding bias parameters estimated by Luce's choice model when both models (B1C1S1 and Luce's choice model) fitted to experimental data.

Figure 3.11 B shows Luce's similarity parameters of the 45 confusion pairs. There are only 45 confusion pairs because Luce's model assumes that the similarity between letters is symmetrical (i.e., the parameter of similarity between the presented letter C and responded letter O is identical to the similarity parameter between the presented letter O and responded letter C). The results suggest that Luce's similarity between letters estimated from the best model data in the cases of similarity factor is activated is similar to Luce's similarity between letters estimated from experimental data. It is also evident that the agreement is not as strong as in the case of Luce's bias comparison (Figure 3.11 A). This is not unexpected since I used the OC matrix to capture the similarity between letters that follows the pattern of the OC matrix. Hence, Luce's model will capture a particular pattern of similarity when fitted to the best model data (i.e., the pattern of OC matrix) whereas Luce's model captures different and a variety of similarity patterns for each observer when fitted to experimental data. Nevertheless, the reasonable agreement between Luce's similarity of the best model and experimental data suggests that the general and common similarity pattern in experimental data is the one that follows the OC matrix (Figure 3.12 B).

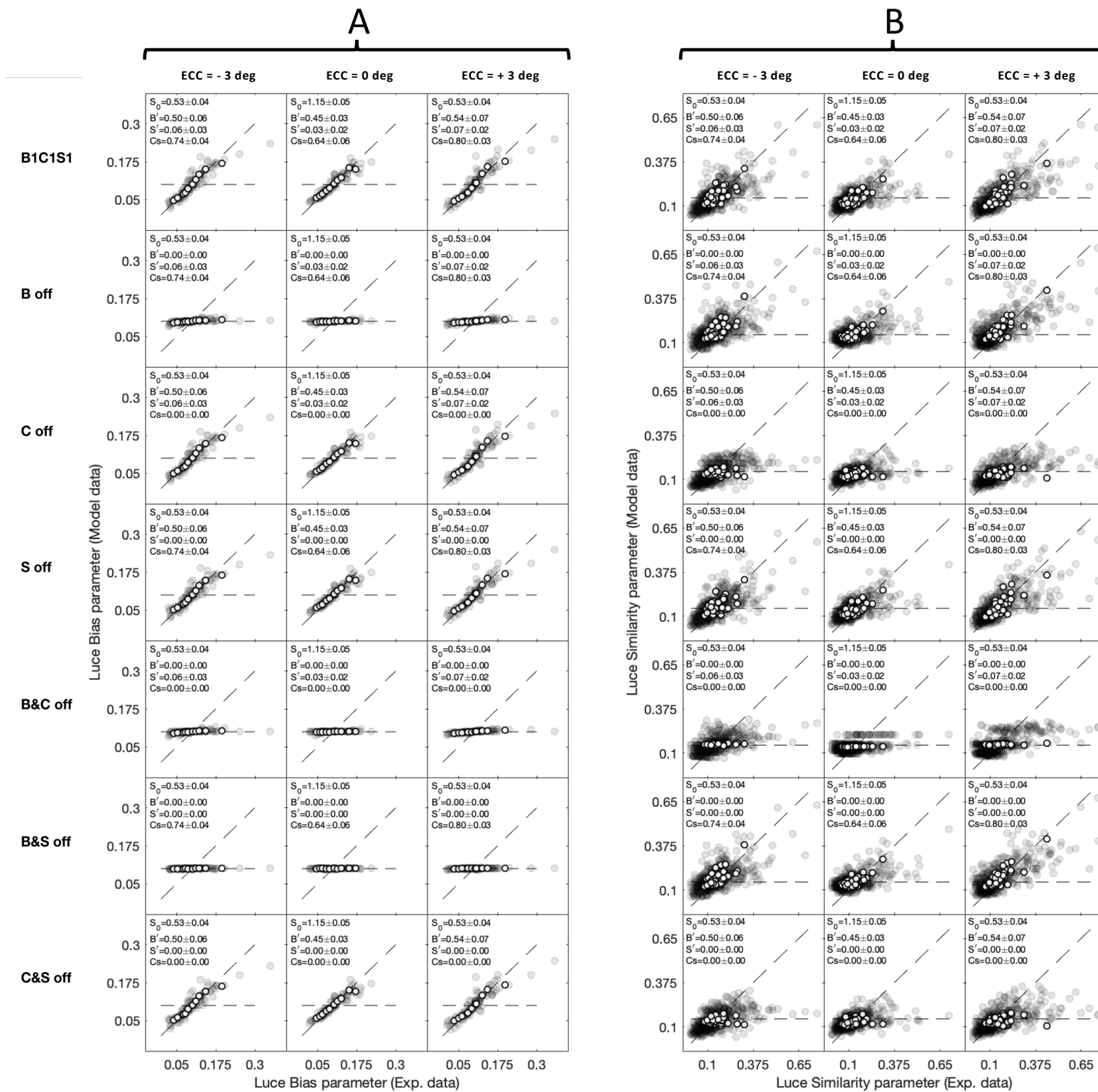


Figure 3.11 Luce's parameters of the best model vs experimental data.

A) the agreement between Luce's bias parameters estimated from the best model and experimental data. B) shows the agreement between Luce's similarity parameters of the model and experimental data. The horizontal dashed line in (A) is the no-bias case where all letters show Luce's bias parameter of 0.1. To calculate the average, Luce bias parameters estimated from experimental data were rank-ordered first for each observer (from low to high bias). Then the ranked biases were averaged across observers. The averages of Luce bias parameters estimated from model data were calculated across the ranked biases of observers using the same ranking (i.e., based on ranking the Luce bias parameters of the experimental data). The white data points show the agreement between the two averages. The horizontal dashed line in (B) is the no-similarity case where all pairs show Luce's similarity parameter of 0.15. It is calculated by assuming that for a given test location the proportion of incorrect responses is 50% out of the total frequency of responded letters (i.e., 600). Therefore, the frequency of incorrect responses is 300 and each confusion pair in this case (and in case of symmetry similarity assumption (see text)) would have a frequency of $300/45 = 6.67$ responses which corresponds to $6.67/45 = 0.15$ chance ratio. The grey data points here are for the 45 confusion pairs for each observer. The white data points are the averages of the parameters across observers.

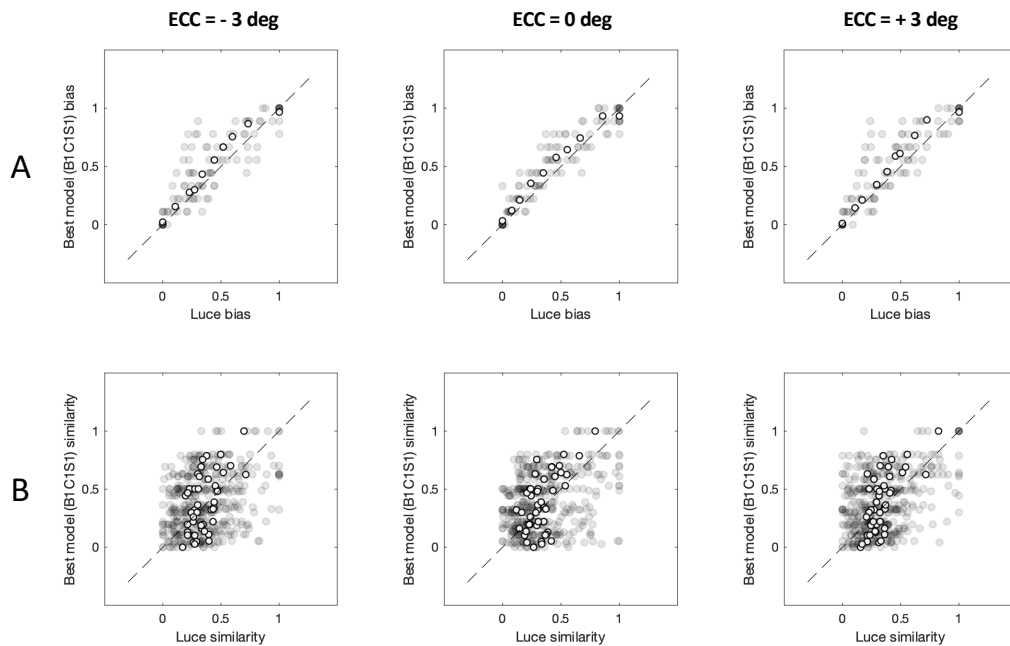


Figure 3.12 Bias and similarity of Luce's model vs the best model.

A) the agreement between Luce's bias and the best model (B1C1S1) bias parameters at the individual (grey data points) and group level (white data points). B) the agreement between Luce's similarity and the best model (B1C1S1) similarity parameters at the individual and group levels. The bias parameters of the Luce model were rank-ordered first for each observer, averaged across observers, and then compared to the corresponding average of the bias parameters of the best model (white data points in A). Same procedure was performed for Luce similarity parameters. i.e., The average was performed for the corresponding confusion pairs (45 pairs) across observers (white data points in B). All data in this figure were normalized between 0 and 1.

3.3.3.1 The effect of bias, sensitivity, and similarity on letter acuity measurement

Model B1C1S1 gave an excellent account of the differences in the proportion of correct responses between letters. This enabled me to examine whether estimating letter acuity from the PMFs of B1C1S1 (that takes account of bias, sensitivity, and similarity) differs from the other variants of the model that use only one, two, or none of these factors (i.e., B1C0S0, B0C1S0, B0C0S1, B1C1S0, B1C0S1, B0C1S1, B0C0S0). For each model, letter acuity for each observer was calculated as follows: the PMFs of the model were averaged across letter identities and letter acuity (in minutes of arc (')) was then estimated as the letter size that gave 55% correct performance (halfway between guessing (10%) and perfect (100%) performance). Table 3.6 shows the average letter acuities (across observers) at the three test locations. These results indicate that models considering any of the three factors individually, two factors, or the three factors, did not significantly alter the estimated letter acuity when compared with the default model (B0C0S0) that had no bias, sensitivity, or similarity factor.

Table 3.6 Letter acuity (letter stroke width in min of arc (')) calculated from the eight fitted models and averaged across observers \pm SE.

Models	ECC = -3 deg	ECC = 0 deg	ECC = 3 deg
B0C0S0	2.98' \pm 0.22	1.19' \pm 0.06	3.04' \pm 0.26
B1C0S0	3.05' \pm 0.25	1.19' \pm 0.06	3.10' \pm 0.27
B0C1S0	2.79' \pm 0.22	1.15' \pm 0.05	2.78' \pm 0.24
B0C0S1	2.87' \pm 0.23	1.17' \pm 0.06	2.96' \pm 0.26
B1C1S0	2.78' \pm 0.21	1.15' \pm 0.05	2.82' \pm 0.26
B1C0S1	2.93' \pm 0.23	1.19' \pm 0.06	3.05' \pm 0.28
B0C1S1	2.73' \pm 0.21	1.13' \pm 0.06	2.73' \pm 0.24
B1C1S1	2.79' \pm 0.21	1.15' \pm 0.05	2.83' \pm 0.26

3.3.3.2 The effect of bias, sensitivity, and similarity on the *spread* of the underlying PMF

I also examined the spread of the PMFs of the eight models. The spread σ is defined as

$$\sigma = \psi^{-1}(1 - \delta, p, S_0, B', Cs, S') - \psi^{-1}(\gamma + \delta, p, S_0, B', Cs, S'), \quad (\text{Eq. 3.6})$$

where ψ^{-1} is the inverse of the PMF (i.e., the letter size that gives a certain proportion correct), δ is an arbitrary number (e.g., 0.01) that sets the start and the end points of σ , and γ is the guessing rate (0.1 in our experiment) (Prins, 2016). According to this, σ is the letter size range (letter stroke width in ') within which the proportion correct goes from 0.11 to 0.99. Calculating σ reveals that the models with the sensitivity and similarity factors individually had substantially higher σ (i.e., shallower slope) especially at the periphery (Table 3.7). Moreover, when fitted with bias (i.e., B1C0S1), the sensitivity effect became negligible on the σ . However, this is not the case with similarity where the wider σ persists even when fitted with other factors (i.e., B1C1S0 and B0C1S1). Therefore, it is plausible to conclude that the wider σ in the best model B1C1S1 is mainly due to similarity.

Table 3.7 Spread of the model psychometric functions (σ) averaged across observers (\pm SE), for each of the eight models.

Models	ECC = -3 deg	ECC = 0 deg	ECC = 3 deg
B0C0S0	4.23' \pm 0.32	1.45' \pm 0.07	4.34' \pm 0.38
B1C0S0	4.30' \pm 0.35	1.47' \pm 0.07	4.40' \pm 0.40
B0C1S0	5.25' \pm 0.59	1.66' \pm 0.11	5.67' \pm 0.70
B0C0S1	5.00' \pm 0.56	1.58' \pm 0.10	5.12' \pm 0.56
B1C1S0	5.28' \pm 0.54	1.63' \pm 0.12	5.62' \pm 0.67
B1C0S1	4.54' \pm 0.43	1.51' \pm 0.08	4.66' \pm 0.47
B0C1S1	5.59' \pm 0.61	1.68' \pm 0.10	5.96' \pm 0.67
B1C1S1	5.44' \pm 0.61	1.65' \pm 0.10	5.69' \pm 0.71

The greater σ for models with similarity (i.e., B0C1S0, B1C1S0, B0C1S1 and B1C1S1) was mainly found in the upper half of the PMFs (upper spread: from 0.56 to 0.99 proportion correct, Table 3.8). Negligible differences were found in the lower half of the PMFs (lower spread: from 0.11 to 0.54 proportion correct, Table 3.9). Hence, one might expect the presence of similarity between letters to lead to higher test-retest variability (i.e., less precise measurements) in letter acuity estimation, when using an adaptive method (letter by letter presentation). Test-retest variability is an important factor to be considered and minimised when the goal is efficient monitoring of change in a visual acuity over time.

Table 3.8 Average of the upper spread of model psychometric functions (from 0.56 to 0.99 proportion correct) across observers (\pm SE).

Models	ECC = -3 deg	ECC = 0 deg	ECC = 3 deg
B0C0S0	1.80' \pm 0.14	0.56' \pm 0.03	1.88' \pm 0.16
B1C0S0	1.83' \pm 0.15	0.57' \pm 0.03	1.89' \pm 0.17
B0C1S0	2.94' \pm 0.42	0.80' \pm 0.08	3.37' \pm 0.51
B0C0S1	2.65' \pm 0.40	0.69' \pm 0.05	2.69' \pm 0.36
B1C1S0	2.97' \pm 0.37	0.78' \pm 0.07	3.27' \pm 0.47
B1C0S1	2.14' \pm 0.24	0.61' \pm 0.04	2.16' \pm 0.24
B0C1S1	3.34' \pm 0.45	0.83' \pm 0.07	3.70' \pm 0.47
B1C1S1	3.10' \pm 0.43	0.80' \pm 0.07	3.34' \pm 0.50

Table 3.9 Average of the lower spread of model psychometric functions (from 0.11 to 0.54 proportion correct) across observers (\pm SE)

Models	ECC = -3 deg	ECC = 0 deg	ECC = 3 deg
B0C0S0	2.36' \pm 0.18	0.88' \pm 0.05	2.44' \pm 0.21
B1C0S0	2.41' \pm 0.19	0.88' \pm 0.04	2.47' \pm 0.22
B0C1S0	2.26' \pm 0.18	0.84' \pm 0.05	2.25' \pm 0.20
B0C0S1	2.29' \pm 0.18	0.86' \pm 0.05	2.38' \pm 0.21
B1C1S0	2.25' \pm 0.17	0.84' \pm 0.04	2.28' \pm 0.20
B1C0S1	2.33' \pm 0.19	0.88' \pm 0.04	2.44' \pm 0.23
B0C1S1	2.18' \pm 0.17	0.82' \pm 0.04	2.19' \pm 0.20
B1C1S1	2.26' \pm 0.18	0.83' \pm 0.04	2.27' \pm 0.21

3.3.3.3 Simulations to examine the effect of bias, sensitivity, and similarity on letter acuity measurements and variabilities

In the following, I used the NTM and conducted a Monte Carlo simulation for an adaptive method to examine the effect of bias, sensitivity, or similarity on the estimation of the letter acuity and test-retest variability of the measurements. Employing the Fixed Step Size (FSS) staircase, the Monte Carlo simulation was used to generate responses for randomly selected letter-by-letter presentations with a modified Up / Down staircase paradigm to estimate the letter acuity with the following parameters. The starting letter size was set to the expected letter acuity from Table 3.6 (i.e., $\sim 1.2'$ for the fovea, and $\sim 3'$ for the periphery). A one Up / one Down staircase with a step-up size of $0.5'$ and step-down size of $0.409'$ was chosen to target 55% correct responses according to the following equation (Garcia-Pérez, 1998):

$$\psi = \left(\frac{\Delta^+}{\Delta^+ + \Delta^-} \right)^{\frac{1}{D}} \quad (\text{Eq. 3.7})$$

where ψ is the target percent correct, Δ^+ is the step-up size ($0.5'$), Δ^- is the step-down

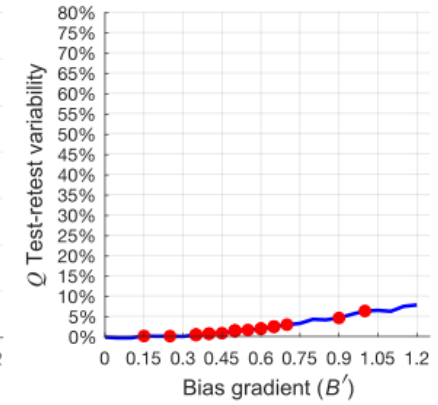
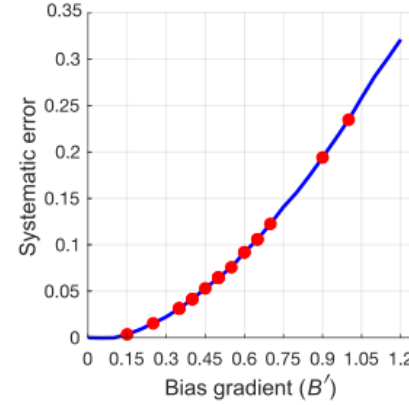
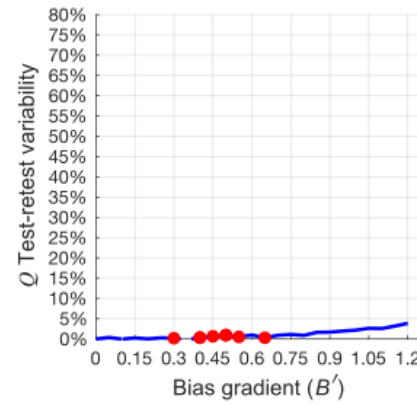
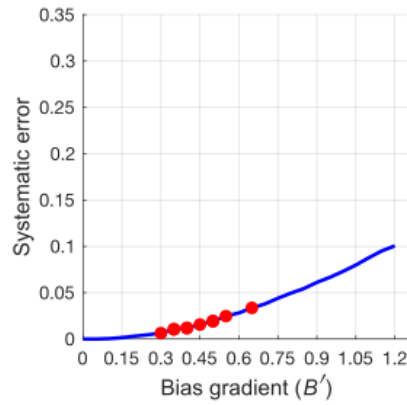
size (0.409'), and D is the number of consecutive correct responses after which a step-down is to be made (1 in the current simulation). Ideally, and especially if I need to run fewer trials, I should choose a larger step-up size (between $\sigma/2$ and σ , e.g., 1' for the fovea). However, to obtain more precise results, I used a smaller step-up of 0.5' and compensated for that by increasing the number of trials. I set each run to stop after 50 trials, and letter acuity was calculated as the average over the last five reversals.

I simulated the effect for a wide range of biases, sensitivities, and similarities (Figure 3.12). The letter acuity was estimated 100,000 times for each parameter, and then the systematic error of the estimated mean of the letter acuity and the change of the test-retest variability was calculated over these 100,000 measurements. I also showed the potential systematic error and the change in test-retest variability that each observer might have according to their estimated parameters from the best model fitting (red dots in Figure 3.13) at the fovea and the periphery. The results show that the systematic error (although negligible) and the change in test-retest variability that result from the similarity between letters (Figure 3.13 B) would be higher than those resulting from bias or sensitivity. For example, the observers that show the highest degree of similarity between letters at the periphery (i.e., $C_s = 1$), can cause a systematic error in estimated letter acuity of $\sim 0.3'$ and a change in the test-retest variability by $\sim 25\%$ (Figure 3.13 B). The effect is less strong at the fovea for the highest similarity (systematic error $\sim 0.1'$ and test-retest variability $\sim 15\%$). However, the observers that show the highest degree of bias can cause a systematic error of the estimated letter acuity by $\sim 0.25'$ and a change in the test-retest variability by only $\sim 7\%$ (Figure 3.13 A). Similarly, observers that show the highest degree of sensitivity differences between letters cause a systematic error of the estimated letter acuity by $\sim 0.02'$ and a change in the test-retest variability by only $\sim 10\%$ (Figure 3.13 C).

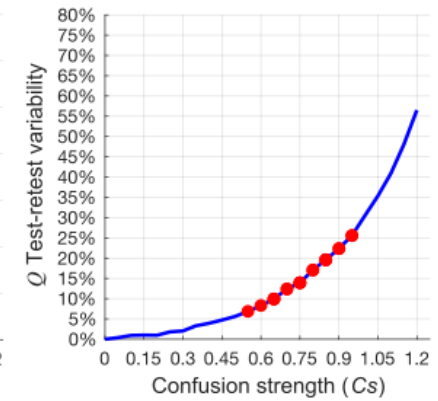
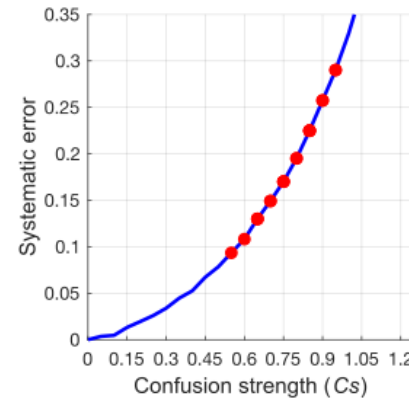
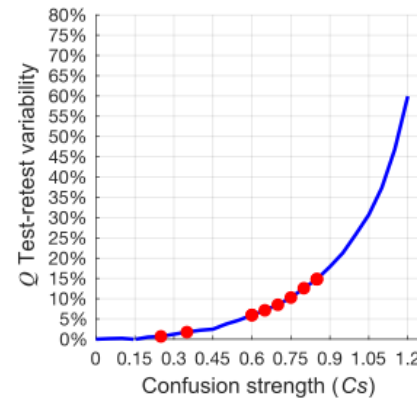
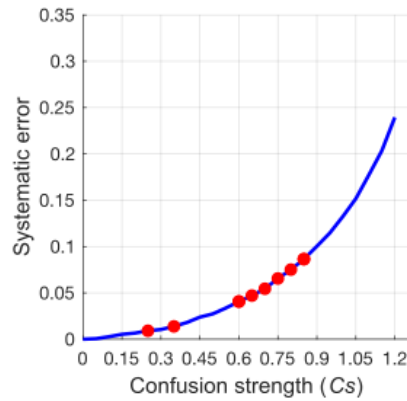
ECC = 0 deg

ECC = ±3 deg

A



B



C

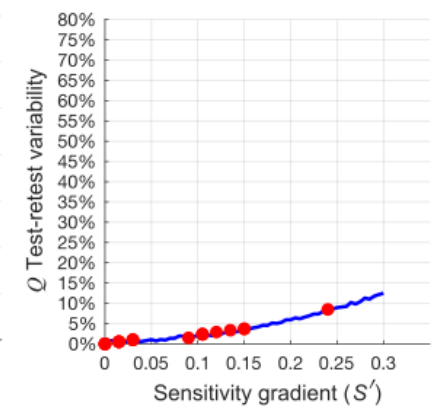
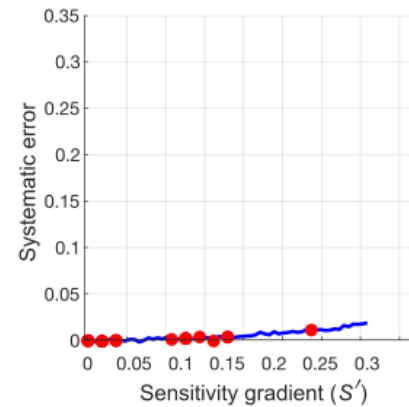
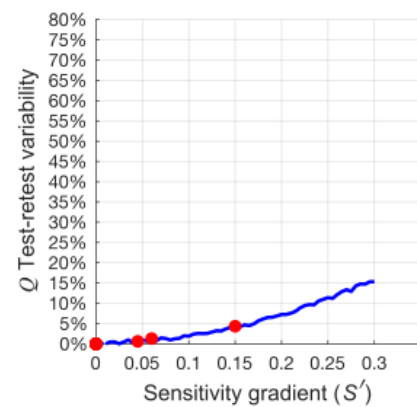
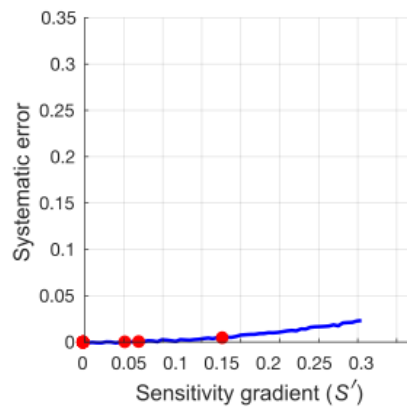


Figure 3.13 The effect of bias, sensitivity, and similarity on letter acuity.

For 100,000 simulated measured letter acuities, the systematic error and the change (Q in the plots) in test-retest variability were calculated across a wide range of bias (A), similarity (B) and sensitivity (C). The red dots are the observers' estimated parameters from best model (B1C1S1) fitting to experimental data.

3.4 Discussion

3.4.1 Best model

Our previous NTM (Georgeson et al., 2023) model predicted variations in visual performance across letters, mainly (but not exclusively) through the patterns of bias that varied between letters and between observers. Here I continue this work by incorporating *similarity* between letters into the NTM. I present an extended version of the NTM to reveal the joint role of bias, sensitivity differences, and similarity between letters *simultaneously* (instead of bias and sensitivity only). I fit eight variants of the NTM that include bias, sensitivity and/or similarity, or none. The results clearly show that the model with three factors (bias, sensitivity, and similarity, i.e., B1C1S1) is the best of the eight models at the peripheral test locations and gives a very good account of the pattern of responses (correct and incorrect) in the experimental data (Figure 3.9, Table 3.1). At central vision, although model B1C1S1 improved the fitting to experimental data (Figure 3.9), AIC analysis shows that the improvement is not significant compared to model B1C1S0 (Table 3.1). Further analysis shows that the joint role of bias and similarity is significantly higher than that of either bias and sensitivity or similarity and sensitivity (Figure 3.9, 3.10 & 3.11, Table 3.3). Results also show that bias is the key factor that accounts for the proportion of correct and incorrect responses in the experimental data (Figure 3.9, 3.10 & 3.11, Table 3.4). The fact that bias is found to be the key factor in the current and our previous work (Georgeson et al., 2023) explains the success of our previous model (bias and

sensitivity only). However, the current work demonstrates that the joint effect of bias and similarity is superior to the joint effect of bias and sensitivity (Table 3.3) at the three test locations.

We propose that the (initially unknown) ordering of letter templates, from least- to most biased, is reflected in the letter usage (total) with which observers choose different letters as a response (whether correctly or not). The biased ordering of an observer's letter template is determined by the rank-ordering of the letter usage (total). The rank order is unique for each observer and for each test location. If we further assume that the bias values are equally spaced across this ordering, we establish the concept of a *bias gradient* across letter templates. This greatly simplifies the modelling because each set of 10 different letter biases is reduced to a single number, the bias gradient (B') that varies between observers and test locations. A similar concept is applied to the sensitivity gradient (S') (Georgeson et al., 2023).

3.4.2 Validating the NTM using Luce's choice model

Luce's choice model (Luce, 1963) has been used by numerous previous studies and it performs very well in capturing response bias and similarity in letter identification tasks (Nosofsky, 1991; Smith, 1992; Mueller & Weidemann, 2012; Coates, 2015; Hamm et al., 2018). Here I employed Luce's choice model to further validate our model. Firstly, I investigated whether the best-performing model (B1C1S1) is efficient in capturing bias and similarity by calculating and comparing bias and similarity parameters from model B1C1S1 and the experimental data using Luce's choice model. Results show that model B1C1S1 was remarkably efficient (especially for the bias) in capturing the bias and similarity as shown by the agreement in Luce's bias and

similarity parameters estimated from model B1C1S1 and experimental data at the three test locations (Figure 3.11). Secondly, I compared the bias and similarity parameters of model B1C1S1 to the bias and similarity parameters of Luce's choice model when both are fitted to experimental data.

3.4.2.1 Response bias

I find that the bias parameters estimated by Luce's choice model are in excellent agreement with the bias parameters estimated by model B1C1S1 at individual and group levels (Figure 3.12 A). This suggests that both models capture the same type of bias that is based on the letter usage (total). Furthermore, since the bias computation in Luce's model does not assume the pattern of the bias across letters to be linear, the excellent agreement seen in Figure 3.12 A supports the assumptions of the bias modelling (i.e., a *linear bias gradient* across letter templates).

3.4.2.2 Similarity between letters

Nevertheless, I did not find a similarly compelling agreement when comparing the similarity parameters between Luce's choice model and model B1C1S1 (Figure 3.12 B). The reason might be rooted in the different approaches to defining and calculating the similarity of Luce's choice model and the NTM. In the NTM, I assume that the pattern of similarity between letters for any observer must follow the OC matrix pattern. However, Luce's choice model computes the similarity between letters by referring to the maximum likelihood estimates of the responses without imposing any restriction on the pattern of the resultant similarity. Hence, it is expected that Luce's choice model produces a unique similarity pattern for different observers. This might

explain the weaker agreement between the two models concerning the similarity parameters (Figure 3.12 B).

3.4.3 Source of error in similarity estimation using Luce's choice model

Given that Luce's choice model uses the maximum likelihood estimates of the experimental data (i.e., confusion matrix) to compute both bias and similarity, I examined the hypothesis that it is possible for the similarity calculation of Luce's choice model to be contaminated by the existence of substantial biases among letters, resulting in various similarity patterns. Figure 3.14 shows the agreement between the parameters of the *best* model and the estimated parameters using Luce's choice model of *simulated* data using the *best* model where all factors are activated (Figure 3.14 A), bias off (Figure 3.14 B) and similarity off (Figure 3.14 C). Results show that the *similarity* parameters' agreement has improved substantially when calculated using Luce's choice model while the bias is off (Figure 3.14 B) compared to the case where the bias is on (Figure 3.14 A). This suggests that the calculation of the similarity using Luce's choice model is substantially influenced by the bias in the data. However, that is not the case for the *bias* calculated by Luce's choice model with similarity activated (Figure 3.14 A) or similarity off (Figure 3.14 C), where in both cases the agreement is excellent between the parameters of the *best* model and the estimated parameters using Luce's choice model. This suggests that the bias calculation via Luce's choice model is robust against the existence of similarity between letters.

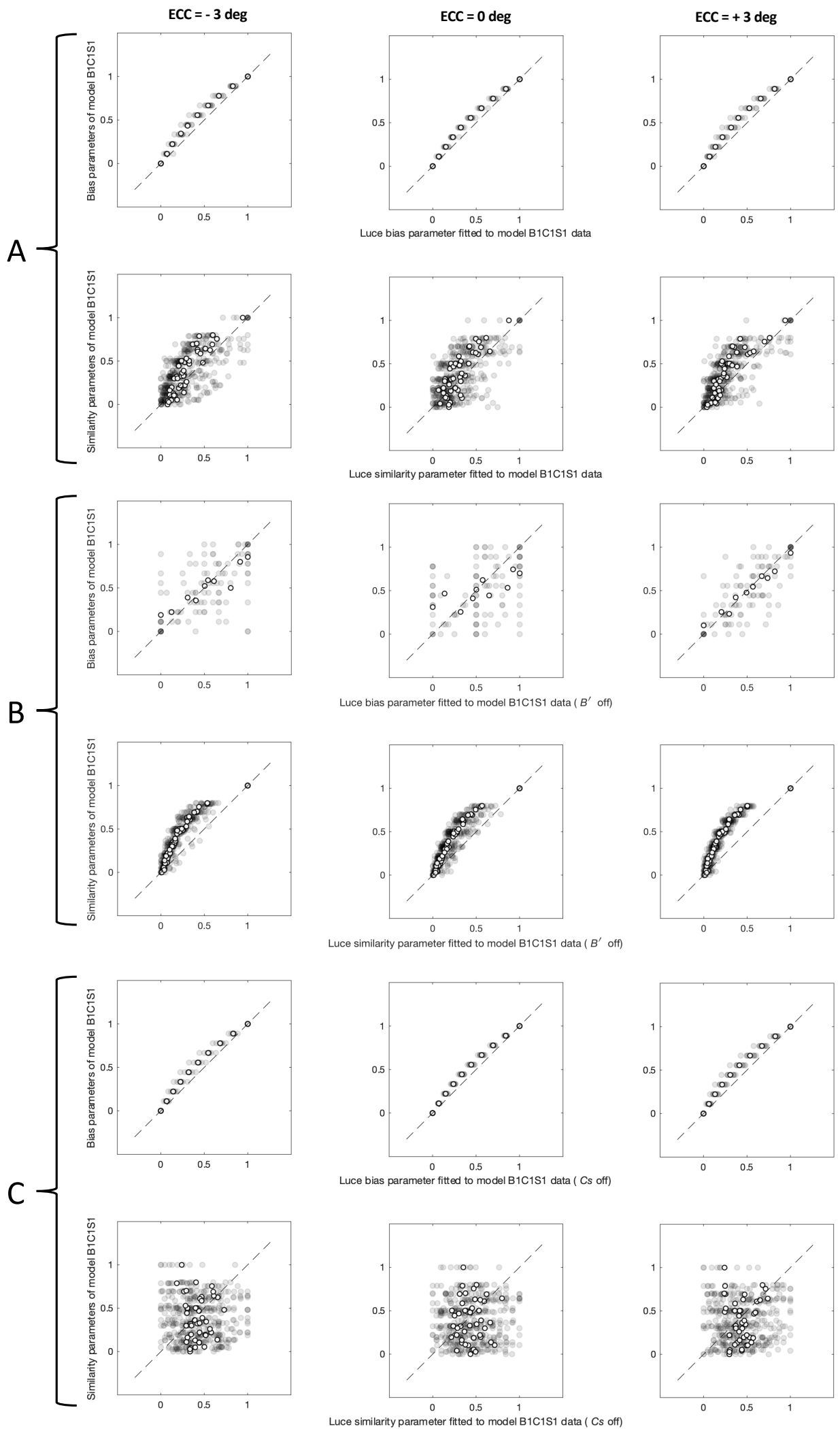


Figure 3.14 The influence of bias on Luce's similarity calculation.

Figure shows the agreement between the parameters of the best model and the estimated parameters using Luce's choice model of simulated data using the best model in case of all factors active (A), Bias off (B), and similarity off (C).

To further investigate these interesting results, I examined the effect of different conditions on the estimated Luce's bias and similarity parameters. I can show that bias and sensitivity (to a smaller degree) factors can substantially influence the similarity calculations via Luce's choice model at the fovea (Figure 3.15 A & B). The results also show that the bias calculation via Luce's choice model is robust against the existence of even strong similarity and sensitivity in the data (Figure 3.15 C & D). Similar results were found at the periphery (Figure 3.16). In this case, the *apparent unique* similarity pattern found for each observer calculated by Luce's model might be regarded as just an artifact of the effect of calculating biases as a similarity.

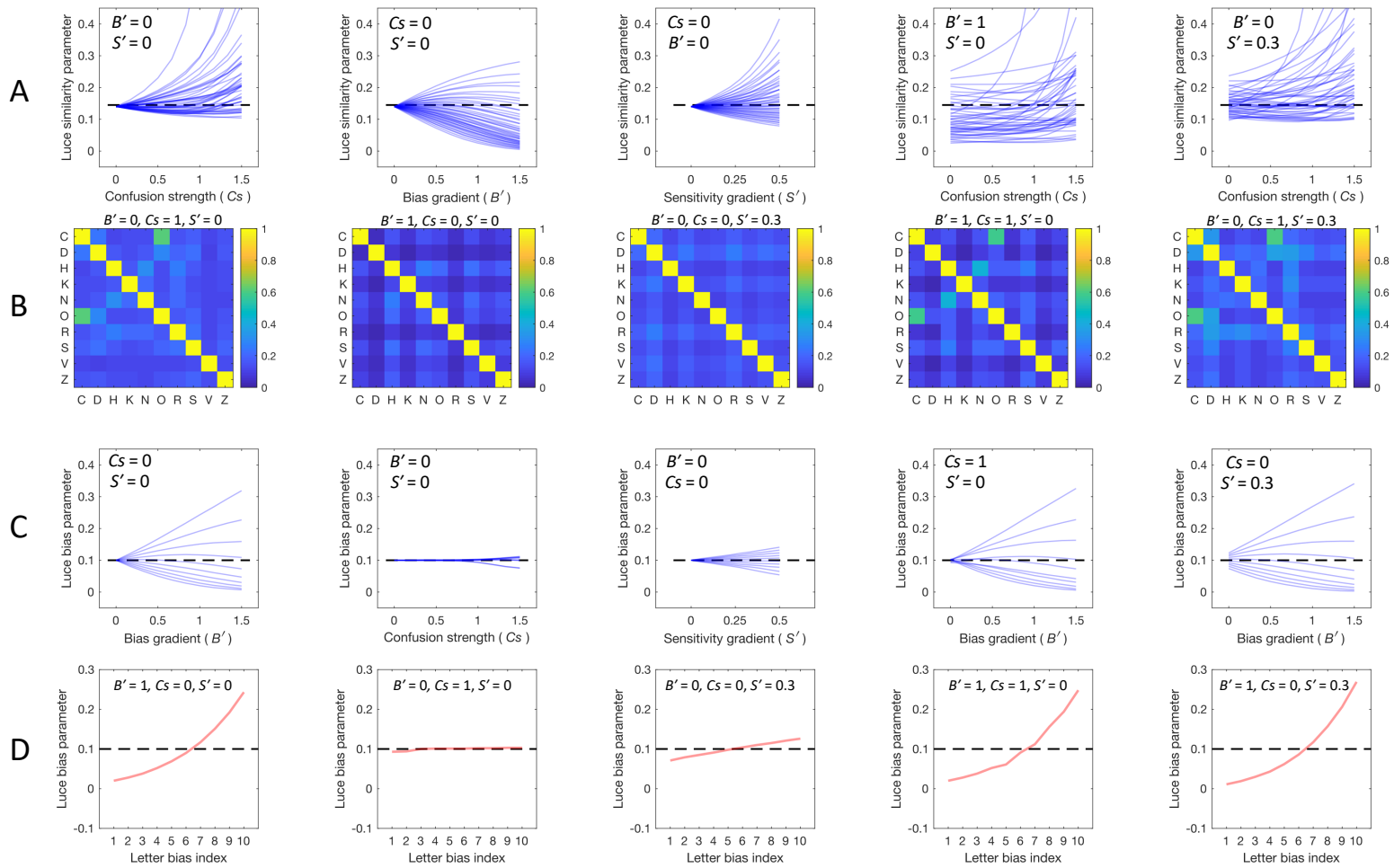


Figure 3.15 The effect of bias and sensitivity on the Luce's similarity (0 deg).

A) Luce's similarity parameters of the 45 confusion pairs for different conditions. B) the resultant similarity pattern C) Luce's bias parameters of the 10 letters for different conditions. D) The bias parameters for each letter indexed from 1 (least biased) to 10 (most biased).

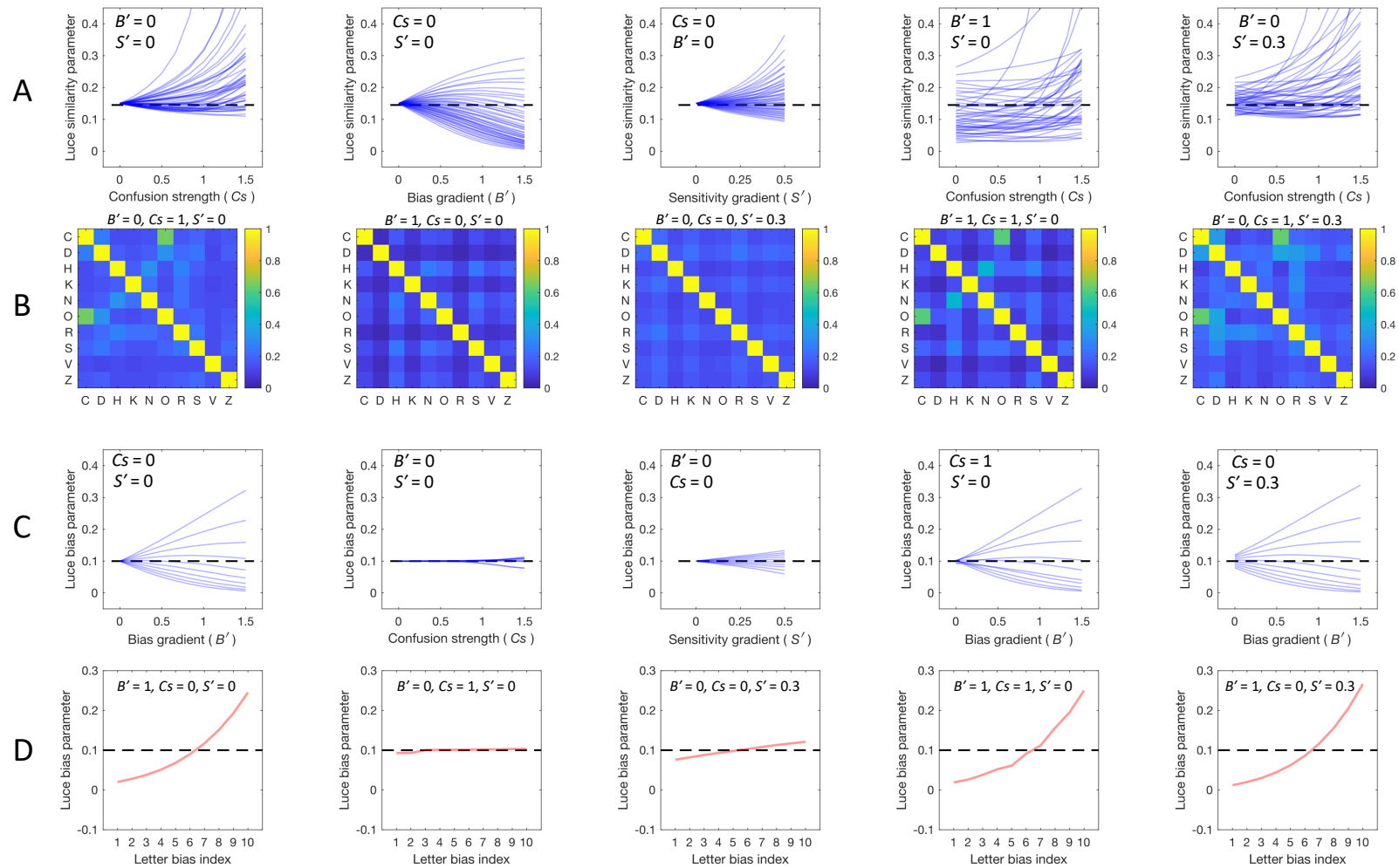


Figure 3.16 The effect of bias and sensitivity on the Luce's similarity (± 3 deg).

A) Luce's similarity parameters of the 45 confusion pairs for different conditions. B) the resultant similarity pattern C) Luce's bias parameters of the 10 letters for different conditions. D) The bias parameters for each letter indexed from 1 (least biased) to 10 (most biased).

3.4.4 Does the random error mimic the effect captured in experimental data?

3.4.4.1 Similarity between letters

It is very unlikely that the pattern of similarity observed in the experimental data and captured by the NTM model is a result of random error. Via simulations, I examined the hypothesis that random error alone (i.e., noise only with no effects of similarity, sensitivity or bias) can produce an *apparent* effect of letter similarity in the data. To examine this, I simulated data (trial by trial Monte Carlo simulation using the NTM without bias, sensitivity difference, or similarity between letters) that mimics the experimental task at the fovea (i.e., $p = 2.5$, $S_0 = 1.1$, 10 letters, 10 simulations per letter for each letter size of 0.3', 0.438', 0.64', 0.94', 1.37', 2') and at the periphery (i.e., $p = 2$, $S_0 = 0.5$, 10 letters, 10 simulations per letter for each letter size of 0.5', 0.79', 1.26', 1.99', 3.15', 5'). Subsequently, I fit model B0C1S0 using the same procedure described in the "Methods" section (Model fitting) to estimate the Cs in the simulated data. I repeated the procedure 10,000 times (i.e., 10,000 estimations of Cs) at the fovea and at the periphery separately. Finally, I compared the results to the Cs parameters estimated from the experimental data. Figure 3.17 shows that the similarity pattern observed in experimental data was very unlikely to be due to the random error alone at the fovea ($p = .0026$) and the periphery ($p < .0001$).

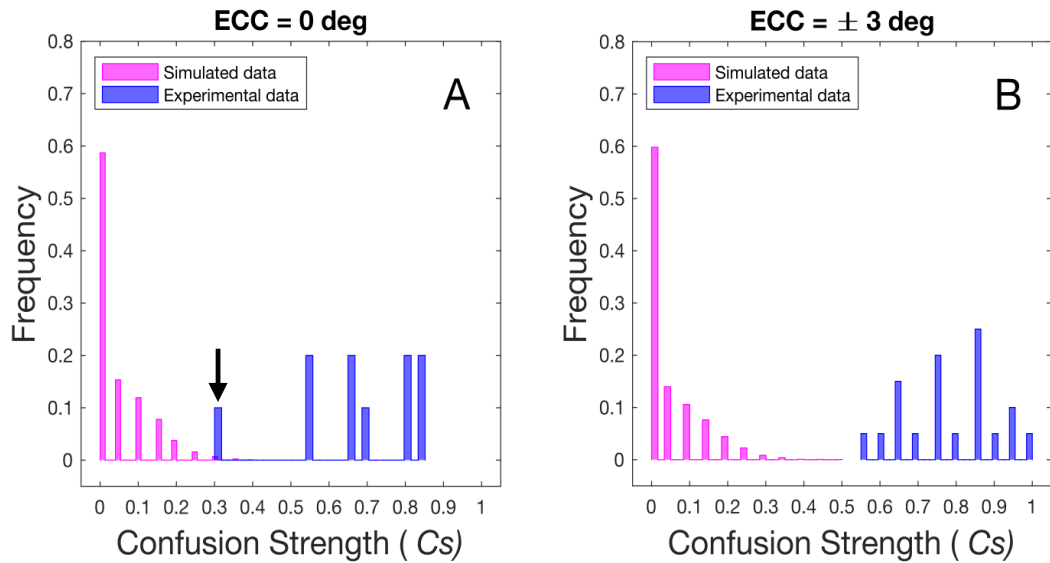


Figure 3.17 Frequency of estimated Cs from simulated and experimental data.

For experimental data we have 10 Cs parameters estimated at ECC = 0 deg and 20 Cs parameters at ECC = ±3 deg. The data were simulated without bias, sensitivity, or similarity. The Model B0C1S0 was unable to capture a substantial similarity between letters (indicated by low Cs) due to random error alone in the simulated data. Hence, the substantial similarity captured by the model B0C1S0 in the experimental data was very unlikely to be due to the effect of the random error at the fovea ($p = .0026$) (A). It was found that out of the 10,000 runs, only 26 showed Cs higher than the lowest Cs observed in the experimental data (i.e., only one observer showed Cs of 0.3, indicated by the black arrow) (A). A similar result was found at the periphery (B) ($p < .0001$).

3.4.4.2 Letter usage differences

In our model, we assumed that the differences in letter usages (total) are caused by either biases and/or sensitivity differences between letters. Here, through simulations, I test the hypothesis that the observed differences in letter usages (total) might be due to random error only and not as a result of the actual biases and/or sensitivity differences between letters.

Smooth curves in Figure 3.18 A show the outcome for the simulated data using the parameters of the basic model (B0C0S0) in which all templates had no bias and no sensitivity differences, but there was only *one trial* per condition (i.e., one trial for each test letter at each letter size, for each of 10 observers). There is a strong trend in the predicted proportion of letter usage (total) across the rank-ordered letters, but there were no biases or sensitivity differences in the model (i.e., B0C0S0). This is undoubtedly the sampling artefact in question. But in panel E we see that with 10 trials/condition (as in the real experiment), this artefactual trend is reduced by a factor of 3 or 4. The effect is almost completely absent with 100 trials/condition (panel I). In short, increasing the number of trials averages out the artefact.

The second column of Figure 3.18 shows almost the same outcome when the model had the sensitivity differences across templates, but still no bias (B0C0S1). With one trial per condition, the artefact was large, and the trend was similar to the real data - but this is surely a false trend, since the fits to the real data were poor when there were 10 or 100 trials per condition. A useful model should get better with more trials, not worse.

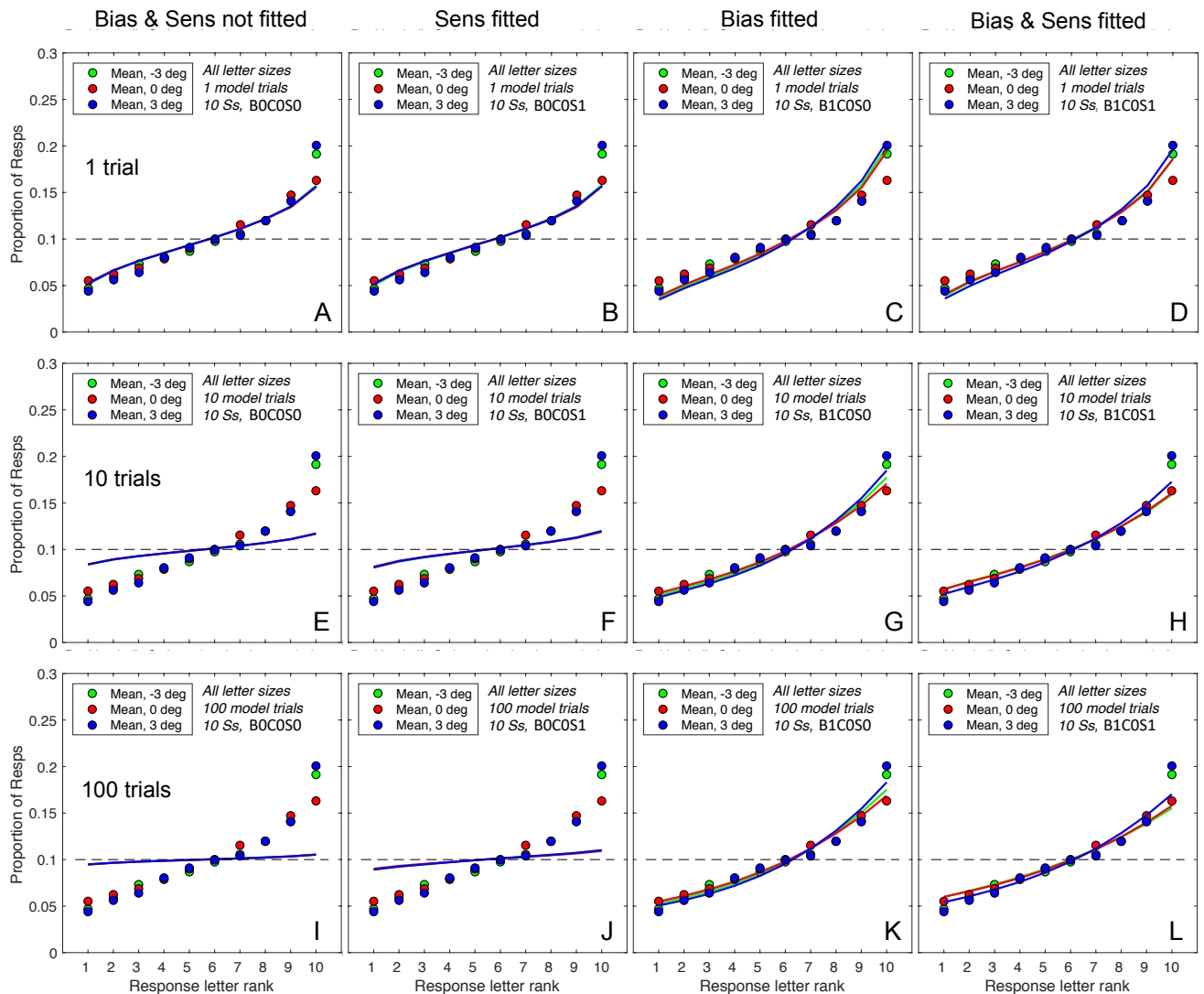


Figure 3.18 Simulated differences in letter usage (total) with different conditions.

We used our model as a tool to show that the profile of letter usage in our experiment arose mainly from biases and was not seriously skewed by noise. Data points are experimental values and are the same in all panels as already seen in Figure 3.1 C. Smooth curves are simulated letter usages (total) that differ as follows. Columns 1- 4 illustrate four versions of the model in which gradient parameters for bias B' and sensitivity S' were or were not fitted, as indicated (and denoted as B0C0S0; B0C0S1; B1C0S0; B1C0S1). There were always 10 model observers and six letter sizes, as in the experiment. Each run of a model (e.g., B0C0S1) used the best-fitting parameters for that model, as in Figure 3.9, but was intended as a simulation of the real experiment, using a Monte Carlo version of model response generation in order to include the effects of sampling variation, and 200 repetitions to smooth the final result. In practice, the influence of this last step was very small. Top row: just one trial per condition; middle row: 10 trials per condition (as in the real experiment); bottom row: 100 trials per condition.

The 3rd and 4th columns of Figure 3.18 show predicted patterns of letter usages for the two models with bias (B1C0S0; B1C0S1). With 10 or 100 trials, both models fit the usage data very well. Importantly, this tells us that for these two models, 10 trials are enough to eliminate the artefact. If that were not so, then the plots for 10 and 100 trials would be different, but they are almost identical (compare panel G with K, and H with L). It is with only one trial/condition, in panels C and D, that we see some trace of the artefact where the models over-estimate the experimental trend by a modest amount. However, this was eliminated with 10 trials/condition (as in the current experiment).

3.4.5 The effect of bias, sensitivity and similarity on letter acuity measurement and test-retest variability

There is a strong agreement in the proportion of correct responses between the *best* model (B1C1S1) and the experimental data where the bias played a major role (Figure 3.10). However, when correct responses are pooled across individual letters, the results show that letter acuity estimated from model B1C1S1 is not substantially different from those estimated from the other seven variants of the model (Table 3.6). This shows that bias, sensitivity, and similarity (in model B1C1S1) have negligible effects on estimated letter acuity when performance is assessed, as is usual in letter acuity studies, by pooling data over all letters.

Nevertheless, when I observe the spread of the PMFs (i.e., σ) of the eight models, I find that those taking account of the similarity factor (i.e., B1C1S1, B1C1S0, B0C1S1, B0C1S0) show a substantially higher and persistent spread than those without the similarity factor (Table 3.7). The increase in spread for these models (B1C1S1, B1C1S0, B0C1S1, B0C1S0) was found entirely in the upper half of the PMF (Tables 3.8 & 3.9).

This raises the question of what the impact is of this increase in σ on letter acuity estimation. It is well known that the slope of the PMF (reciprocal of σ) has an impact on the test-retest variability of the threshold when estimated by an adaptive method: a decrease in the slope (increase in σ) will increase the test-retest variability in the threshold estimation. To examine this hypothesis, I conducted Monte Carlo simulations of an adaptive method to investigate the effect of the observed increase of σ on the systematic error of average letter acuities and the change in test-retest variability over many measurements (100,000 simulated experiments). The

simulations show that, unlike bias or sensitivity, the similarity between letters increased the test-retest variability of the measurements substantially (especially at the periphery) when estimated with an adaptive method (i.e., a staircase, see Figure 3.13 B). However, the results also show that the effect of bias, sensitivity, or similarity on the systematic error was negligible (Figure 3.13 A & C). The test-retest variability affects the precision of threshold estimation of a given test in monitoring the progression of a clinical condition. High test-retest variability thus requires the test to be repeated more often to detect a clinically significant change. Therefore, it is important to minimise the test-retest variability in any test that is used to monitor the progression of a clinical condition.

Many factors have been found to increase the test-retest variability in visual acuity measurements, such as the visual acuity chart design, the optotypes used, the degree of optical defocus, the presentation time, the scoring and the termination rule used (Vanden & Wall, 1997; Siderov & Tiu, 1999; Rosser et al., 2003; Rosser et al., 2004; Shah et al., 2011). When adopting letters as optotypes in visual acuity measurements, many studies found that similarity between letters was a major source of test-retest variability. It has been shown that the test-retest variability can significantly be reduced by correcting the letter acuity scoring for similarity or by choosing a set of dissimilar letters (Grimm et al., 1994; McMonnies & Ho, 1996; Fülep et al., 2017; McMonnies & Ho, 2000).

In this study, I investigated, via simulation, the potential effect of bias, sensitivity and similarity on the test-retest variability expressed as the percentage increase in test-retest variability. In line with previous studies (Grimm et al., 1994; McMonnies & Ho, 1996; Fülep et al., 2017; McMonnies & Ho, 2000), our simulation results show that similarity between letters is a major cause of the increase in test-retest variability, and

with high similarity between letters, the increase can be as high as ~25%. With high bias or sensitivity, the increases in test-retest variability were much lower, about ~7% and ~10% respectively (Figure 3.13).

3.4.6 Conclusion

To summarise, this study aimed to reveal the role of bias, sensitivity differences, and similarity in a letter identification task simultaneously and to investigate the impact of these three factors on letter acuity measurements. Results show that the joint role of bias and similarity are more influential than the joint role of bias and sensitivity as previously shown (Georgeson et al., 2023). The results also demonstrate that similarity between letters is a major cause of the increase in test-retest variability when letter acuity is assessed with an adaptive method. Therefore, it is very important to minimise the effect of similarity between letters either by developing a new scoring system (McMonnies & Ho, 1996; Fülep et al., 2017; McMonnies & Ho, 2000) or by introducing a new letter set with minimal similarity (Grimm et al., 1994). This is crucial if the aim is to monitor the deterioration of visual acuity in eye conditions within a short period of time. In the following chapter, I will introduce a new letter set that shows high dissimilarity to be employed as stimuli in a paradigm to measure the sensitivity of the macula in a clinical condition.

Chapter 4 . Acuity Perimetry with Speech Input for Mapping Macular Visual Field in Glaucoma

4.1 Motivation

In this final study, I will examine the feasibility of measuring macular sensitivity in healthy and glaucomatous eyes using an acuity perimetry. The acuity perimetry is an approach that I have introduced, where letters are used as stimuli and speech recognition is used as the input method to enable observers to independently perform the task. The subjective experience of the experiment will also be investigated via a questionnaire. Extensive analyses will be conducted to explore the efficiency of different aspects of the approach and the paradigm used for the experiment. Additional analysis will be carried out to investigate potential sources of errors in letter acuity estimation using the proposed approach.

4.2 Methods

4.2.1 Observers

11 eyes from 11 healthy observers with no history of a systemic disorder or learning disability (10 females, mean age 26.60 ± 9.12 (SD), age range: 18 - 47 years) and 10 eyes from six primary open-angle Glaucoma observers with no history of a learning disability (three females, mean age 69.30 ± 8.62 , age range: 56 - 81 years, macular mean deviation = -8.40 ± 5.00 dB; M-pattern, Octopus 900, Haag-Streit, Switzerland) were recruited for the study (observers with diabetes and high blood pressure conditions but no history of strokes were included). The mean best corrected visual acuity and the mean refractive error (spherical equivalent) were 0.01 ± 0.09 logMAR and -1.11 ± 1.39 DS for the healthy eyes and were 0.12 ± 0.15 logMAR and 1.40 ± 1.50 DS for the glaucomatous eyes respectively. Written informed consent was obtained from all observers, and the study was approved by the University of Plymouth Ethics Committee. All experiments were conducted following the Declaration of Helsinki.

4.2.2 Stimuli

Nine letters that follow the Sloan design (i.e., A, T, O, L, N, S, U, Y, Z) were adopted (i.e., The stroke width = $1/5$ of the letter's height. The height and the width of the letter are equal. See Figure 2.1 B). They were black letters (luminance = 2.2 cd/m²) on a white background (luminance = 215 cd/m²), resulting in 99% Weber contrast according to Eq. 2.1. The letters were chosen by informal multiple pilot experiments to ensure that they show high dissimilarity between each other *visually* and

phonetically. The visual dissimilarity between the letters was examined by referring to the optotype correlation matrix of the 26 letters (Fülep et al., 2017). The letters' phonetic dissimilarity was examined by multiple pilot experiments carried out by the author. The phonetic dissimilarity was found to be highly subject-dependent and the validation test to choose only reliable letters (among the nine letters) was necessary.

4.2.3 Apparatus

An Apple MacBook Pro (2019) with a 1.4 GHz Intel Core i5 processor and an Intel Iris Plus Graphics 645 - 1536 MB graphic card was used to run the experiment. The stimuli were presented on the MacBook Pro screen (LED-backlit retinal display of 13.3-inch diagonal size and 2560×1600 native resolution). Observers viewed the targets at a viewing distance of 70 cm while sitting on a chair without using a chin or forehead rest. The experimenter checked the distance frequently to ensure a constant viewing distance. At this distance, one pixel subtended $0.54'$. A speech recognition algorithm (Selvaraju, 2022) was employed to capture the responses given verbally by the observers via a regular microphone attached to a headphone (model: Jabra Evolve 20 HSC016, speaker frequency range; wideband: 150 Hz to 6800 Hz; narrowband: 300 Hz to 3400 Hz, microphone frequency range: 100 Hz to 8000 Hz).

4.2.4 Software

The stimulus generation, data collection, statistical analysis and simulations were performed using *Matlab* 2019a (MathWorks, Natick, Massachusetts, USA). Routines from the Psychtoolbox-3 were used to present the stimuli (Brainard, 1997; Pelli, 1997;

Kleiner et al., 2007).

4.2.5 Procedure

4.2.5.1 Data collection

All tests were conducted monocularly. The left or right eye was selected at random for healthy observers. However, for Glaucoma observers, the eye that showed a history of macular visual field damage was selected for the experiment. In all cases, the fellow eye was occluded using an opaque eye patch. The task for the observer was to recognise individual letters randomly presented at eccentricities of 0, 1, 3 and 4 deg. A fixation cross was presented in the middle of the screen; dimensions: length/width 3.27', stroke width 1.09'. 13 test locations were examined (Figure 4.1). The presentation time of the stimulus was 250 milliseconds. The screen angle was adjusted so that it was parallel to the plane of the observer's face. The fixation cross disappeared when a letter was presented foveally and reappeared immediately when the letter presentation was finished. After the presentation of the letter, observers had a response window of 1500 milliseconds as indicated by two auditory beep sounds at the start and end of the window. The observers were asked to give their verbal responses within the time in-between these two beeps (Figure 4.2). Additionally, they were asked to try their best to make the intonation of verbal responses as consistent as possible. This was important to ensure that the speech recognition algorithm worked efficiently. After each response (i.e., after the second beep), the observer needed to press the spacebar to start the next trial. This gave them the chance to pause at any point during the experiment for a break. The headphone with the attached microphone was used for all observers to hear the beeps and to enter verbal responses. The speech recognition was

programmed so that it recognised only the letters from the letter set. If the observer gave a response different from the adopted letter set, the speech recognition algorithm chose the letter that was closest to the response speech pattern of the nine letters set. Prior to the experiment, the observers were shown the adopted nine-letter set with a large size (1 logMAR) (further details below). No instructions were given to observers to restrict their response to this nine-letter set during the main experiment.

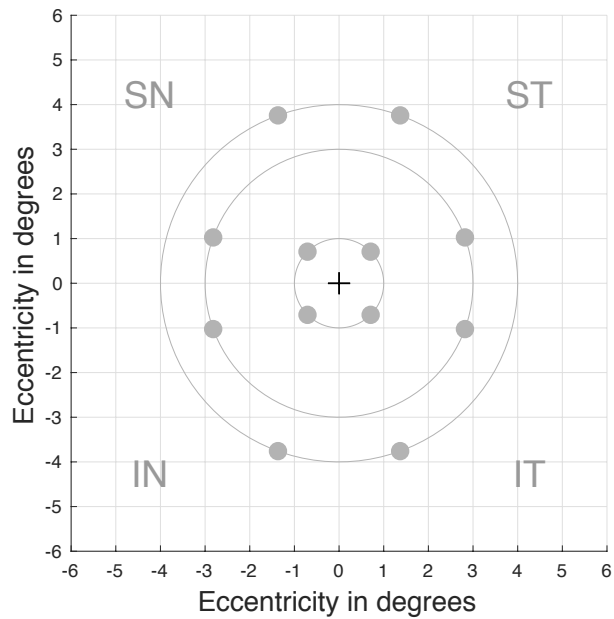


Figure 4.1 The fixation cross and the test locations (grey circles). SN, ST, IT, and IN stand for the superior nasal, superior temporal, inferior temporal, and inferior nasal quadrant respectively.

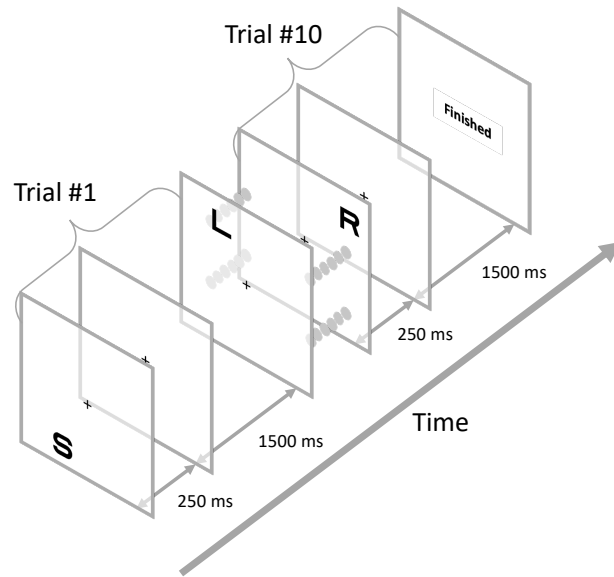


Figure 4.2 shows the experimental paradigm.

The experiment was carried out under room illumination of 160 lux without noise-proofing. Multiple pilot trials (performed by the author) showed that the detection of the verbal responses via the regular headphone microphone used in the current experiment was very efficient to filter out the background noise in an office-like environment.

The experiment was divided into two parts. After a brief verbal explanation of the task, the participant started the first part of the experiment. The first part consisted of the training and validation stage of the speech recognition algorithm, immediately followed (without perceived distinction) by the first session of the visual field experiment (i.e., acuity perimetry). The training and validation stage of the speech recognition algorithm started by presenting large letters (1 logMAR) randomly at any of the 13 test locations. The participants were naïve to the purpose of this stage. 10 audio recordings were obtained for each letter. These recordings were subsequently used to build the subject-specific acoustic models for the nine letters. The efficiency of the algorithm was tested immediately after the recording stage by presenting each

letter 10 times with a large size (1 logMAR) randomly at any of the 13 test locations. The verbal responses to these presentations were captured and processed by the speech recognition algorithm. For each letter, the efficiency was evaluated by the correct responses achieved out of the 10 presentations. Only the letters that achieved at least nine out of 10 correct responses in the validation stage were chosen for the acuity perimetry experiment. In two cases, the algorithm failed to achieve nine out of 10 correct responses for any of the letters and therefore I adopted a rate of eight out of 10 for these two cases.

After the speech recognition training and validation stage was completed, the first session of the acuity perimetry experiment started. After the first experiment was finished, the second session was initiated. The second session consisted of the acuity perimetry experiment using only the validated letters from the first part. The observers were asked to give responses to all the presented letters, and they were asked to guess in case of uncertainty. The observers were asked to maintain fixation on the fixation cross. Fixation compliance was frequently encouraged by the examiner before and between the two sessions. The observers were not aware of the various stages of the experiment. Each observer independently (i.e., without supervision) completed two sessions of the experiment. On average, the first part took 20 minutes (i.e., speech recognition validation stage + the first acuity perimeter experiment) and the second part took 10 minutes (i.e., the second acuity perimeter experiment) to complete.

4.2.5.2 Observers' experience of the experiment

The feasibility of the test was investigated by monitoring the ability of the observers to carry out the full experiment (i.e., two parts) without additional help from the

examiner. After the two parts were completed, each observer filled out a questionnaire that assessed the feasibility of the test. Structured questions were developed to investigate the observers' perception of different aspects of the experiment such as the length of presentation time, the verbal response method, the letter recognition task, the presentation of very small letters, the need to guess in case of uncertainty, the maintain of fixation and the occlusion of the unexamined eye. The responses to the questions were quantified using a Likert scale (see Appendix B).

4.2.5.3 Letter acuity measurements

At each test location, the letter acuity was estimated using the Fixed Step Size (FSS) staircase. A modified staircase of the one up / one down paradigm was used with step-down size = $0.8081 \times$ step-up size to target ~55% percent correct of the underlying psychometric function (PMF) (Eq. 3.7, Garcia-Pérez, 1998). The step-up size was set to 0.25 logMAR for eccentricities 0 and 1 deg and to 0.4 logMAR for eccentricities 3 and 4 deg. The starting letter size of the staircase was 0.3 logMAR for all the test locations. All staircases terminated after 10 trials per location and the letter acuity was estimated as the average across the last four reversals. At each test location, the average letter acuity of the two sessions was used in the data analysis. The results obtained from the left eyes were mirrored to match the corresponding points in the right eyes. The results were analysed and presented using right eye quadrants notation. An additional visual field test was carried out for glaucomatous eyes for comparison (macular visual field test; M-pattern, standard strategy, Octopus 900, Haag-Streit, Switzerland).

4.3 Results:

4.3.1 Letter acuity measurements

Figure 4.3 A shows the average estimated letter acuity (\pm SE) at each test location across the 11 healthy eyes. Qualitatively, the results follow the expected gradual decrease in acuity from the fovea towards the periphery. The letter acuity ranged from -0.13 ± 0.03 logMAR at fovea to 0.36 ± 0.05 logMAR at 4 deg eccentricity (Figure 4.3 C).

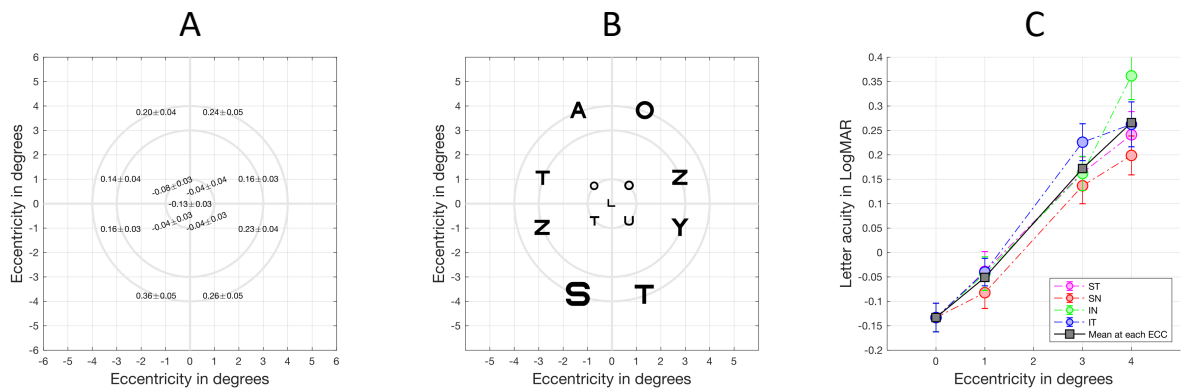


Figure 4.3 Results of acuity perimetry for healthy eyes.

A) the average (\pm SE) of the estimated letter acuity (logMAR) across healthy eyes for each test location. B) the letters (chosen at random) relatively scaled to the corresponding average letter acuity and C) the estimated letter acuity as a function of eccentricity for each quadrant (the error bars are the SE).

Results show a good topographical correspondence between the results of the M-pattern test and the current acuity perimeter in all glaucomatous eyes (Figure 4.4).

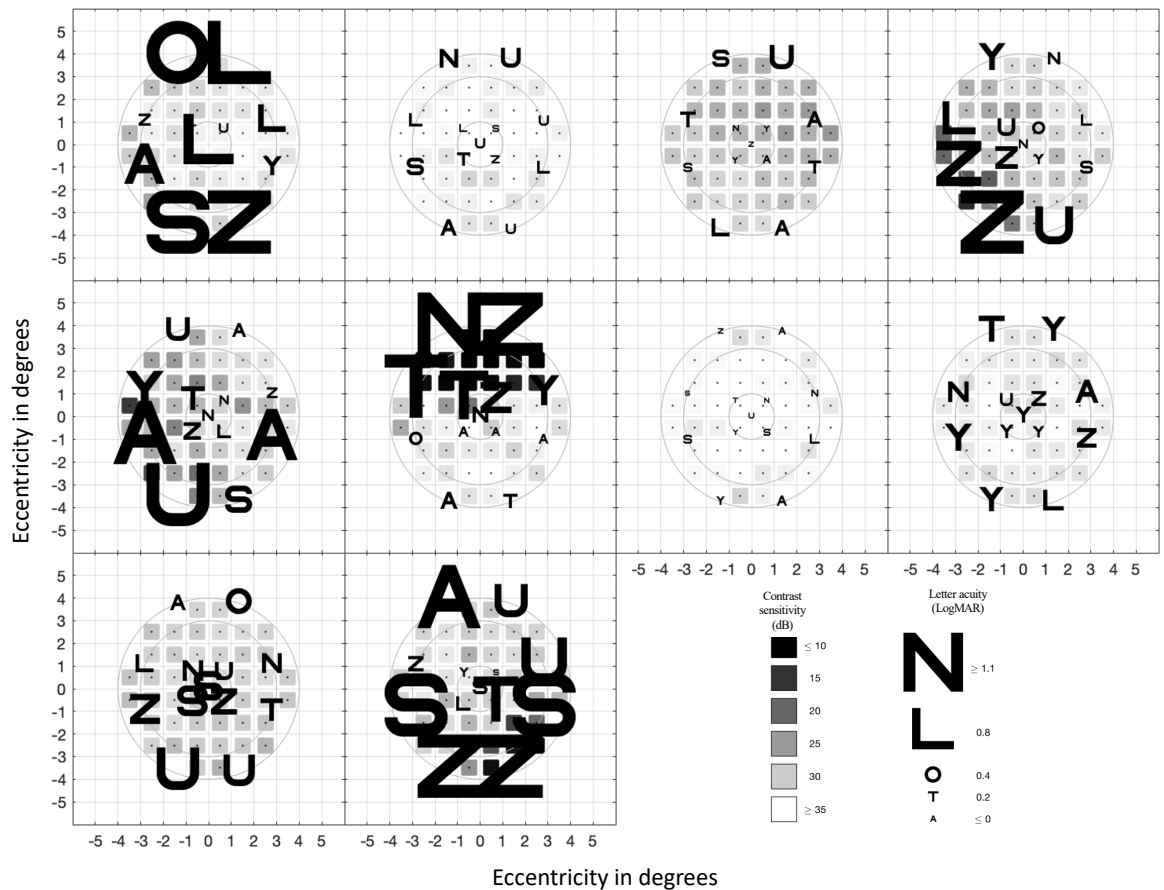


Figure 4.4 Results of acuity perimetry and M-pattern test for glaucomatous eyes. Figure shows the topographical correspondence between the estimated letter acuity (logMAR) (i.e., letters scaled to the corresponding acuity) and the contrast sensitivity (dB) (i.e., grayscale map). Each plot is the result of one glaucomatous eye. The dots in the middle of the grey-shaded squares are the test locations of the M-pattern visual field test.

To further explore the estimated letter acuity in the glaucomatous eyes, a pointwise agreement in measurements between the letter acuities at 12 test locations (data from fovea were excluded because of missing data from the M-pattern test) and the contrast sensitivity of the corresponding points of the M-pattern test was conducted. Since there were no exact corresponding test points between the acuity and M-pattern perimeters, the pairs with the closest proximity were considered corresponding points (Figure 4.5).

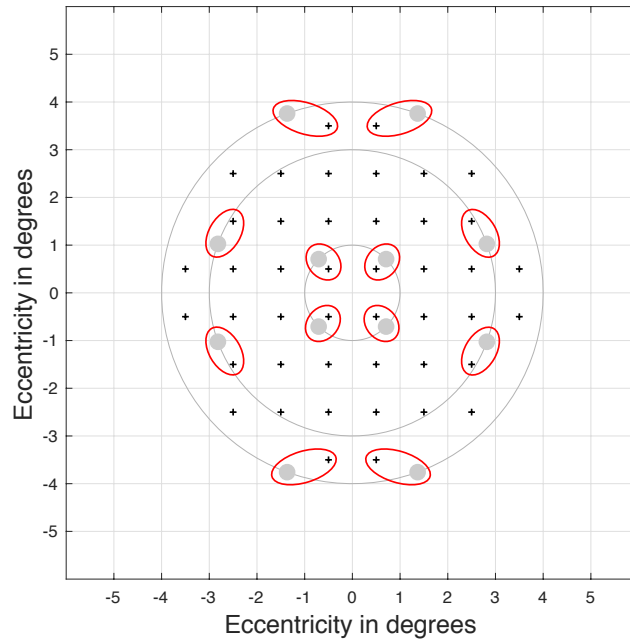


Figure 4.5 Corresponding test points of the acuity and the M-pattern perimeters. The test locations of the acuity perimeter and the M-pattern are the grey circles and the black crosses respectively. The corresponding points are circled with red oval shapes.

Although the correlation (Pearson correlation, $n = 10$) was significant at only five locations (ranged from $r = -0.68$, $p < .05$ to $r = -0.93$, $p < .001$), the pointwise comparison shows a clear trend (i.e., negative correlation) between the acuity (logMAR) and contrast sensitivity (dB) for all test locations (Figure 4.6).

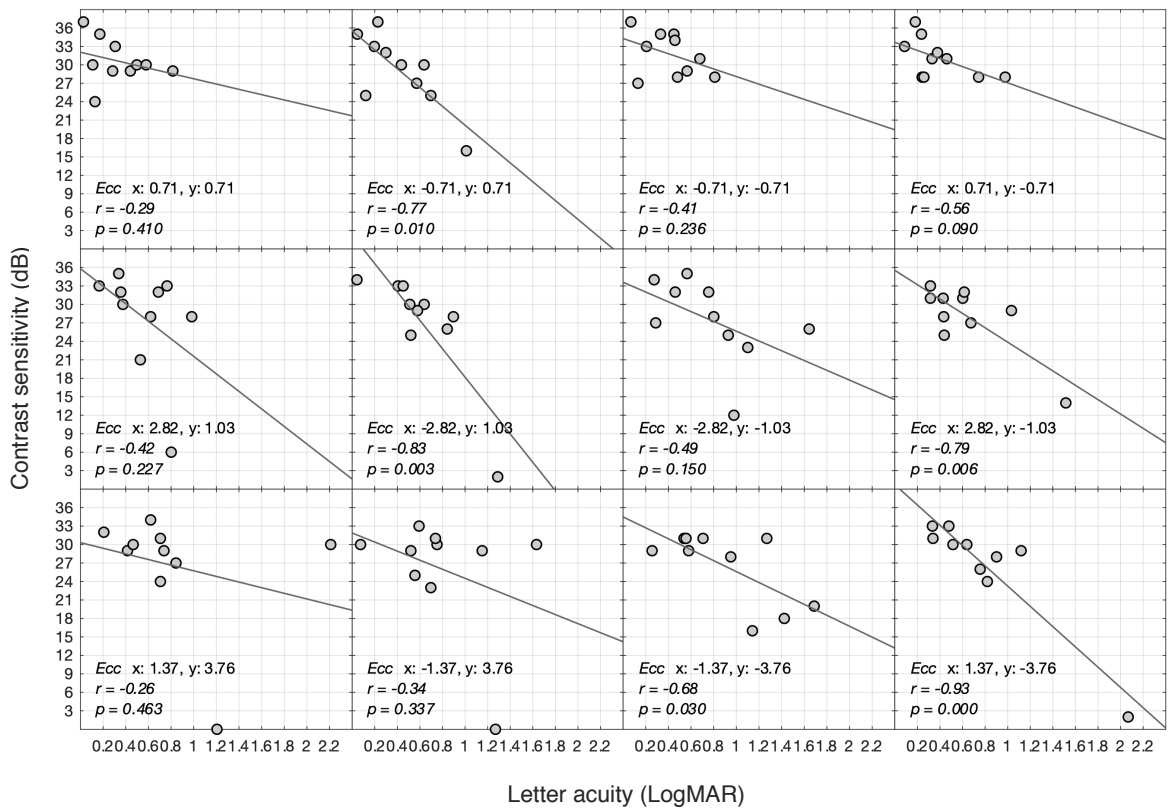


Figure 4.6 The pointwise agreement of the letter acuity vs contrast sensitivity.

Figure shows results of the 12 test locations. Each panel represents the data (from the 10 eyes with Glaucoma) from one test location of letter perimeter indicated by the x and y coordinates in degrees and the corresponding test location of the M-pattern test.

Additionally, I examined the correlation between the acuity and the M-pattern perimeters in measurements averaged across test locations of each quadrant. The correlation (Pearson correlation, $n = 10$) was found to be significant at the SN ($r = -0.80$, $p < .05$), IT ($r = -0.85$, $p < .05$) and IN ($r = -0.73$, $p < .05$) quadrants. However, the correlation was not significant at the ST ($r = -0.48$, $p = .17$) quadrant (Figure 4.7).

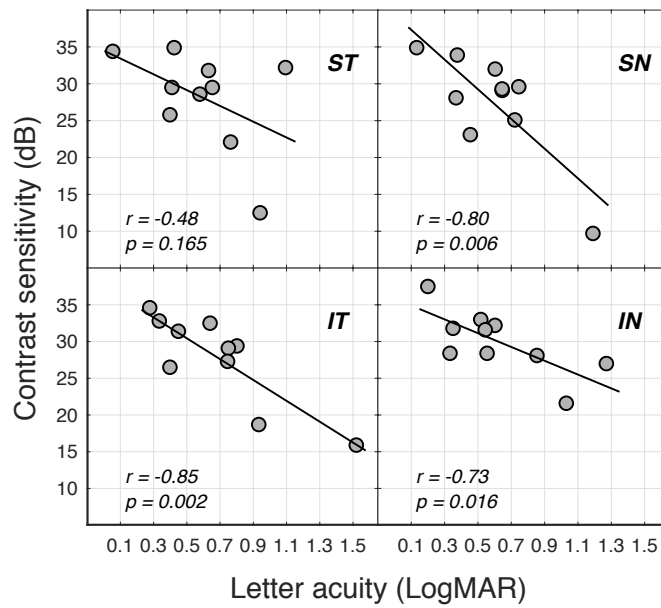


Figure 4.7 The correlation of letter acuity vs contrast sensitivity at each quadrant. Measurements averaged across test locations of each quadrant.

Further analysis showed that the decrease in letter acuity was strongly associated with the decrease in contrast sensitivity as revealed by the strong and statistically significant correlation between the acuity perimetry and M-pattern measurements (Pearson correlation; $r = -0.90$, $n = 12$, $p < .001$) (Figure 4.8).

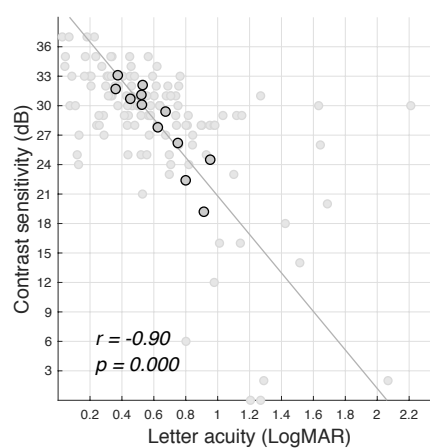


Figure 4.8 The correlation of letter acuity vs contrast sensitivity at all locations. The correlation was calculated between the letter acuities of the 12 test locations (rank-ordered from low to high acuities and then averaged across observers) and the corresponding measurements from the M-pattern test (dark grey circles). The faint grey circles show the individual data points of each observer.

4.3.2 Observers' experience of acuity perimetry

These results demonstrate that the acuity perimetry approach was reliable in terms of the task (i.e., letter recognition) and the input method (i.e., speech recognition algorithm) in healthy and glaucomatous eyes. However, to further assess the observers' experience of the experiment, the observers were asked to complete a questionnaire after completion of the experiment. The questionnaire consisted of 11 items, and it showed good reliability and internal consistency (Cronbach's Alpha = 0.76). Using a Likert scale, on average (median, interquartile range (IQR)), observers with healthy eyes (median, 9.00; IQR, 8.00 to 10.00) and those with glaucomatous eyes (median, 8.00; IQR, 7.00 to 8.00) reported that the experiment was very easy to perform (where 0 = extremely difficult, 10 = extremely easy). Observers also reported that the letter recognition task and the speech recognition method have a low negative impact on their performance in the experiment (Figure 4.9). The most difficult part of the test was reported to be the presentation of very small letters where guessing was required. Maintaining fixation was also reported to have a medium negative effect. Nevertheless, all observers were happy to sit for the experiment again and even weekly if a shorter version of the test would be developed. Table 4.1 shows the median and IQR for each test group.

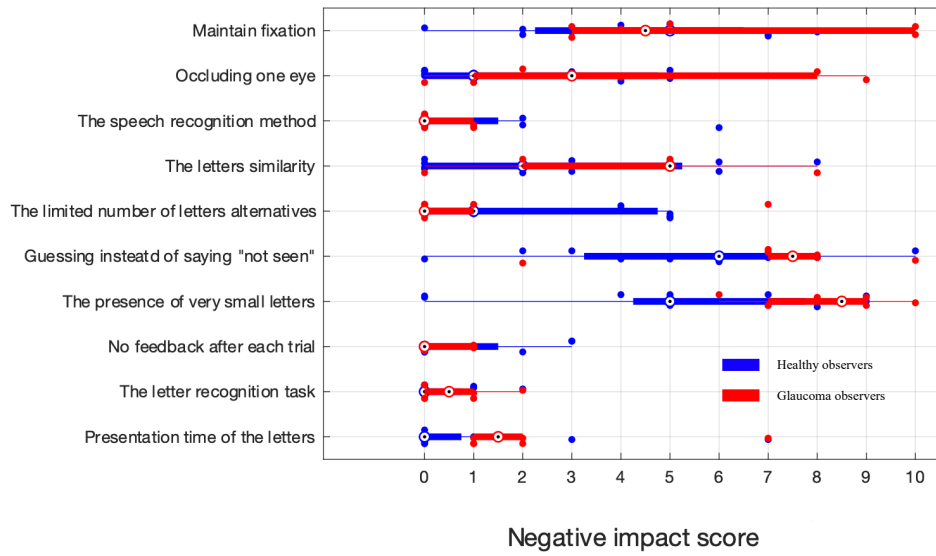


Figure 4.9 Observers' responses to the questionnaire.

The Figure shows; for different factors (y-axis), the individual (red and blue dots) and group (blue and red circles as medians and bars as IQR for healthy and Glaucoma observers respectively) negative impact scores (scores from 0 to 10 where 0 = no effect and 10 = huge negative effect) in healthy and Glaucoma observers.

Table 4.1 shows the median, 25th & 75th percentile quartiles, and the statistical test results of comparing scores of responses to 10 items of the questionnaire in healthy and Glaucoma observers.

Item	Healthy observers			Glaucoma observers		
	Median	25th Q	75th Q	Median	25th Q	75th Q
Maintain fixation	5.00	2.00	6.00	4.50	3.00	10.00
Occluding one eye	1.00	0.00	5.00	3.00	1.00	8.00
The speech recognition method	0.00	0.00	2.00	0.00	0.00	1.00
The letters similarity	2.00	0.00	5.00	5.00	2.00	5.00
The limited number of letters alternatives	1.00	0.00	5.00	0.00	0.00	1.00
Guessing instead of saying "not seen"	6.00	3.00	7.00	7.50	7.00	8.00
The presence of very small letters	5.00	4.00	8.00	8.50	7.00	9.00
No feedback after each trial	0.00	0.00	2.00	0.00	0.00	1.00
The letter recognition task	0.00	0.00	1.00	0.50	0.00	1.00
Presentation time of the letters	0.00	0.00	1.00	1.50	1.00	2.00

Notation: 25th Q, 25th percentile quartile; 75th Q, 75th percentile quartile.

4.3.3 The reliability of the speech recognition algorithm

The speech recognition algorithm employed in the current experiment was subject-dependent. I, therefore, performed a validation test for each observer before they proceeded to the acuity perimetry. Only the validated letters were used in the experiment. On average, the validated letters were found to be 6.70 ± 1.60 in healthy eyes and 5.6 ± 2.60 in glaucomatous eyes out of nine letters. Since I parametrised the FSS staircase to target $\sim 55\%$ of the underlying PMF assuming that the guessing rate was 11.11% (i.e., one-interval 9AFC task), it was important to verify that the number of alternatives (i.e., letters) adopted by observers in responses during the experiment were not affected by the number of presented (i.e., validated) letters. Although less than nine letters were validated in some cases, the observers continued to guess from on average 8.00 ± 0.80 letters in healthy observers and 8.00 ± 1.16 letters in Glaucoma observers (Figure 4.10). Since the validated letters were those which achieved at least nine out of 10 tested letters in the validation stage, I assumed some lapses to occur (i.e., 2.00 %) due to the potential failure of the speech recognition algorithm to detect some well visible letters. The lapse rate of 0.02 will maintain performance at $\sim 55\%$ in the case of adopting eight alternatives.

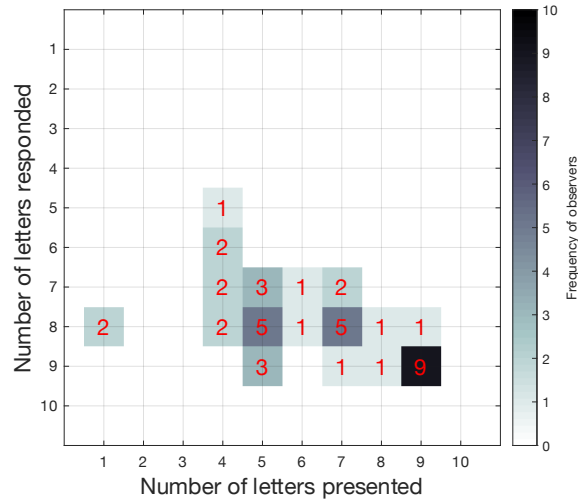


Figure 4.10 The number of letters presented (i.e., validated) vs responded to. The grey colour scale shows the frequency of observers for each cell where darker colour represents higher frequency. The number of observers for each cell (i.e., numbers in red) was also added.

4.3.4 Reliability of FSS staircase

4.3.4.1 Performance of FSS staircase

The parameters used in the FSS staircase in the current experiment were found to be efficient to produce on average 5.73 ± 1.55 (mean \pm SD) reversals within 10 trials for all observers (5.80 ± 1.5 and 5.60 ± 1.6 reversals in healthy and glaucomatous eyes respectively) (Figure 4.11 A). In the current experiment, the letter acuity was estimated by the average across the last four reversals. However, 21 out of 286 (i.e., 7.3%) staircases in healthy eyes and 27 out of 260 (i.e., 10.4%) staircases in glaucomatous eyes produced less than 4 reversals (Figure 4.11 B). 15 (i.e., 5.2%) staircases in healthy eyes and 19 (i.e., 7.3%) staircases in glaucomatous eyes produced three reversals; 6 (i.e., 2.1%) staircases in healthy eyes and 5 (i.e., 1.9 %) staircases in glaucomatous eyes produced two reversals; and none of staircases in healthy eyes and only 2 staircases (i.e., 0.8%) in glaucomatous eyes produced one reversal. In all these cases (i.e., with reversals less than four per staircase) the letter acuity was estimated by the

average across the available reversals. In glaucomatous eyes only, one staircase (i.e., 0.4%) failed to generate any reversal. This might be due to a deep scotoma (i.e., almost blind area) at that test location. In the latter case, the letter acuity was estimated by averaging the last two trials. On average, the number of reversals was not influenced substantially by the estimated letter acuity except for three extreme cases observed in glaucomatous eyes (Figure 4.11 B).

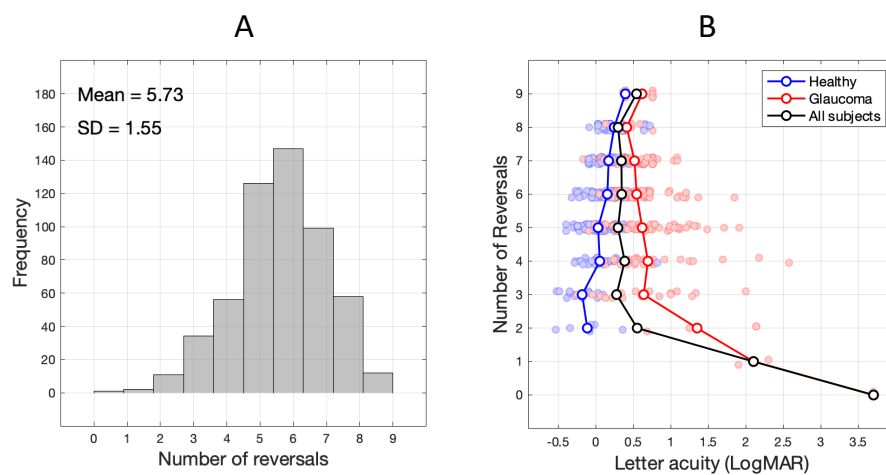


Figure 4.11 The number of reversals in all staircases.

A) Shows the distribution of the number of reversals of staircases obtained from all experiment sessions for all observers. B) The individual number of reversals (faint-coloured circles) for each estimated letter acuity in the experiment. The connected central data points show the average across the corresponding group. Jittering was added along the y-axis to improve the visibility of individual data points.

4.3.4.2 Test–retest variability

The reliability of the FSS paradigm in estimating the letter acuity was examined by calculating the correlation (Pearson correlation; $n = 21$ (normal and Glaucoma observers)) in letter acuity between the two sessions at each test location. The correlations in letter acuities between the two sessions were found to be statistically significant for 12 test locations (at $p = .001$ for six locations and at $p = .05$ for six

locations). The correlation was not statistically significant at the fovea ($r = 0.31$, $p = .175$) despite a clear visual agreement (Figure 4.12).

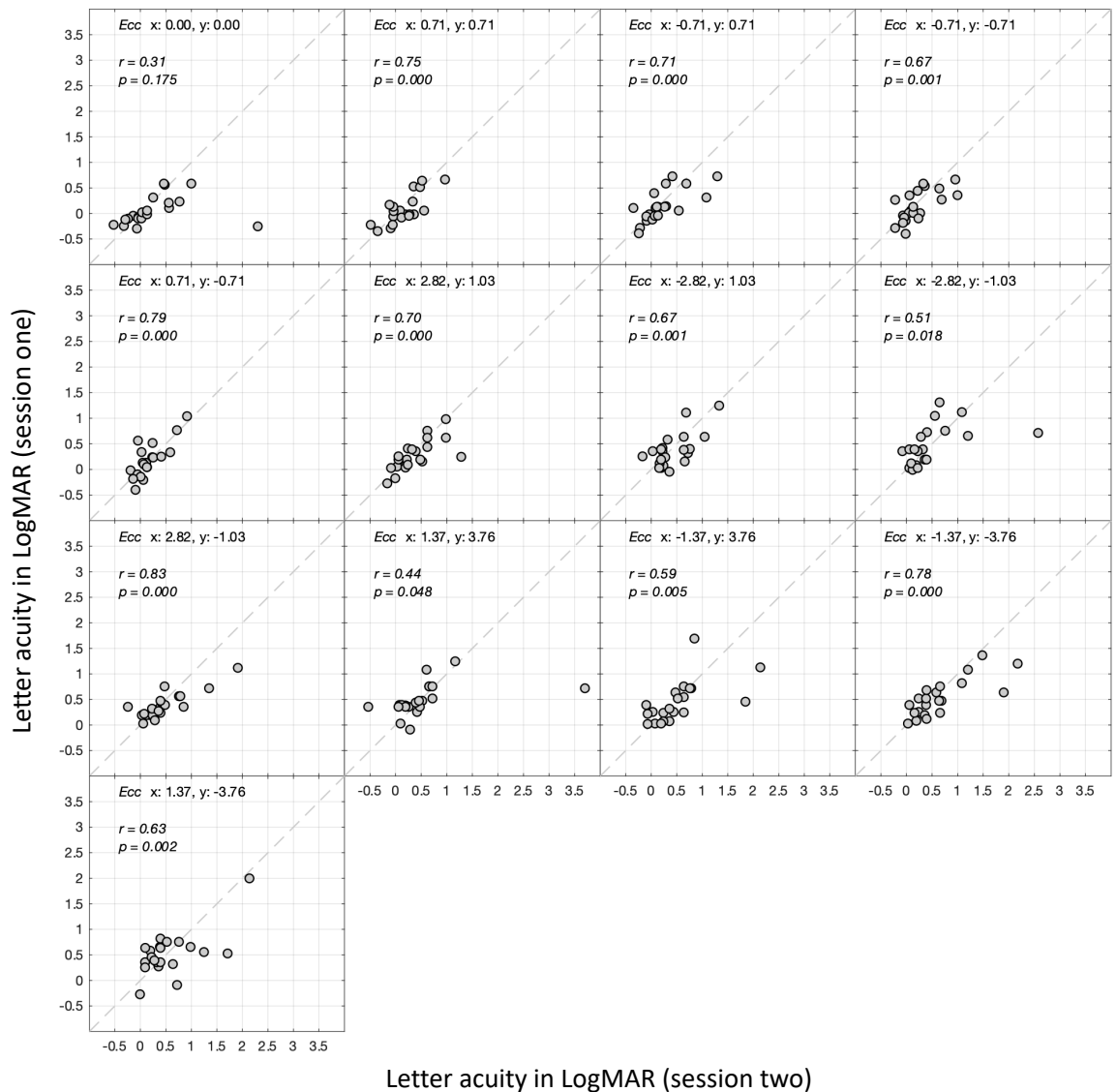


Figure 4.12 The agreement in measured letter acuity between the two sessions. The Figure shows the results at each test location (expressed as Eccentricity (Ecc in the plot) in x and y coordinates in degrees). The grey diagonal dashed line is a 1:1 line, i.e., perfect agreement.

Further investigation was conducted using the Bland-Altman analysis (Bland & Altman, 1999) to determine the mean of differences in the measurements between the two sessions and to calculate the Smallest Detectable Change (SDC) in letter acuity in the current experiment. SDC was calculated as the range between the 95% limits of

agreement between the two sessions. The Bland-Altman analysis was performed for each test location separately (Figure 4.13). For each location, the difference in measurements that exceeded three standard deviations from the mean of differences was considered an outlier and excluded from the analysis (Parrinello et al., 2016). The procedure was repeated until there were no outliers. Using this method, I identified six outliers in total; one outlier was found at each of the four different test locations and two outliers were found in one of the test locations (depicted as black asterisks in Figure 4.13).

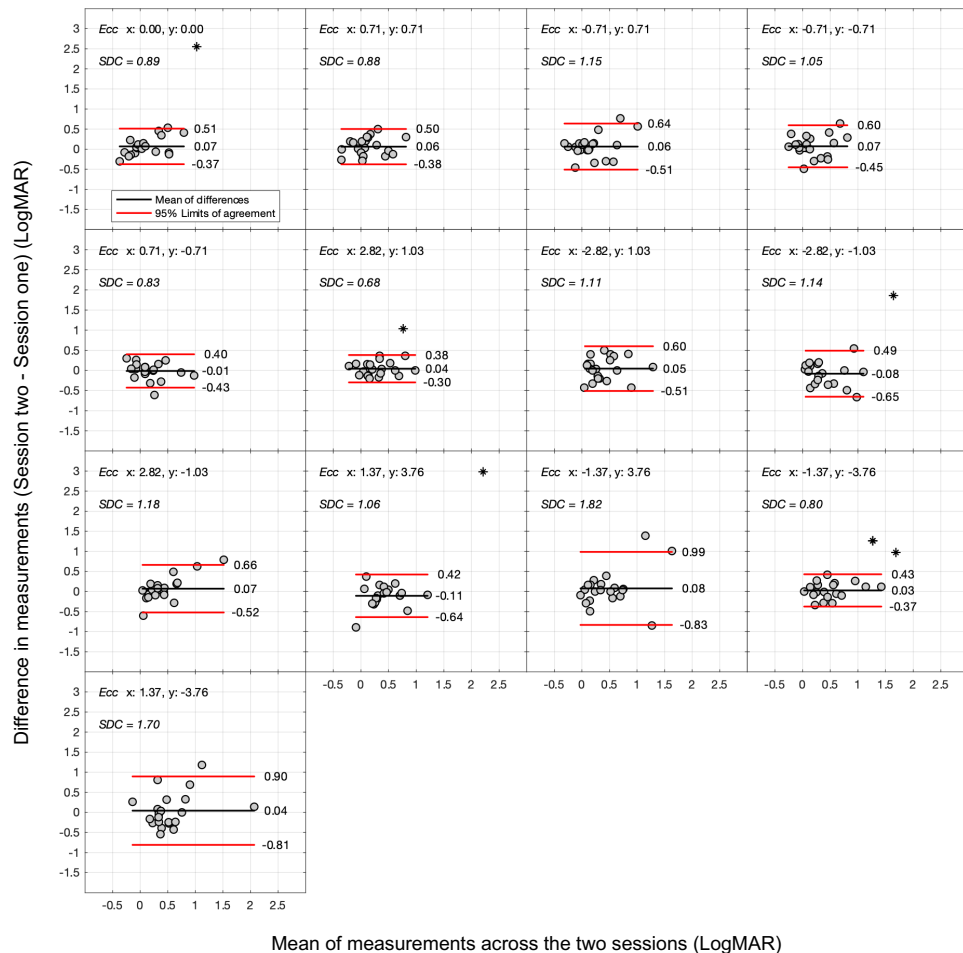


Figure 4.13 The Bland-Altman plots at each test location.

The grey circles are data from individual observers. The black asterisks are the identified outliers (see text).

The Bland-Altman analysis showed that the absolute means of differences in letter

acuity measurements between the two sessions were negligible at all test locations (average 0.06 ± 0.025 ; ranged from 0.01 to 0.11 logMAR). However, the SDC was found to be high at all test locations (average 1.10 ± 0.33 ; ranging from 0.68 to 1.82 logMAR).

4.3.5 The effect of the lapse rate

4.3.5.1 Lapse rate in the experimental data

In the current experiment, I investigated the hypothesis that the source of the high SDC is primarily due to the high lapse rate that is associated with the speech recognition method. To explore this hypothesis, I calculated the lapse rate in the experimental data. I defined the lapse rate as the proportion of incorrect responses out of the total number of the *well-visible* letters. The well-visible letters are the presented letters with sizes that are larger than the estimated letter acuity by at least 0.1 logMAR for each test location and each observer. I calculated the lapse rate at the group level and the results are shown in Table 4.2 for a range of letter sizes at which the well-visible letter is defined (i.e., from 0.1 to 0.4 in 0.05 logMAR steps larger than the estimated letter acuity at each staircase) (Table 4.2). The lapse rate was ~11% when calculated for the inner two eccentricities (i.e., fovea and 1 deg), the outer two eccentricities (i.e., 3 and 4 deg), or for all test locations. It is not unexpected to observe such a high lapse rate in the current experiment since I employed the speech recognition algorithm with an accuracy tolerance of 90% (80% in two sessions; see section 4.2.5.1). Additionally, I set a response window of 1500 milliseconds between trials which might have increased the possibility of lapsing because the correction for accidental errors was impossible (i.e., re-entering the responses for mispronounced but correctly perceived letters).

Table 4.2 shows the calculated lapse rates at different letter sizes that are larger than letter acuity.

	Test location (deg)	Letter sizes larger than letter acuity by (logMAR):						
		0.10	0.15	0.20	0.25	0.30	0.35	0.40
Number of well-visible letters	0 & 1	751	569	419	299	222	150	94
	3 & 4	1168	981	803	602	442	318	232
	All	1919	1550	1222	901	664	468	326
Number of correct responses to the well-visible letters	0 & 1	663	508	378	265	198	135	83
	3 & 4	1019	875	720	539	401	286	205
	All	1682	1383	1098	804	599	421	288
Lapse rate	0 & 1	12%	11%	10%	11%	11%	10%	12%
	3 & 4	13%	11%	10%	10%	10%	10%	12%
	All	12%	11%	10%	11%	10%	10%	12%

4.3.5.2 Simulating the effect of lapsing on letter acuity measurements

To investigate the potential effect of lapsing on the mean of differences in the letter acuity measurements between the two sessions and on the SDC in the current experiment, I carried out further analysis via simulation. A Monte Carlo simulation was performed to generate data from 42 sessions (two sessions of 21 eyes; similar to the case in the experimental data). The data were generated assuming that the underlying PMF of the current task is a Log-Quick function with a slope (β) of 3.5 for the two inner eccentricities and 3 for the two outer eccentricities. The guessing rate (γ) was set to 11.1% (one-interval 9AFC). For initial exploration, the ground truths were set to 0.10 and 0.40 logMAR for the inner and outer eccentricities respectively for all simulations (i.e., the threshold of the underlying PMF (α) was set to be 0.10 logMAR for the inner and 0.40 logMAR for the outer eccentricities). Data were generated using a FSS staircase similar to the paradigm used in the experiment. The step-up size was 0.25 logMAR for the inner eccentricities and 0.40 logMAR for the outer eccentricities, and the step-down size = $0.8081 \times$ step-up size to target ~55% percent correct of the corresponding underlying PMF. The starting letter size was set to 0.3 logMAR for all eccentricities. The termination rule was set to be after 10 trials and the letter acuity was estimated as the average across the last four reversals. I simulated the results with different lapse rates (λ) ranging from 0 to 15% in 5% steps, and subsequently calculated the mean of differences in letter acuity measurements between the two sessions and SDC using the Bland-Altman analysis for each lapse rate separately. The procedure was repeated 10,000 times and the mean of differences in letter acuity measurements and the SDC values of the 10,000 simulations were obtained. Figures 4.14 & 4.15 show the average and the distribution of the values of the 10,000

simulations at each condition (i.e., different λ) for the inner and outer eccentricities. The results showed that the high lapse rate has a negligible effect on the mean of differences in letter acuity measurements between the two sessions at the inner and outer test locations (Figure 4.14 & 4.15; A, B, C & D). The absolute mean of differences in letter acuity measurements between the two sessions is very unlikely to exceed 0.03 (when $\lambda = 0$), 0.03 (when $\lambda = 0.05$), 0.04 (when $\lambda = 0.10$), or 0.06 (when $\lambda = 0.15$) for the inner test locations and 0.03 (when $\lambda = 0$), 0.05 (when $\lambda = 0.05$), 0.06 (when $\lambda = 0.10$), or 0.09 (when $\lambda = 0.15$) logMAR for the outer test locations ($p < .05$). On the other hand, the results showed that the high lapse rate has a substantial effect on the SDC at the inner and outer test locations (Figure 4.14 & 4.15; E, F, G & H). The random error alone (i.e., when $\lambda = 0$) due to employing the FSS staircase can result in SDC (on average; in logMAR) of 0.33 ± 0.09 for inner eccentricities and 0.33 ± 0.13 for outer eccentricities. It was also very unlikely for letter acuity estimation using the FSS staircase to produce SDC higher than 0.50 for inner eccentricities and higher than 0.55 logMAR for outer eccentricities due to the random error only (i.e., $\lambda = 0$) ($p < .05$). However, in the case of lapsing, on average, the SDC was found to be 0.37 ± 0.10 (when $\lambda = 0.05$), 0.46 ± 0.13 (when $\lambda = 0.10$), and 0.61 ± 0.18 (when $\lambda = 0.15$) for the inner test locations and was found to be 0.48 ± 0.15 (when $\lambda = 0.05$), 0.68 ± 0.21 (when $\lambda = 0.10$), and 0.94 ± 0.28 (when $\lambda = 0.15$) logMAR for the outer test locations. In these cases of lapsing, it was found that the SDC can be as high as 0.54 (when $\lambda = 0.05$), 0.70 (when $\lambda = 0.10$), or 0.94 (when $\lambda = 0.15$) at the inner test locations and 0.74 (when $\lambda = 0.05$), 1.05 (when $\lambda = 0.10$), or 1.46 (when $\lambda = 0.15$) logMAR at the outer test locations (at $p = .05$). Therefore, I can attribute the observed high SDC in the experimental data primarily to the high lapse rate associated with the lack of accuracy in the speech recognition algorithm.

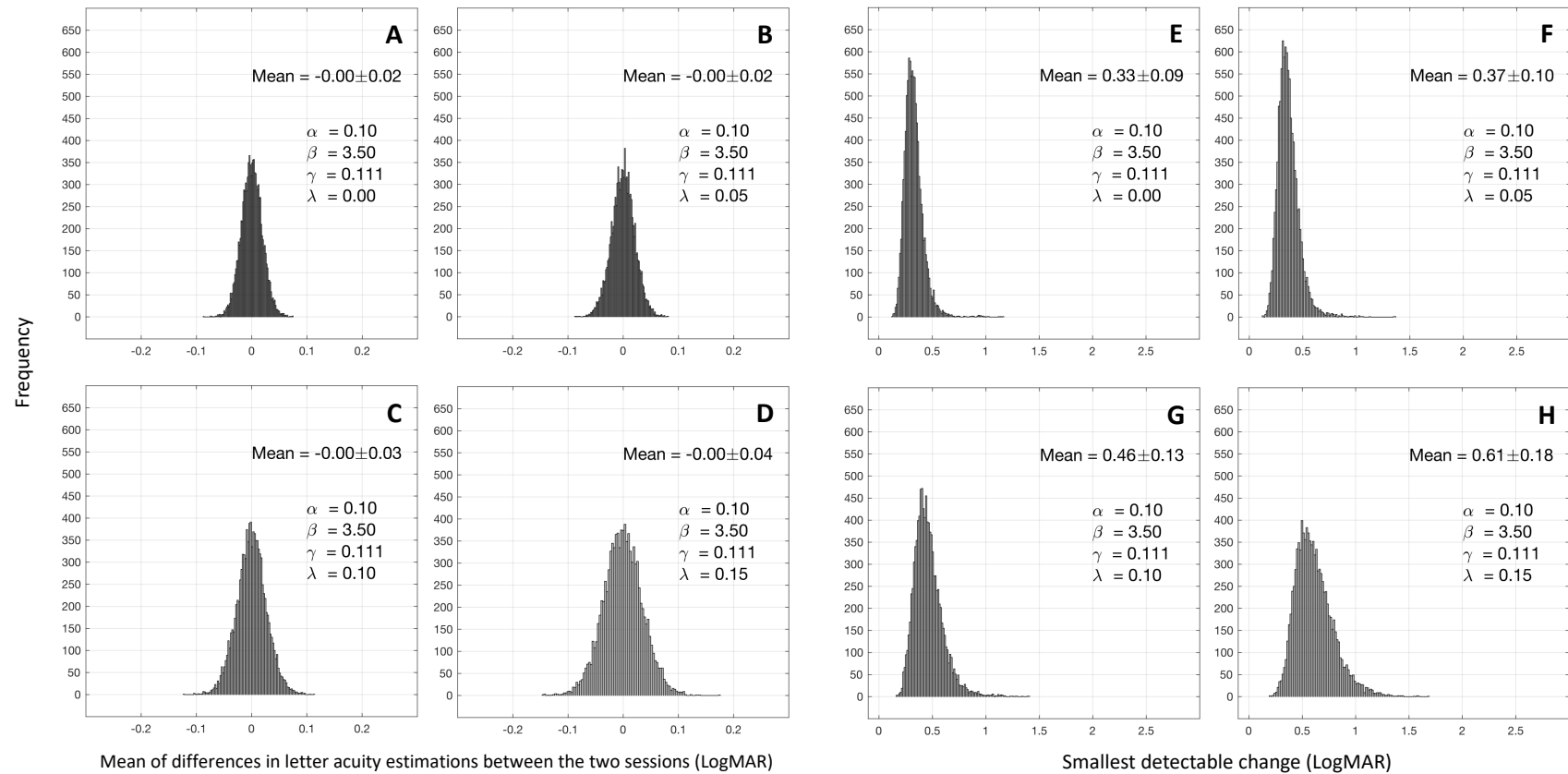


Figure 4.14 The effect of lasing on letter acuity for inner eccentricities.

The Figure shows; for different lasing rates; the mean and distribution of 10,000 simulated mean of differences in letter acuity measurements between the two sessions (A, B, C & D) and the smallest detectable changes (E, F, G & H).

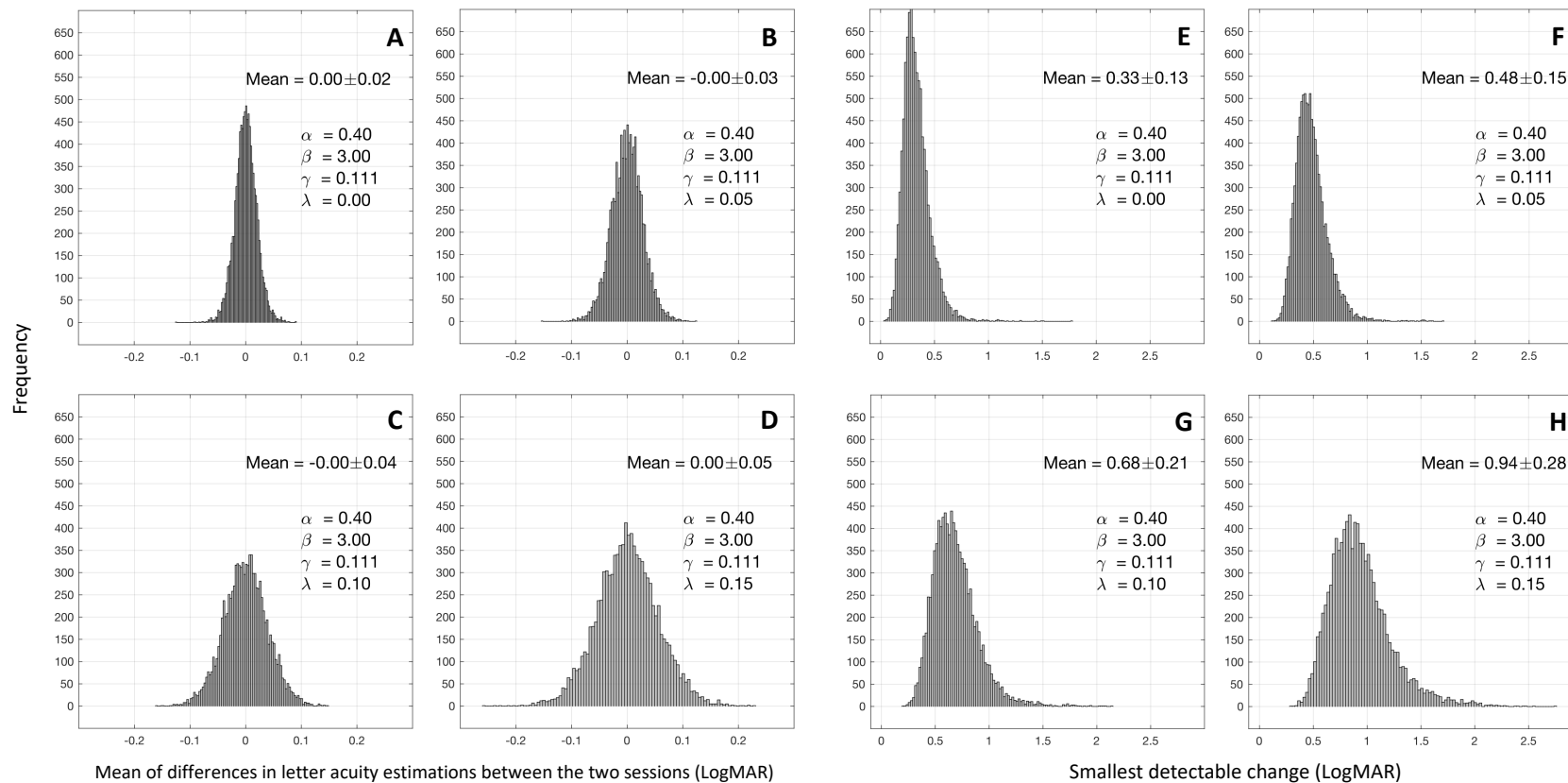


Figure 4.15 The effect of lapsing on letter acuity for outer eccentricities.

The Figure shows; for different lapsing rates; the mean and distribution of 10,000 simulated mean of differences in letter acuity measurements between the two sessions (A, B, C & D) and the smallest detectable changes (E, F, G & H).

In the experimental data, on average, the SDC was found to be relatively higher at the outer eccentricities (1.19 ± 0.40) compared to the inner eccentricities (0.96 ± 0.13) (Figure 4.13). This was also the case for the simulated SDCs where the values were found to be higher at the outer eccentricities compared to the values at the inner eccentricities (Figure 4.14 & 4.15; E, F, G & H). At the outer eccentricities, the observed higher SDC could be due to the performance of the shallow slope (assumed for the simulation) of the underlying PMF (3.00 as opposed to 3.50 at the inner eccentricities) or a result of using a large step-up size (0.40 as opposed to 0.25 at the inner eccentricities). However, the simulation showed that the observed high SDC associated with outer eccentricities is likely because of the larger step-up size (i.e., 0.40) and not because of the shallow slope of the underlying PMF (i.e., 3.00) (Figure 4.16). Further simulations demonstrate that a large step-up size (e.g., 0.50) causes higher SDC where the effect is higher in case of high lapse rates (e.g., $\lambda = 0.15$). All results were consistent in the case of different slopes (4.00, 3.50, 3.00, and 2.5) of the underlying PMF (Figure 4.16).

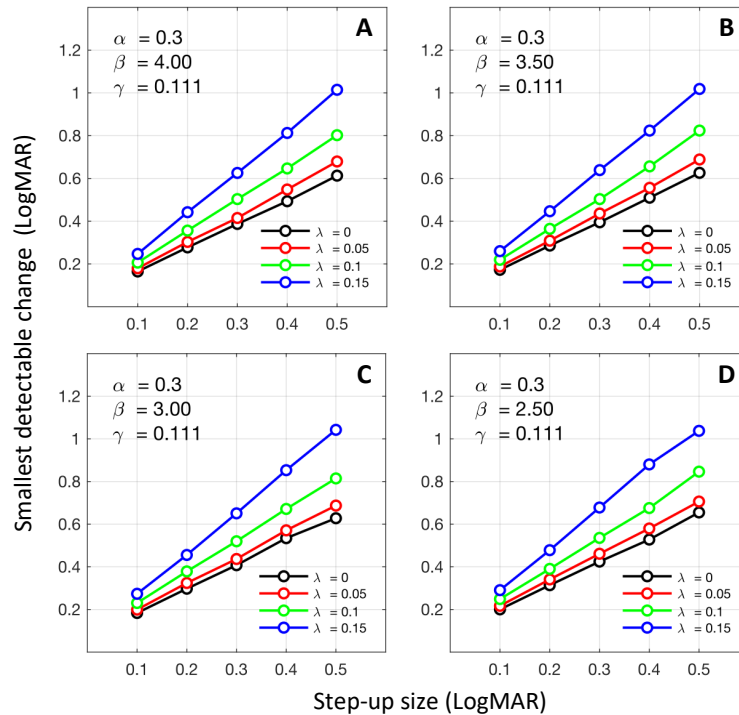


Figure 4.16 The effect of step-up size on the SDC for different lapses.

The Figure shows further simulations to reveal the effect of different step-up sizes (ranged from 0.10 to 0.50 logMAR in 0.10 steps) on the SDC in case of different lapses (ranged from 0 to 0.15 in 0.05 steps) and a fixed ground truth ($\alpha = 0.3$). The results are consistent for different slopes (ranging from 2.5 to 4.00 in 0.50 steps) of the underlying PMF. The curves of the simulated data here and throughout are the average across 500 repetitions for smoother curves.

Furthermore, simulations show that estimating letter acuity that is better than the starting letter size (I used 0.3 logMAR for all the test locations) increases the SDC compared to the case of estimated worse letter acuity than the starting letter size (Figure 4.17; see and compare SDC values of ground truth more than 0.30 vs those of less than 0.30). Additionally, using a larger step-up size would also increase the SDC (Figure 4.17 B). In both cases, the number of well-visible letters would increase, leading to a high possibility for the lapsing to occur. To illustrate, when estimating letter acuities better than the starting letter size at the inner eccentricities (where one would expect to have better letter acuities than 0.30; experimental data showed that this is the case for most inner locations), the staircase would converge from a starting

letter size of 0.30, therefore there would be a large number of well-visible letters presented before converging to the letter acuity. That would increase the likelihood of the lapse occurring (see Figure 4.17 A, especially when the ground truth is smaller than 0.3). Similarly, for the outer eccentricities, when the estimated letters can be equal to or higher than the starting letter size when the staircase converges towards the estimated letter acuity, one would expect that the number of well-visible letters will be small, however because of the large step-up size used at the outer locations, the likelihood of presenting well-visible letters would increase, hence, the likelihood of lapsing would be high as well. This might explain the high SDC associated with all test locations.

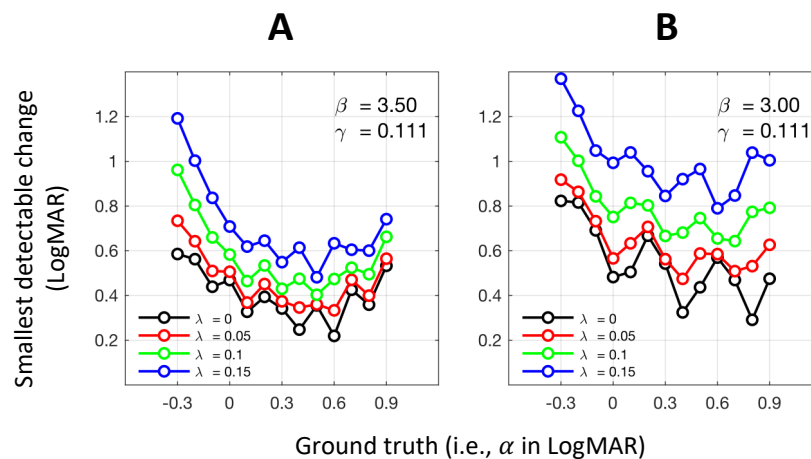


Figure 4.17 The effect of lapsing on SDC for different ground truths.

The Figure shows the effect of lapse rate on the SDC for a fixed starting letter size of 0.30 logMAR and for a range of ground truths (ranged from -0.30 to 0.90 in 0.1 steps) at the inner (A) and outer (B) eccentricities.

Results of simulations for a range of ground truths showed that a high lapse rate can overestimate the letter acuity at the inner and outer eccentricities. Although negligible, the overestimation of letter acuity is more evident, especially at the letter acuity that is smaller than the starting letter size (0.30 logMAR in the current experiment and simulation) by 0.60 logMAR (Figure 4.18). The effect was found to be larger at the

outer eccentricities (Figure 4.18 B). The larger effect observed at the outer eccentricities was found to be a consequence of the larger step-up size (0.4) and not due to the shallower slope (Figure 4.19). However, the overestimation was found to be less than 0.15 logMAR in extreme cases (i.e., $\lambda = 0.15$ and step-up size = 0.5).

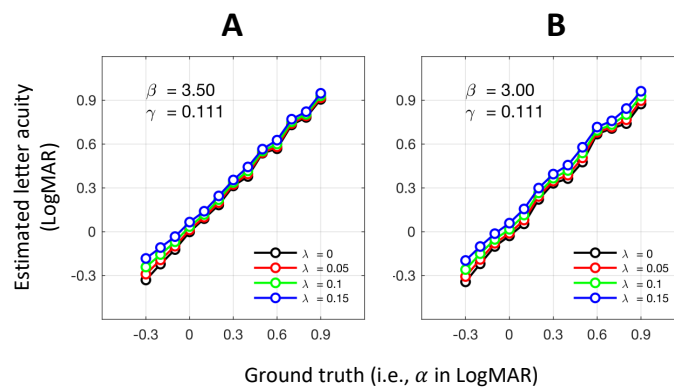


Figure 4.18 The effect of lapsing on letter acuity (by FSS) for different ground truths. The Figure shows the effect of lapse rate on the estimated letter acuity for a fixed starting letter size of 0.30 logMAR and for a range of ground truths (ranged from -0.30 to 0.90 in 0.1 steps) at the inner (A) and outer (B) eccentricities.

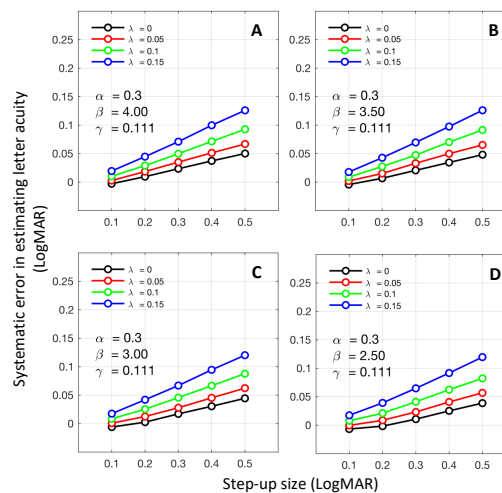


Figure 4.19 The effect of lapsing on letter acuity (by FSS) for different step-up sizes and slopes.

The Figure shows the effect of lapse rate on the estimated letter acuity (expressed as: systematic error = estimated letter acuity – Ground truth (i.e., $\alpha = 0.30$ in the current simulation)) for different step-up sizes (ranger from 0.10 to 0.50 in 0.01 steps) and different slopes of the underlying psychometric function (ranged from 2.5 to 4.00 in 0.50 steps).

A Bayesian approach is known to have less variability (hence, less SDC) compared to FSS and is recommended when the aim is to estimate the threshold with fewer trials in clinical testing (Treutwein, 1995; Phipps et al., 2001), I also investigated (through simulation) the efficiency of employing a Bayesian approach to estimate the letter acuity in case of high lapse rate. More specifically, I investigated the effect of lapse rate on the estimated letter acuity and the SDC on simulated data of 42 sessions (corresponding to two sessions of 21 eyes; similar to the case in the experimental data) using the Zippy Estimation by Sequential Testing, or ZEST method. The ZEST method is a Bayesian adaptive technique. It uses the *mean* of the posterior probability density function to place the stimulus intensity of any trial and to estimate the threshold (King-Smith et al., 1994; Treutwein, 1995; Anderson, 2003).

The underlying PMF of the task remained the same (i.e., Log-Quick function with β of 3.5 for the two inner eccentricities and 3 for the two outer eccentricities. $\gamma = 11.1\%$). For the ZEST method parameters, I assumed that the ZEST likelihood function is a log-Quick function with β_{ZEST} of 3.5 for the two inner eccentricities and 3 for the two outer eccentricities, λ_{ZEST} was set to be 0 and γ_{ZEST} was set to be 11.1%. For all test locations, the prior α_{ZEST} range was set to range from -1.00 to 1.60 in 0.10 (logMAR) steps. In this case, the starting letter size was 0.30 logMAR. The prior probability density function distribution was assumed to follow a uniform distribution. The termination rule was set to be after 10 trials. The letter acuity was determined as the mean of the estimator after 10 trials.

Like the case with the FSS staircase, results show that lapsing can overestimate the letter acuity when estimated using the ZEST method (Figure 4.20 A & B). The amount of overestimation was found to be not more than 0.20 logMAR in extreme cases (i.e., $\lambda = 0.15$ and ground truth = -0.30 ; i.e., less than starting letter size by 0.60 logMAR).

However, re-simulating the results while $\lambda_{\text{ZEST}} = \lambda$, ZEST seems to be very efficient to correct for the overestimation of letter acuity when the exact assumption about the lapse rate in the likelihood function of ZEST (λ_{ZEST}) is made (Figure 4.20 C & D).

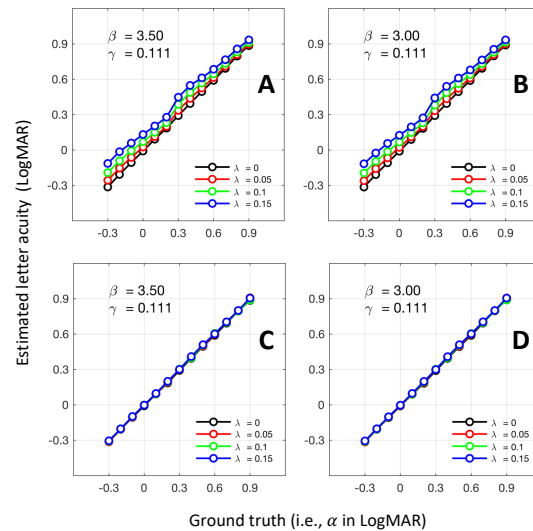


Figure 4.20 The effect of lapsing on letter acuity (by ZEST) for different ground truths. The Figure shows the effect of lapsing on the estimated letter acuity for a fixed starting letter size of 0.30 logMAR and for a range of ground truths (ranged from -0.30 to 0.90 in 0.1 steps) at the inner (left panels) and outer (right panels) eccentricities. Panels A & B show the results estimated by ZEST with $\lambda_{\text{ZEST}} = 0$. Panels C & D show the results when $\lambda_{\text{ZEST}} = \lambda$.

Further simulation showed that SDC estimated via the ZEST method is generally lower than SDC estimated via the FSS staircase when the $\lambda = 0$. Nevertheless, the ZEST method seems to be more sensitive to the presence of lapsing and ZEST produced higher SDC compared to FSS in case of lapsing for the same range of ground truths (see Figure 4.21 A & B and compare to Figure 4.17). Furthermore, when re-simulating the results while $\lambda_{\text{ZEST}} = \lambda$, ZEST was not very efficient to correct for the high SDC when the exact assumption about the lapse rate in the likelihood function of ZEST (λ_{ZEST}) is made (Figure 4.21 C & D).

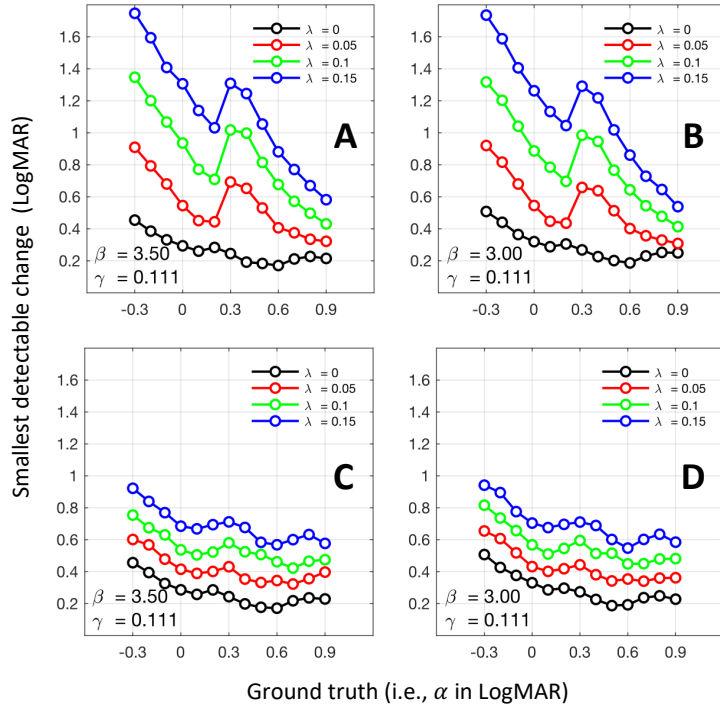


Figure 4.21 The effect of lasing on the SDC (by ZEST) for different ground truths. The Figure shows the effect of lasing on the SDC for a fixed starting letter size of 0.30 logMAR and for a range of ground truths (ranged from -0.30 to 0.90 in 0.1 steps) at the inner (left panels) and outer (right panels) eccentricities. Panels A & B show the results estimated by ZEST with $\lambda_{ZEST} = 0$. Panels C & D show the results when $\lambda_{ZEST} = \lambda$.

ZEST is well known to be robust against inaccuracies in assumptions of β_{ZEST} (Kingdom & Prins, 2010). I investigated the effect of different lapse rates on estimating letter acuity and SDC using ZEST with different β_{ZEST} in the case of different β of the underlying PMF. The results showed that the inaccuracy of the assumption of β_{ZEST} can overestimate the letter acuity by only ~ 0.05 logMAR in the extreme case (i.e., $\lambda = 0.15$, and β_{ZEST} is changed from 2 to 6 for different β of the underlying PMF) (Figure 4.22 A). Additionally, ZEST could estimate letter acuity accurately for different β_{ZEST} but only when assuming $\lambda_{ZEST} = \lambda$ (Figure 4.22 B). On the other hand, I found that the influence of the inaccuracy of the assumption of β_{ZEST} on estimating the SDC can produce a difference of only ~ 0.20 logMAR in the extreme case (i.e., $\lambda = 0.15$, and β_{ZEST} is changed from 2 to 6 for different β of the underlying

PMF) (Figure 4.22 C). However, it seems that ZEST cannot eradicate the influence of lapsing on SDC when assuming $\lambda_{ZEST} = \lambda$ (Figure 4.22 D).

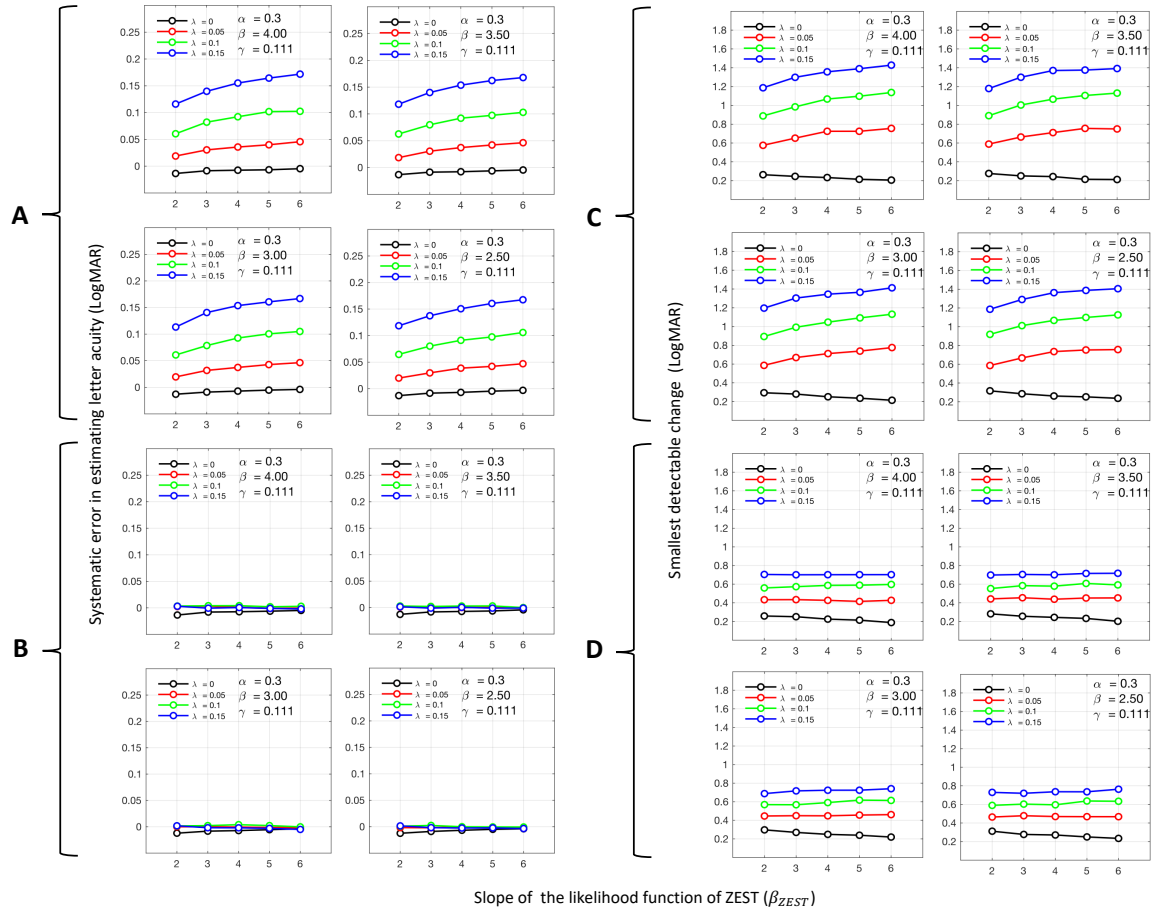


Figure 4.22 The effect of lapsing on letter acuity and SDC (by ZEST) for different conditions. The Figure shows the effect of lapsing on letter acuity estimation (expressed as: systematic error = estimated letter acuity – Ground truth (i.e., $\alpha = 0.30$ in the current simulation)) and SDC via ZEST with different β_{ZEST} ranged from 2 to 6 in 1 steps in case of different β of the underlying PMF (ranged from 2.50 to 4.00 in 0.50 steps). Panels A & C show the results via ZEST when $\lambda_{ZEST} = 0$. Panels B & D show the results via ZEST when $\lambda_{ZEST} = \lambda$.

4.4 Discussion

4.4.1 Letter acuity measurements

The main aim of the current study was to investigate the feasibility of acuity perimetry in healthy and glaucomatous eyes. Results of healthy eyes showed a comparable decrease in letter acuity with increasing eccentricity (Ludvigh, 1941; Hairol et al., 2015) (Figure 4.23).

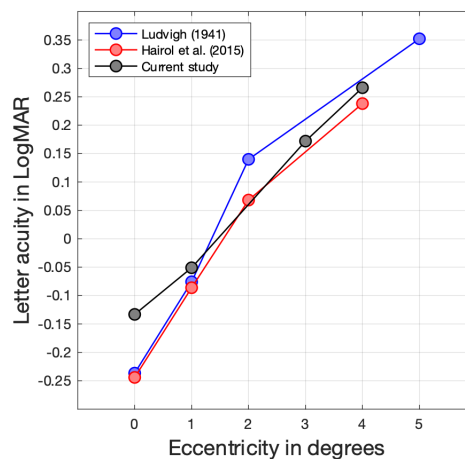


Figure 4.23 Measured letter acuity of the current and two previous studies. The results of the current study are the average across all test locations of each eccentricity for healthy eyes.

In glaucomatous eyes, the letter acuity perimetry was efficient to capture the macular damage confirmed by conventional visual field measurements (i.e., M-pattern) (Figure 4.4, Figure 4.6, Figure 4.7, Figure 4.8). These results demonstrate that the approach is feasible in terms of the visual task (i.e., letter recognition) and the input method (i.e., speech recognition algorithm); as all observers completed the test independently.

4.4.2 Observers' experience of the acuity perimetry

I further investigated the observers' experience of the task via a questionnaire. The questionnaire results show that the novel elements (i.e., letter recognition task and speech recognition algorithm) in the experiment had a low negative impact on the observers' experience without a significant difference between the healthy participants and Glaucoma patients (Figure 4.9, Table 4.1). However, the presentation of very small letters was found to have a more negative effect on the performance of Glaucoma patients compared to healthy observers. This might be because the glaucomatous eyes in the current experiment had manifest macular damage, therefore they would be presented with a larger number of perceived smaller letters compared to those perceived by healthy eyes. Guessing the presented letter instead of saying "not seen" in case of uncertainty negatively affected the performance of all observers without a significant difference between healthy and Glaucoma observers. For the current experiment, guessing instead of saying "not seen" in case of uncertainty was necessary for the analysis of the number of adopted letters (guessing rate) during the experiment for each observer (Figure 4.11).

4.4.3 Performance of the speech recognition algorithm

In the current experiment, I employed a speech recognition algorithm to capture and enter the responses (Selvaraaju, 2022). The algorithm was subject-dependent; therefore, a training stage was needed prior to the experiment for each observer. I set the accuracy tolerance to be 90% (it was 80% in two cases) and the letters that were validated and achieved the accuracy tolerance criterion were chosen for the experiment

(i.e., acuity perimetry). For each observer, only the validated letters were presented as stimuli in the experiment. The number of presented letters did not substantially affect the number of adopted alternatives (responded letters) during the experiment (Figure 4.11). Hence, our assumption about the guessing rate remained valid.

4.4.4 The rationale for employing the FSS staircase in the current study

The FSS staircase was employed in the current experiment and the parameters of the FSS were efficient to produce enough reversals within 10 trials of each staircase (Figure 4.10). The reason for choosing the FSS method was to present observers with letters of different sizes by setting large but within recommended step-up size (larger than half but smaller than the spread of the underlying PMF) (Garcia-Pérez, 1998). Nevertheless, the step-up size at the fovea and 1 deg (i.e., 0.25 logMAR) was larger than the recommended value as our main aim was to map the macular *damage* in glaucomatous eyes, therefore, using a smaller step-up size (e.g., 0.1 logMAR) might fail to produce enough reversals within 10 trials in case of poor letter acuity associated with macular damage (e.g., 1 logMAR). For the outer eccentricities, the recommended step-up size (i.e., 0.4 logMAR) was large enough to converge for poor letter acuities in case of macular damage. Additionally, the presentation of large letter sizes (due to the large step-up size) will act as a confidence enhancer and as a reminder of the letters set for the observers (Kingdom & Prins, 2010). Furthermore, it enabled me to assess the lapse rate that might be associated with the failure of the speech recognition algorithm. In the current experiment, the starting letter size was 0.30 logMAR for all test locations. The reason for using a fixed starting letter size for all the test locations was to examine the ability of the paradigm to converge towards the anticipated letter

acuity departing from the same letter size.

4.4.5 Sources of test-retest variability in the current study

I collected data from all observers twice and the agreement between the measurements of the two sessions at each test location was examined (Figure 4.12). Bland-Altman analysis showed that the mean of differences in measurements between the two sessions at each test location was negligible. However, the SDCs (i.e., 95% limits of agreement) were found to be substantially high (i.e., high test-retest variability) for all test locations (Figure 4.13).

Many factors can increase the test-retest variability in visual acuity measurement such as the optotypes used, the degree of optical defocus, the presentation time, the scoring and the termination rule used (Vanden & Wall, 1997; Siderov & Tiu, 1999; Rosser et al., 2003; Rosser et al., 2004; Shah et al., 2011).

4.4.5.1 Lapse rate as a primary cause of increased test-retest variability

In estimating letter acuity in the current experiment, similar factors might have contributed to the high test-retest variability. However, further analysis of experimental data showed that the lapse rate was high (i.e., ~11%) (Table 4.2). A high lapse rate is known to cause an over-estimation of the threshold and increase test-retest variability (Swanson & Birch, 1992; Wichmann & Hill, 2001). Simulations demonstrated that such high lapse rate could be the primary cause of the high SDC observed in the experimental data (Figures 4.14 & 4.15). In the FSS staircase method, using a large step-up size or estimating better letter acuity compared to the starting

letter size aggravates the effect of lapsing on SDC (Figure 4.16 & 4.17). The reason is simply that in these two cases, the number of well-visible letters increases and therefore the occurrence of failure to respond to those well-visible letters becomes high. Consequently, the variability (hence the SDC) in estimating the letter acuity increases.

A Bayesian approach (i.e., ZEST) was examined as an alternative to the FSS staircase. Simulations showed that estimating letter acuity using ZEST led to lower SDC compared to FSS in case of no lapses ($\lambda = 0$). However, in the case of high lapses, ZEST showed higher SDC compared to FSS (i.e., ZEST is more sensitive to the presence of lapses) (Figure 4.17 & 4.21 A & B). Additional simulations showed that even when corrected for the exact lapse rate, ZEST failed to correct completely for the increase in SDC due to the high lapsing (Figure 4.21 C & D).

As expected, further simulations showed that a high lapse rate causes overestimation of the letter acuity (although negligible) in the case of estimating the letter acuity either using the FSS or ZEST (Figures 4.18 & 4.20 A & B). Furthermore, simulations showed that when corrected for the exact lapse rate, ZEST completely corrected the overestimation of the letter acuity caused by the high lapse rate (Figure 4.20 C & D).

It has been argued that the lapse rate associated with letter acuity measurement (with the possibility to amend responses) is negligible (i.e., 0.0005) (Arditi, 2006).

In the current experiment, I employed a speech recognition algorithm that captures the observers' verbal responses. The paradigm has a response window (1500 milliseconds) within which the responses should be given. Failure to respond within the response window will be recorded as a random response. Additionally, once the response was given, amendment of accidental errors was not possible. These considerations along with the accuracy tolerance of 90% of the speech algorithm increased the possibility

of lapsing in the current experiment. In this case, even with an accuracy of 100% of the speech recognition algorithm, lapsing would be unavoidable if the observers have only one chance to respond without the possibility of correction. Therefore, the optimisation of the speech recognition algorithm should be sought not only to improve the accuracy tolerance (e.g., more than 95%) but also the optimisation should allow for correction in case of accidental errors. In this case, adopting the ZEST method while assuming a lapse rate of 0.05 (or less) for instance, would improve the SDC and it would not substantially alter the performance in case less than the assumed lapse occurred. The risk of assuming a larger lapse rate (e.g., 0.1 or 0.15) is not only that it would not improve SDC significantly but also it would alter the performance (more than 0.1 logMAR change) in case of lower lapse (i.e., less than 0.05) (Figure 4.24).

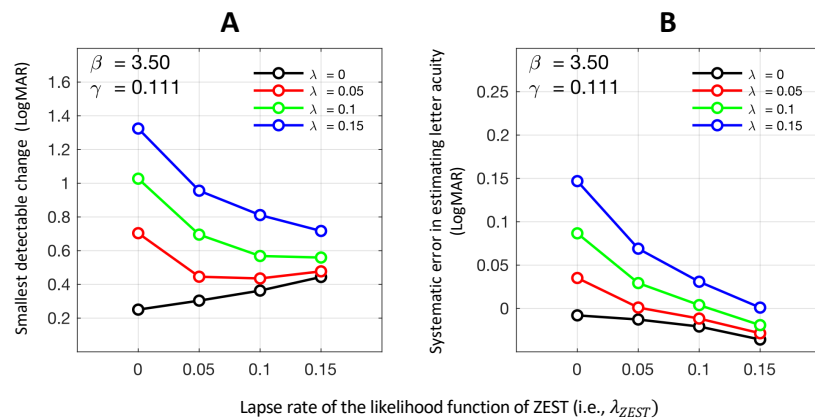


Figure 4.24 The effect of lapsing on letter acuity and SDC (by ZEST) for different λ_{ZEST} . The Figure shows the effect of assumed lapse rate of the likelihood function of ZEST (λ_{ZEST}) on the smallest detectable change (A) and systematic error in estimating the letter acuity (B) for different occurred lapse rates (λ).

4.4.5.2 Other potential sources of increased test-retest variability

In the current experiment, I instructed the observers to fixate on a fixation cross, but I did not monitor the fixation compliance objectively during the experiment. The fixation loss might be an additional source of the observed high SDC in the current experiment (Demirel & Vingrys, 1994; Ishiyama et al., 2014). I tried to minimise the negative effect by reminding the observers frequently (at least two times for each observer). Furthermore, I used a small fixation cross to enhance fixation (Hirasawa et al., 2016). Although it has been shown that loss of fixation has no preference towards the scotoma (Kasten et al., 2006), further investigation is required to confirm that this is the case when mapping the macular damage using acuity perimetry.

In the current experiment, I adopted nine letters as optotypes assuming that they are dissimilar. The similarity between the nine letters was assessed based on the correlation between the undistorted letters (Fülep et al., 2017). However, in the case of presenting letters on an area of macular damage, this might cause the letters to be perceived as distorted and therefore different patterns of the similarity between the adopted letters might be perceived by the glaucomatous eye. In this case, the similarity between letters will be an additional factor that increases the SDC (Grimm et al., 1994; McMonnies & Ho, 1996; McMonnies & Ho, 2000).

4.4.6 Conclusion

To conclude, the current experiment showed that acuity perimetry using letters as optotypes in conjunction with a speech recognition algorithm as an input method is feasible and provides a promising solution for future home-based visual field test.

Future work should aim to optimise the speech recognition algorithm to provide higher accuracy to lower the lapse rate, allow correction for accidental errors, and be subject-independent to be able to include more letters. A Bayesian approach should be employed to further increase the precision of measurements. Additionally, the fixation pattern and the similarity pattern between letters should be quantified and investigated as additional factors of increasing the SDC during mapping the macular damage using the acuity perimeter.

Chapter 5 . Summary

5.1 General discussion

The ultimate aim of this PhD thesis was to explore the possibility and feasibility of using letter acuity-based tests for the potential application in a home-based test for monitoring macular damage. I considered the following for the approach: the task should be sampling-limited (i.e., the task is limited by the neural sampling of the retinal image at the periphery) to strongly reflect the structural damage in the macula and the test should employ an intuitive task. I found a letter identification task fulfils these criteria. I adopted Sloan letters as optotypes and I started by investigating the confounding factors in the Sloan letters identification task (i.e., response bias, sensitivity differences and similarity between letters). I employed Luce's choice model (1963) and our newly developed NTM (Georgeson et al., 2023) to investigate these confounding factors. Luce's choice model uses the maximum likelihood estimates of responses in the presented vs responded letter confusion matrices of the experimental data to compute the response biases and similarity (Chapter 2). However, apart from being not a generative model, Luce's choice model does not estimate the sensitivity differences between letters. Additionally, similarity estimation using Luce's choice model might be affected by the presence of biases (Chapter 3; section 3.4.3).

Therefore, I introduced an original extension to our previously devised the NTM model (Georgeson et al., 2023). Our original model was constructed to quantify the biases and sensitivity differences in the perception of individual letters in letter identification tasks. In Chapter 3, I presented the new version of the NTM model where I incorporated the letter similarity into the model. The new version of the NTM model enabled me to estimate the response bias, sensitivity differences, and similarity between letters simultaneously. The generative nature of the NTM facilitates the investigation of the role of response biases, sensitivity differences, and similarities between letters in the letter identification task. The simulations presented in Chapter 3 showed that letter similarity is the major source of test-retest variability. Therefore, it is of high importance to minimise the similarity between the adopted letters as stimuli in acuity perimetry. I examined the feasibility of acuity perimeter to map the macular sensitivity in healthy observers and Glaucoma patients. I employed dissimilar letters as stimuli and implemented a speech recognition algorithm as an input method. This is paramount to be able to use many letters as stimuli in any potential subject-friendly home-based test. The initial results obtained from glaucomatous eyes were promising (Chapter 4). Nevertheless, the test-retest analysis showed that the test has poor sensitivity to monitor the glaucomatous macular damage (i.e., high smallest detectable change in macular sensitivity). Extensive simulations showed that the primary cause of the poor sensitivity is likely due to the high lapse rate associated with using the current speech recognition algorithm. However, the possibility to have different letter similarity patterns associated with macular damage might also have contributed significantly. Another factor that might have contributed to poor sensitivity is poor fixation compliance during the test.

5.2 General conclusion and future work

The current project showed that acuity perimetry using letters as optotypes in conjunction with a speech recognition as an input method are feasible methods that might be useful for the development of a home-based flagging tool for monitoring macular damage. The proposed acuity perimetry approach is promising especially with the increase in the need to deliver the monitoring services at home where possible (e.g., during COVID for instance). The paradigm can be optimised by further investigating the adopted letters and design to ensure that they are highly dissimilar and have consistent performance for different macular sensitivities.

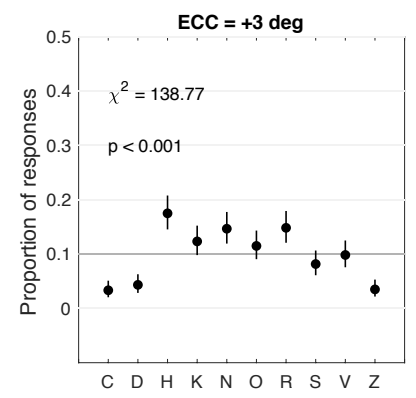
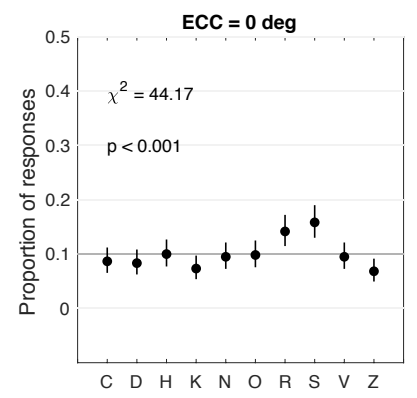
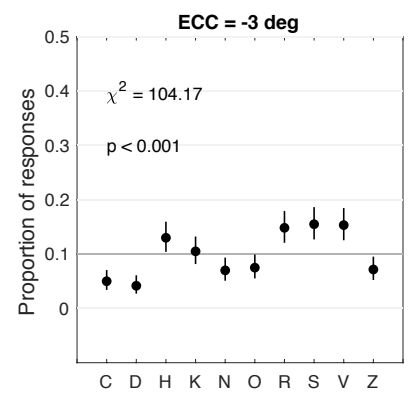
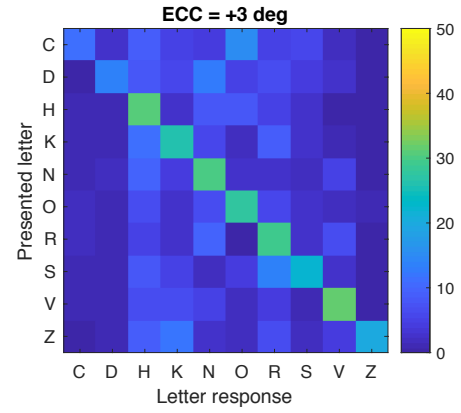
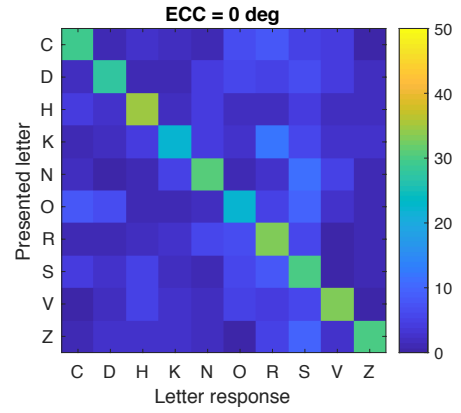
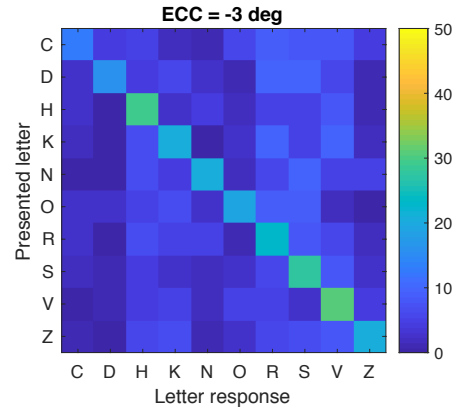
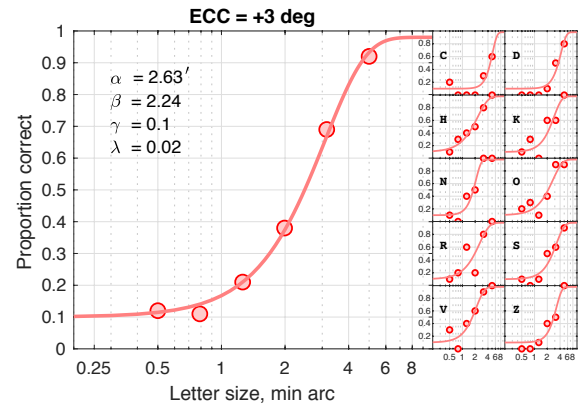
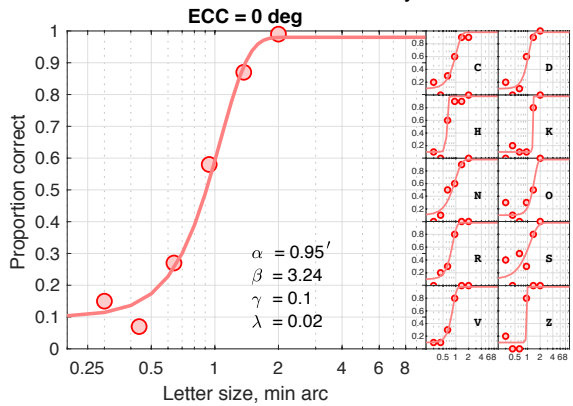
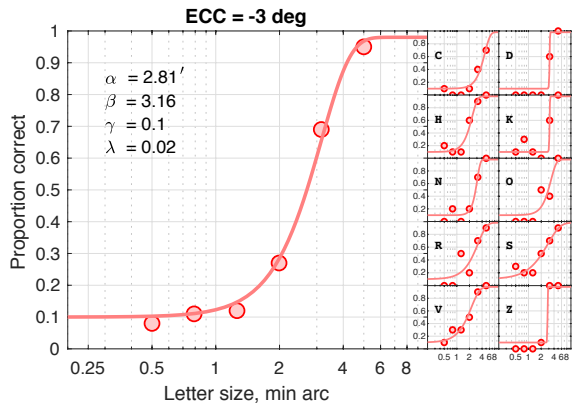
Significant progress has been made in speech recognition technology through the application of deep learning, neural networks, and the utilisation of extensive datasets. These advancements have resulted in improved accuracy for transcribing speech and understanding natural language (Deng et al., 2013; LeCun et al., 2015; Nassif et al., 2019). Future work should aim to optimise the speech recognition algorithm in acuity perimetry.

Furthermore, an optimised version of the acuity perimetry should be tested on a larger scale at home settings to reflect the feasibility of the approach as an efficient home-based visual field test.

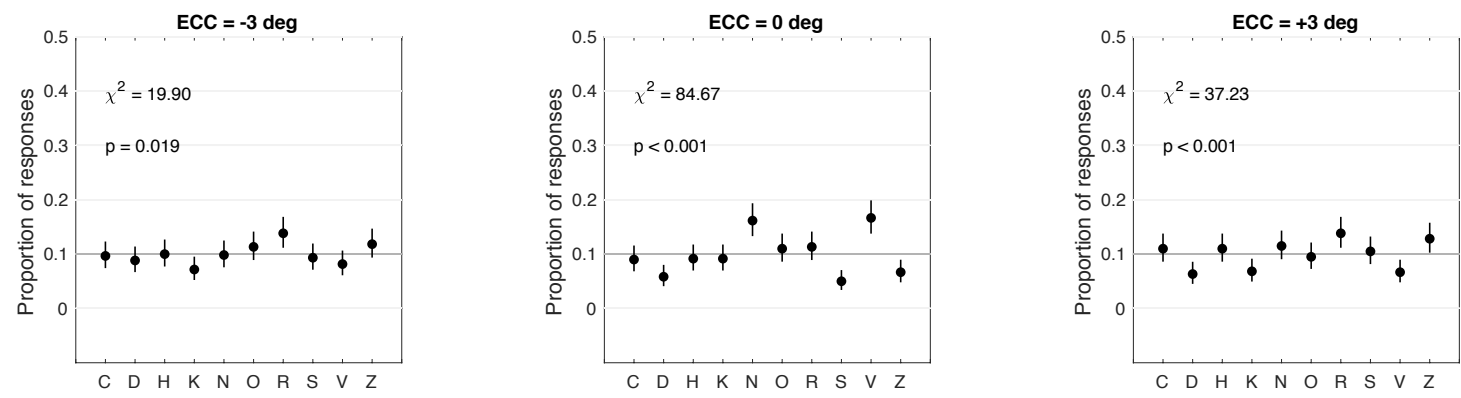
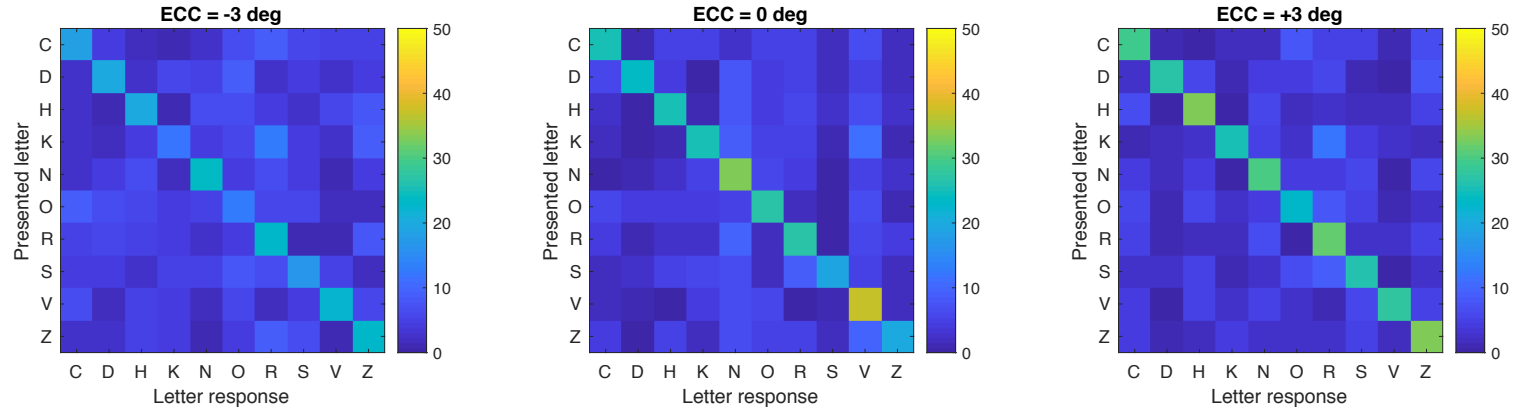
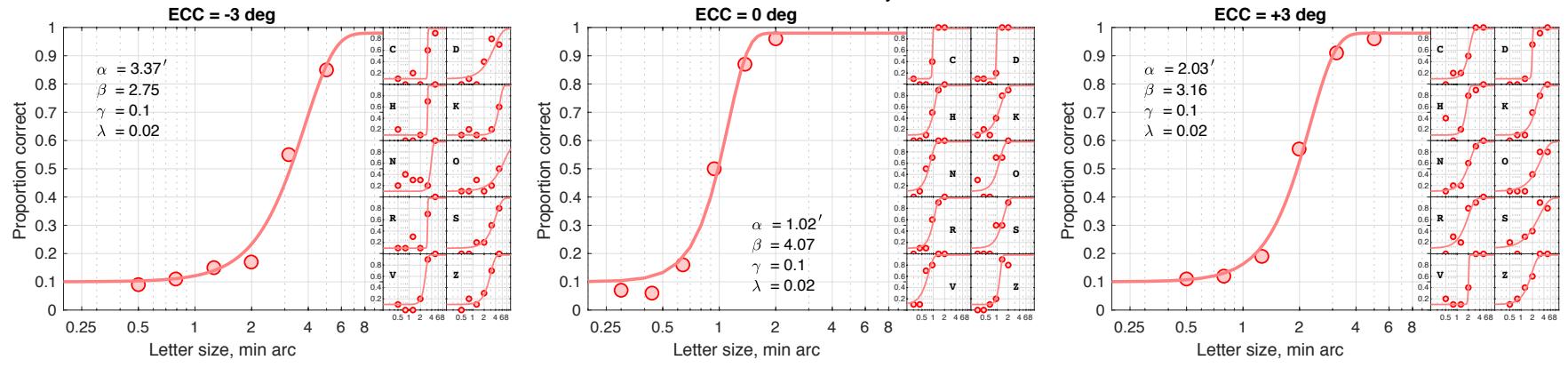
Appendix A

Appendix A shows, for each observer, the psychometric functions of pooled and individual letters (top row), presented vs responded letter confusion matrices collapsed across the letter sizes (middle row), and the proportion of total usage of individual letters at three locations separately (bottom row). These data were used in analysis of Chapter 2 & 3. Each page compiled data for one observer as indicated by their name code.

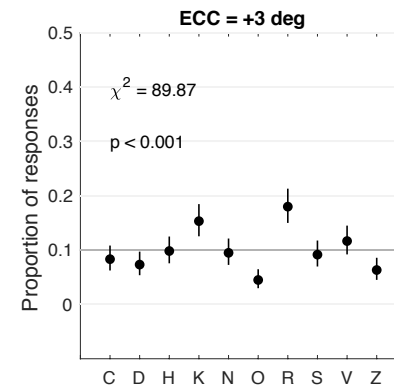
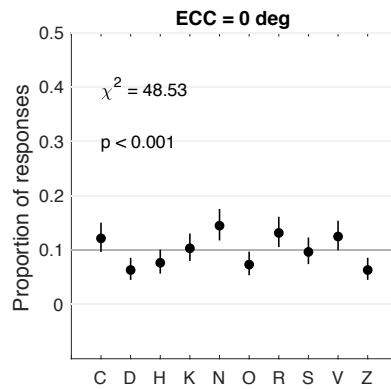
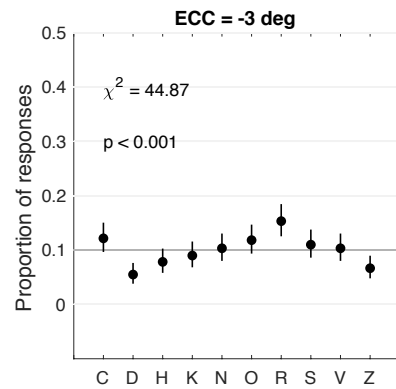
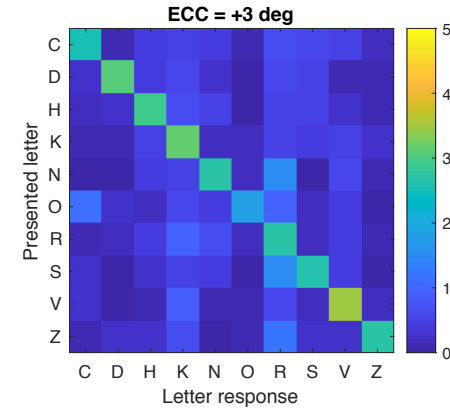
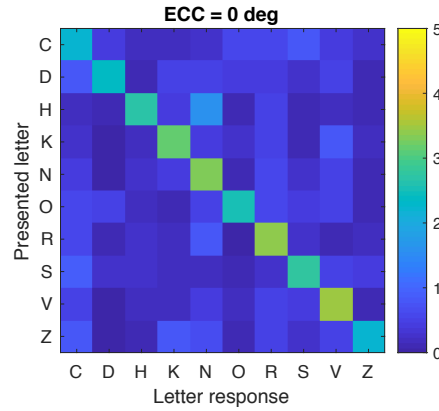
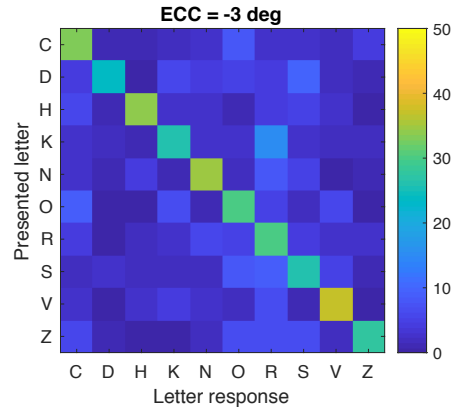
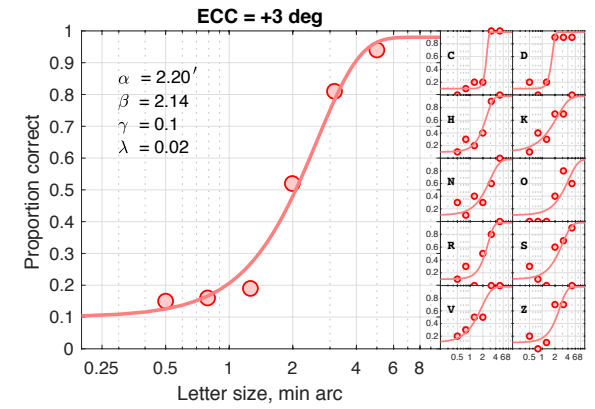
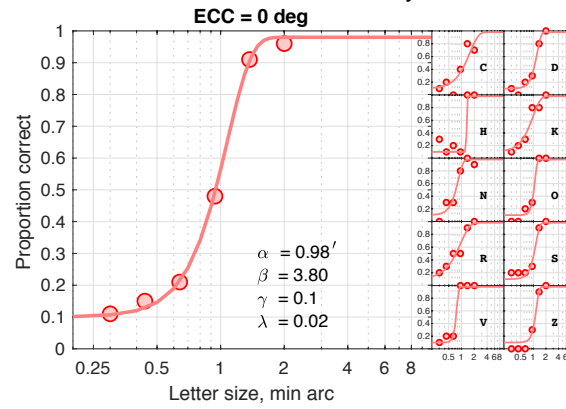
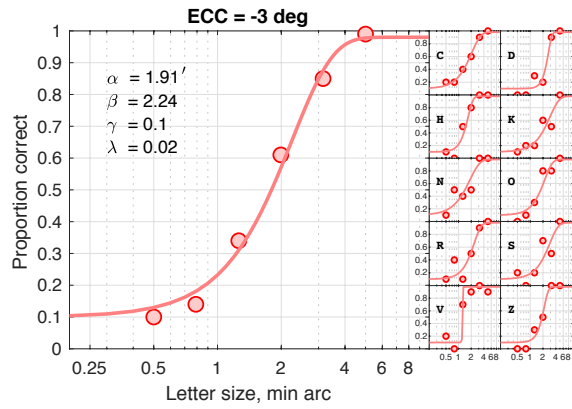
Subject.AB



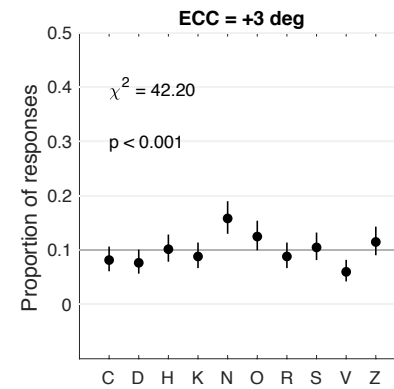
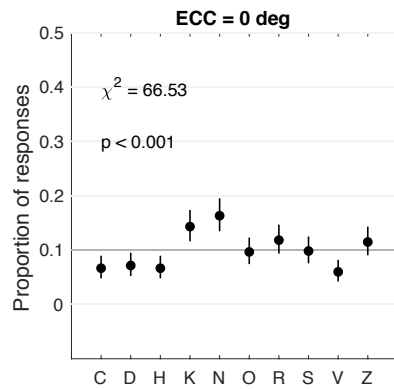
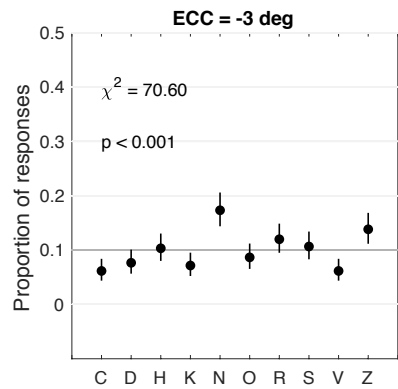
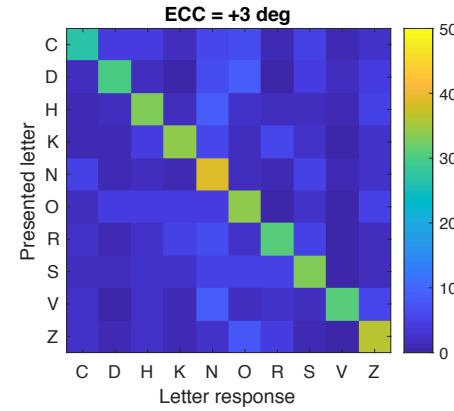
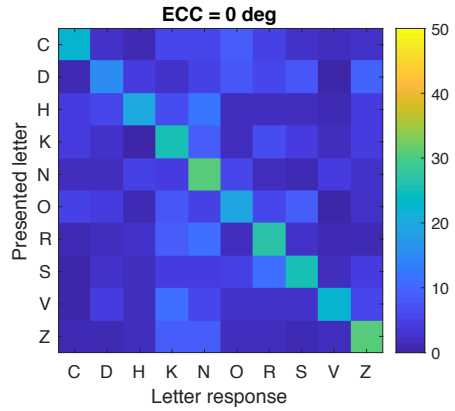
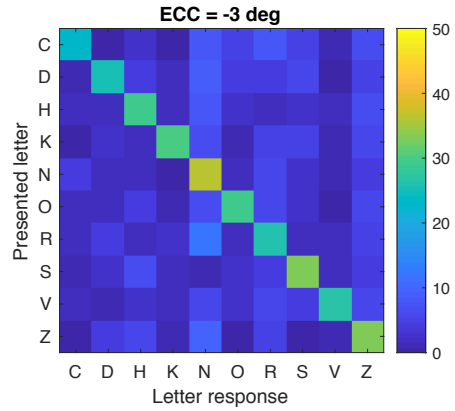
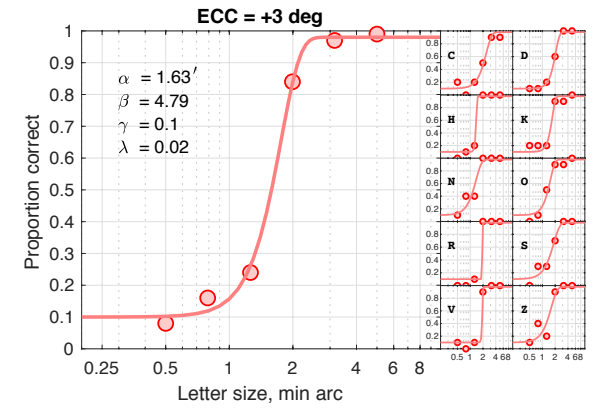
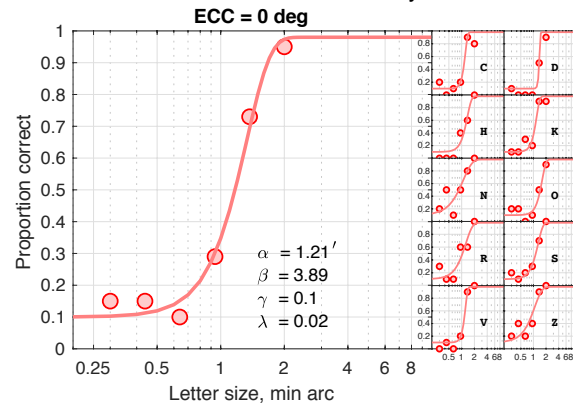
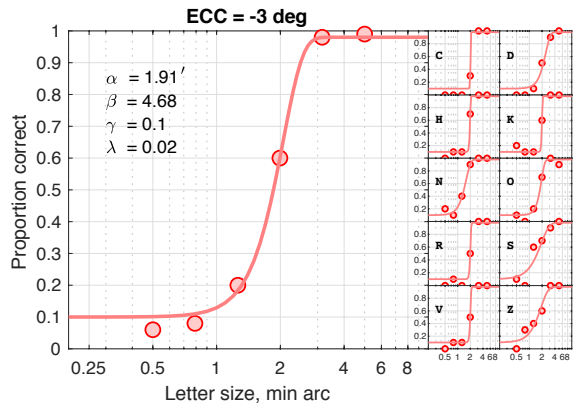
Subject.AM



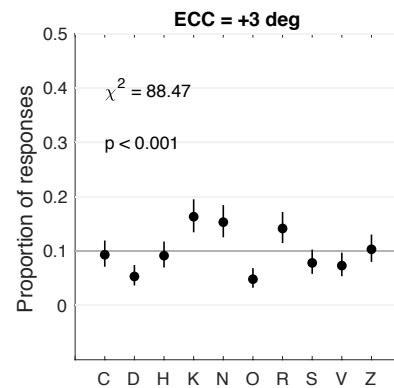
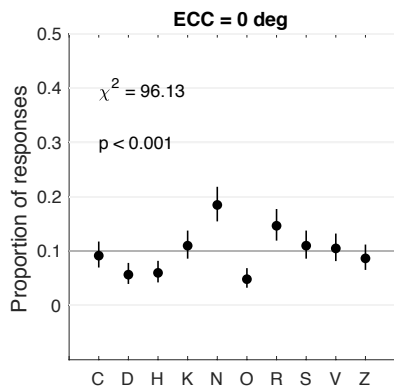
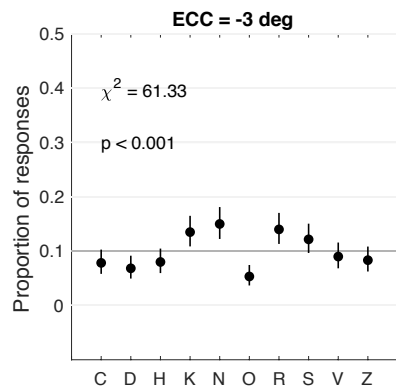
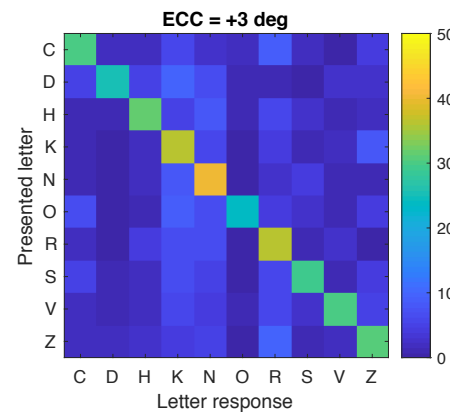
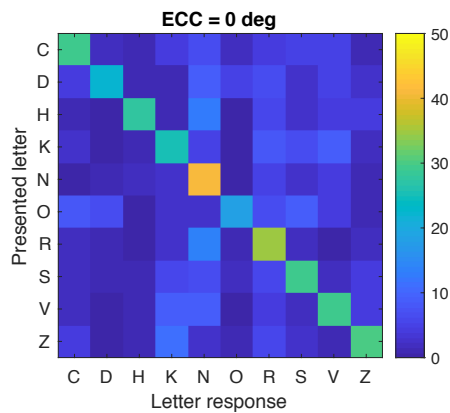
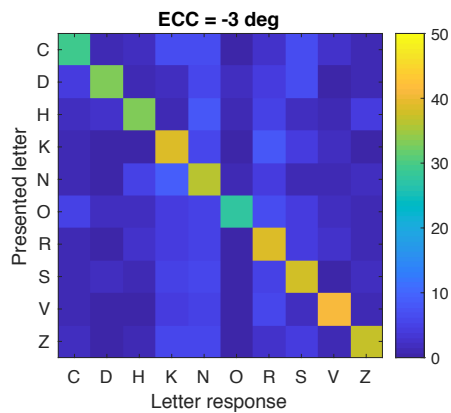
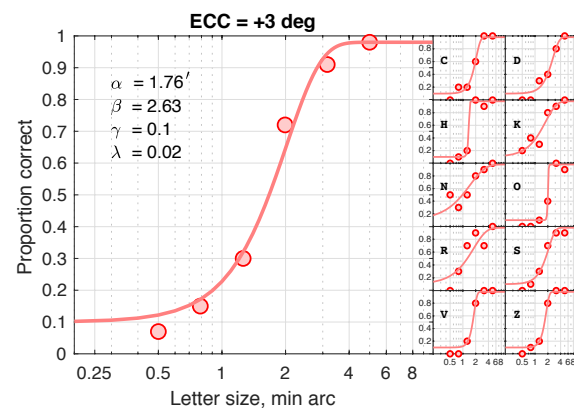
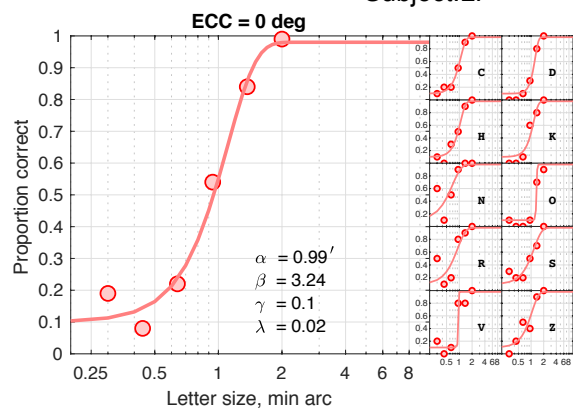
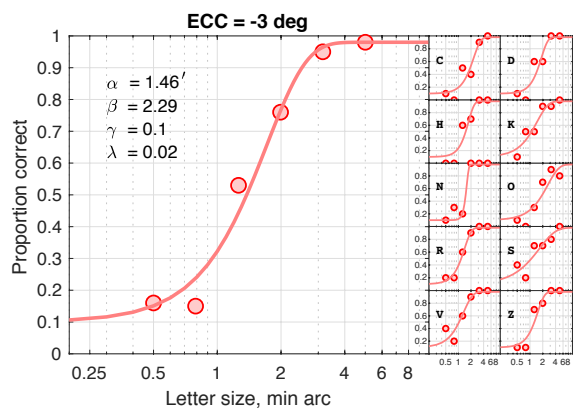
Subject.CW



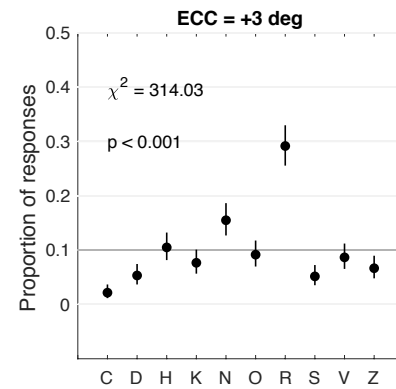
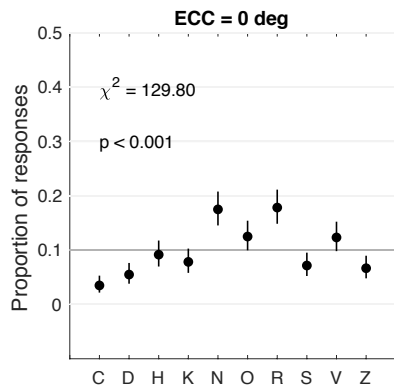
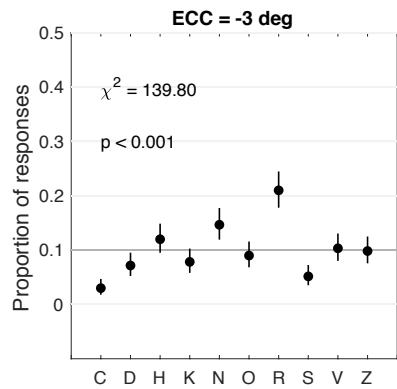
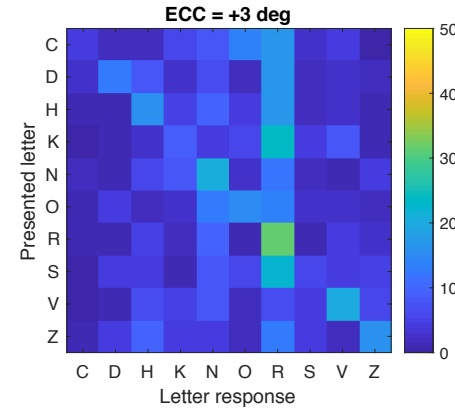
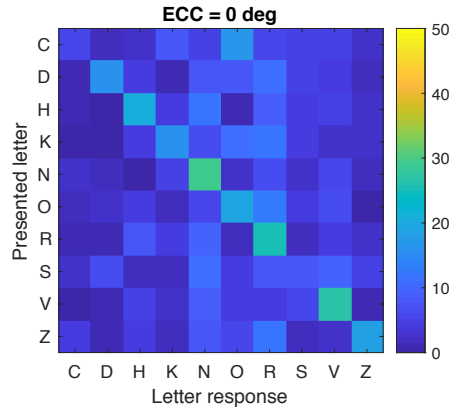
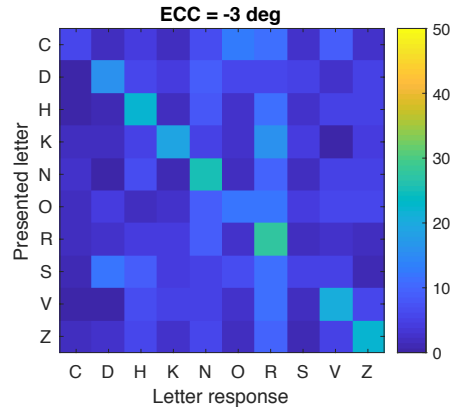
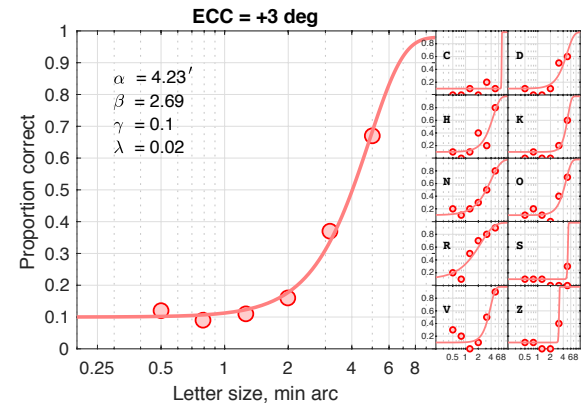
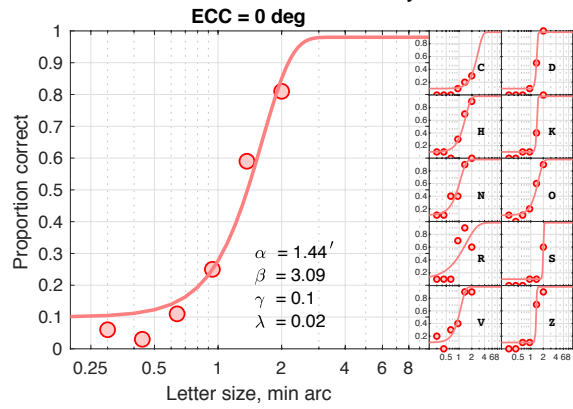
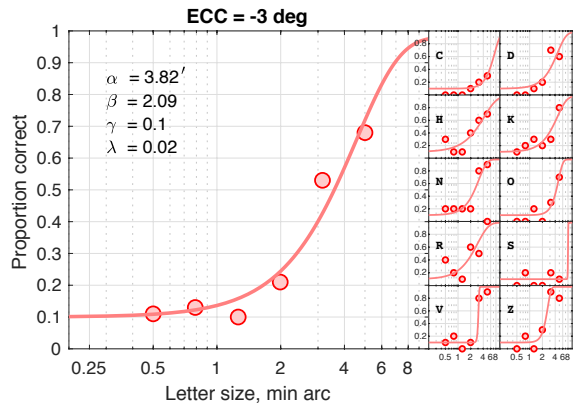
Subject:EM



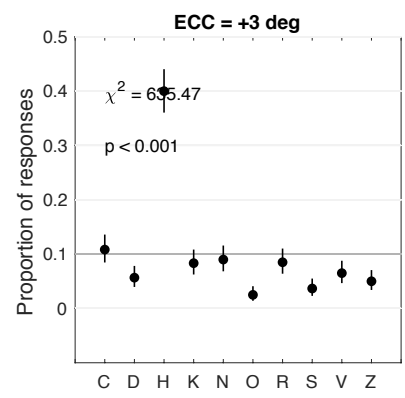
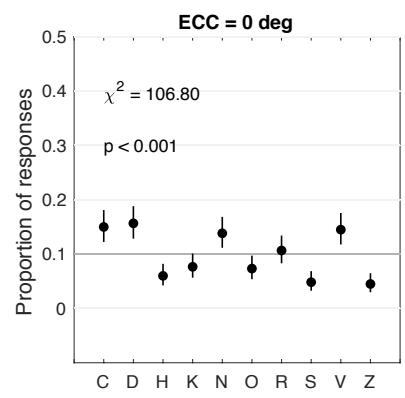
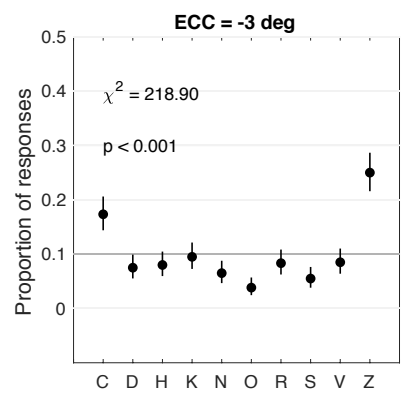
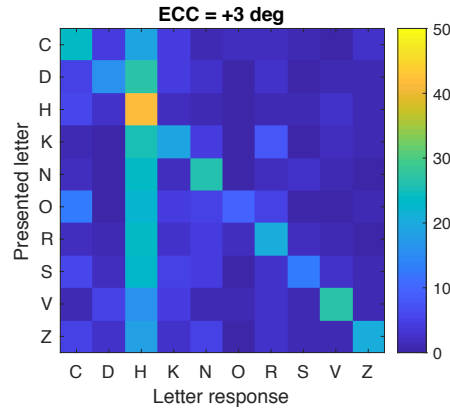
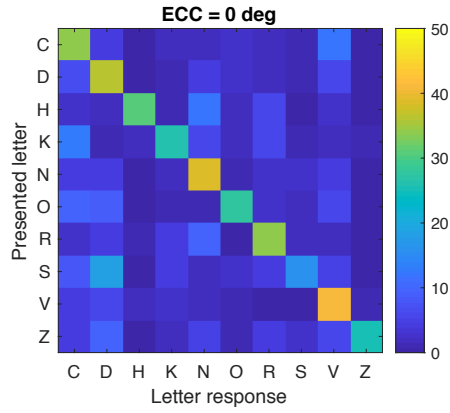
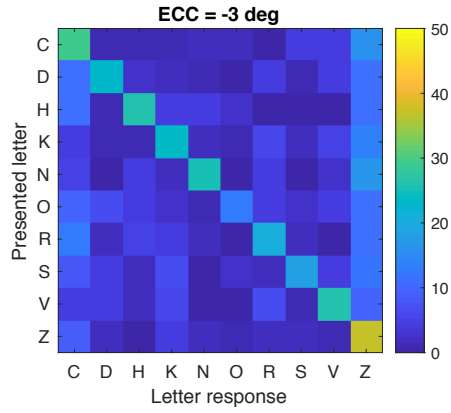
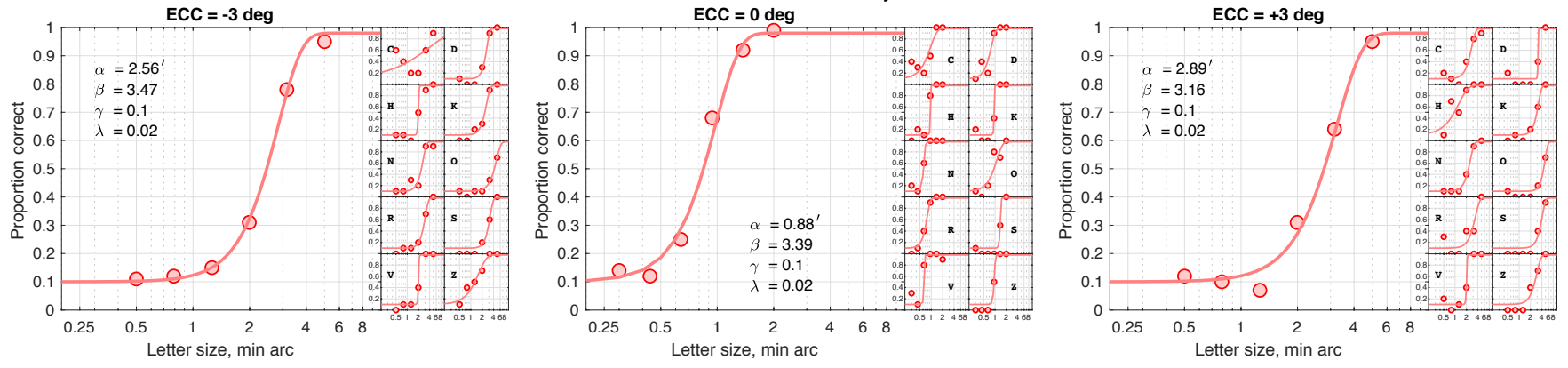
Subject.LP



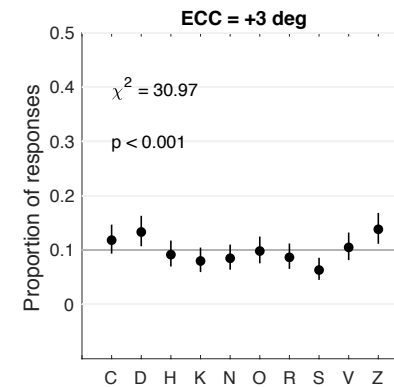
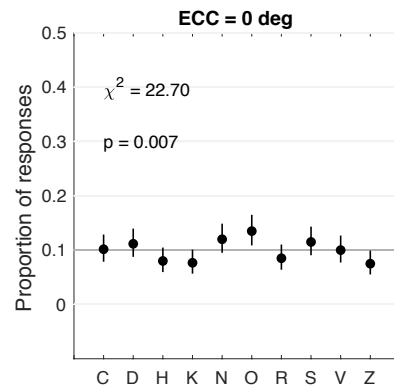
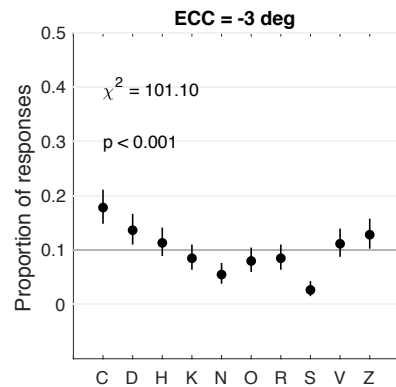
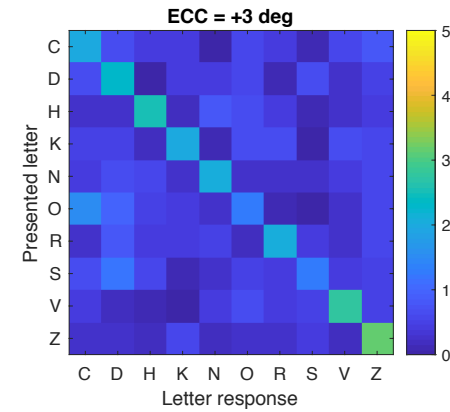
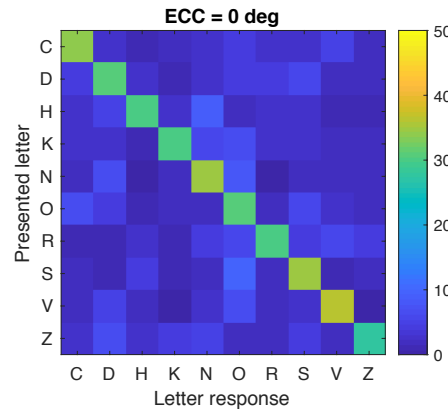
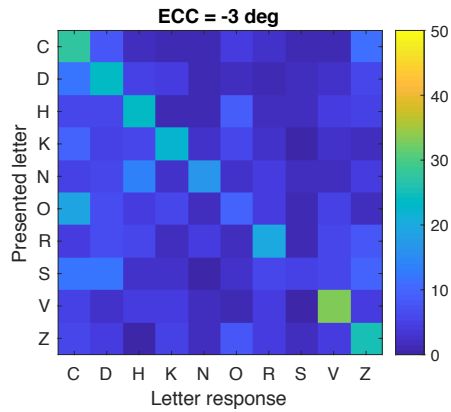
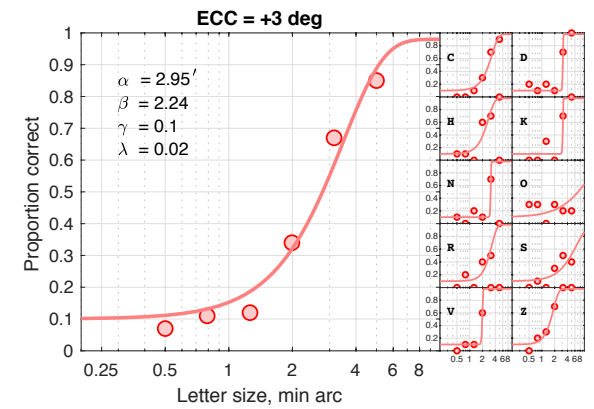
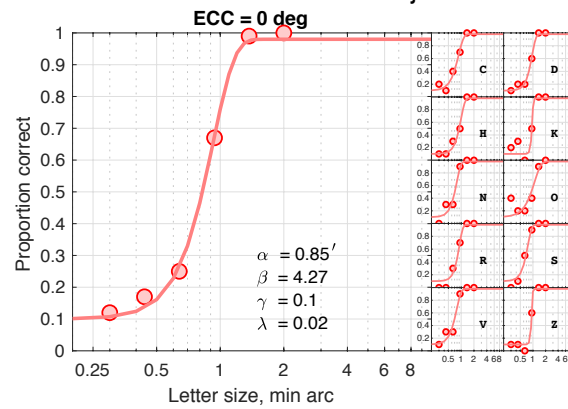
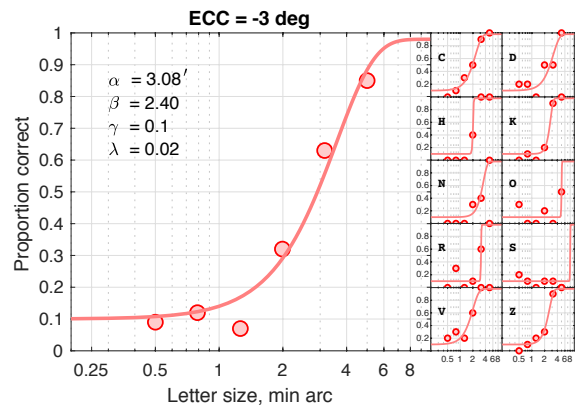
Subject.MT



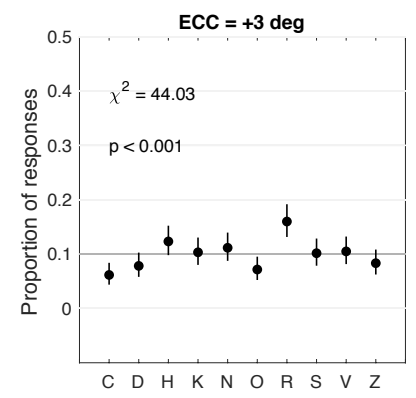
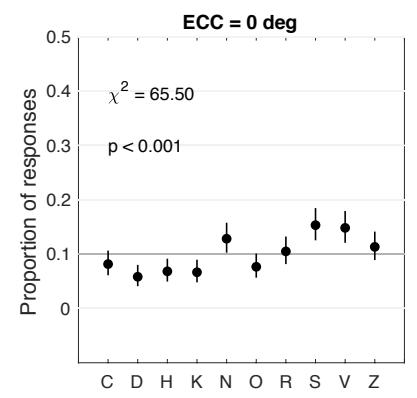
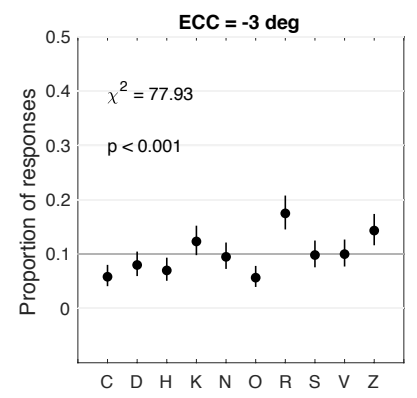
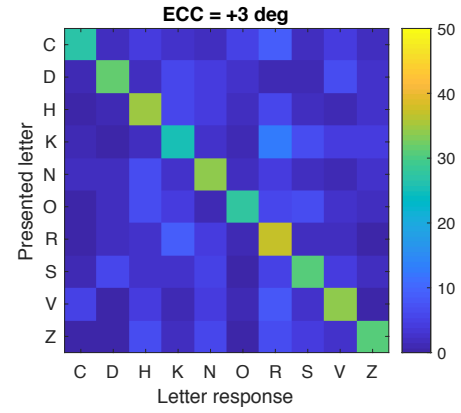
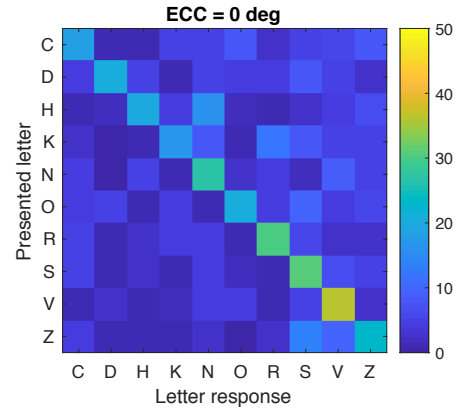
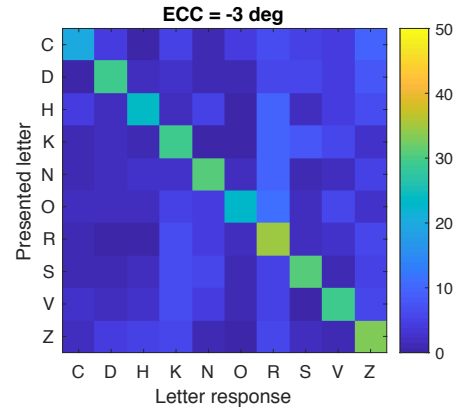
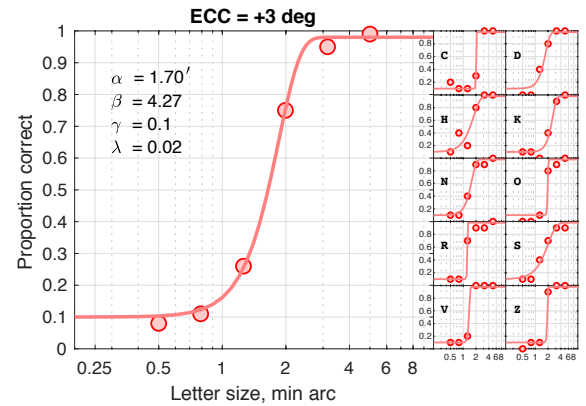
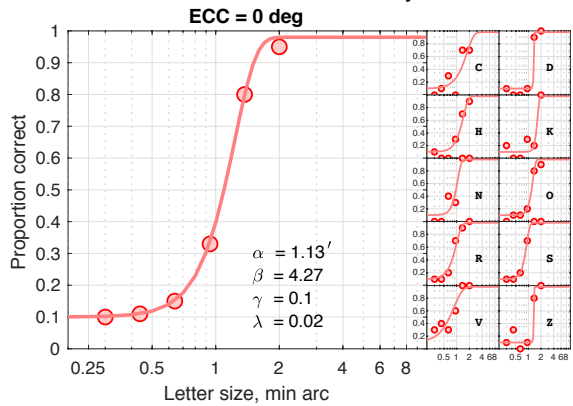
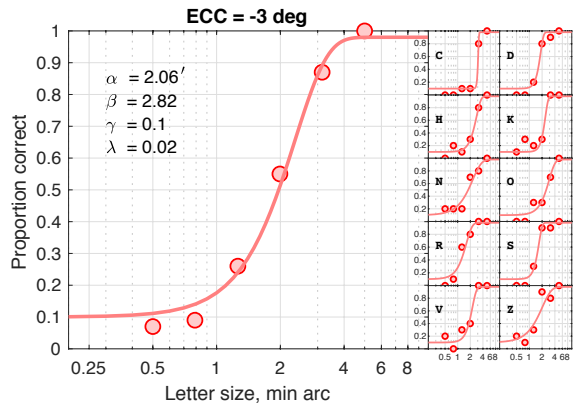
Subject.OO



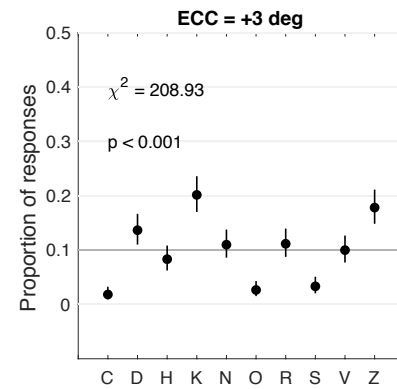
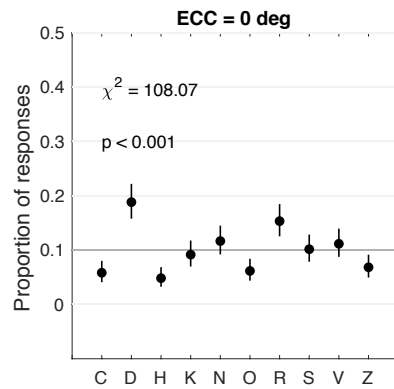
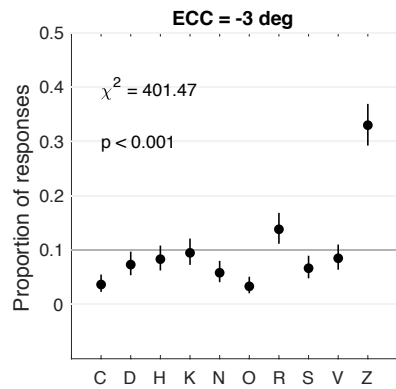
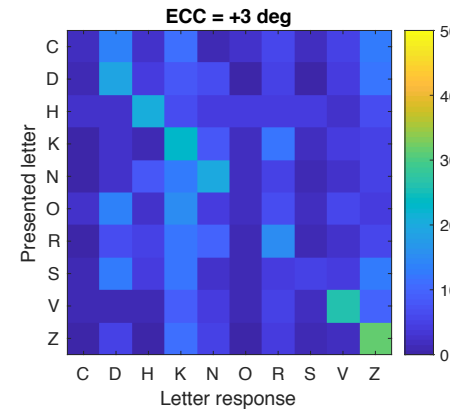
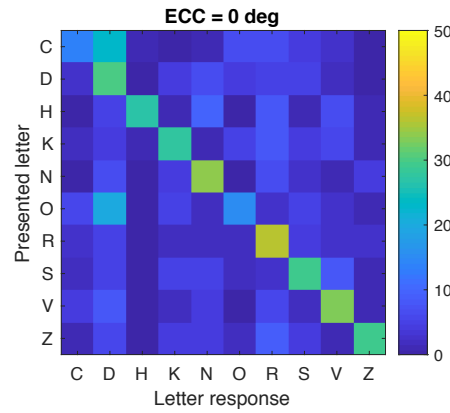
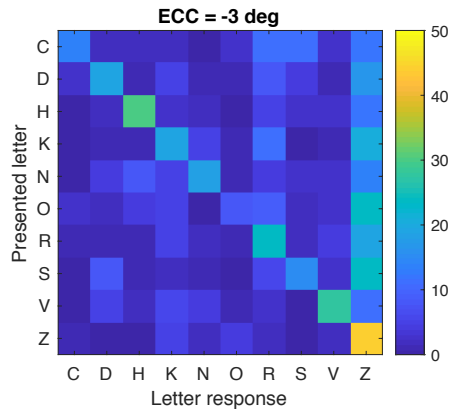
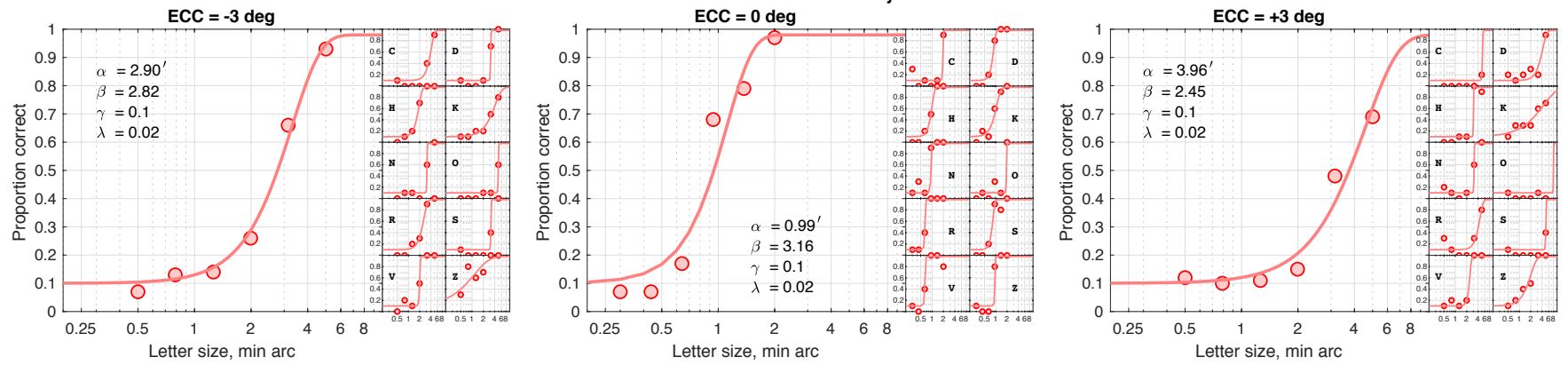
Subject.PK



Subject.SB



Subject.ZC



Appendix B

Appendix B shows the documents of research ethical approval, patient information sheet, patient consent form, patient debrief and the questionnaire of observer's perception of the acuity perimeter.



UNIVERSITY OF
PLYMOUTH

12th November 2020

Hatem S.H. Barhoom
School of Health Professions
Faculty of Health
University of Plymouth
FF01, Peninsula Allied Health Centre
Derriford
Plymouth PL6 8BH

Dear Hatem

**Application for Approval by Faculty Research Ethics and Integrity
Committee**

Reference Number: 19/20-1287

Application Title: Home-Based Central Visual Field Test for
Monitoring of Macular Damage in Glaucoma: A New Design.

The Chair has granted ethical approval to conduct this research.

Please note that this approval is for 3 years (until 11th November 2023), after
which you will be required to seek extension of existing approval.

Please note that if you wish to make any MAJOR changes to your research
you must inform the Committee. Please contact the Faculty Research
Administrator, Maurice Bottomley (email FOHsethics@plymouth.ac.uk).

Yours sincerely

Professor Sarah Neill,
PhD, PGD Res. Deg. Sup., PGDE, MSc, BSc(Hons), RGN, RSCN, RNT
Professor of Nursing
Co-Chair, Research Ethics and Integrity Committee - Faculty of Health

Professor Sarah Neill, PhD
Co-Chair, Faculty of Health Research Ethics and Integrity Committee,
Room 206, 8-11 Kirkby Place, University of Plymouth, Drake Circus, Devon PL4 8AA
T +44(0)1752 586572 E hbsethics@plymouth.ac.uk W www.plymouth.ac.uk



**Home-Based Central Visual Field Test for Monitoring of Macular Damage in
Glaucoma: A New Design.**

Principal Investigator: Hatem S.H. Barhoom

You are invited to take part in a research study. Please read the information below before making your decision. It is important that you understand what the study is about and what will be involved.

What is the aim of this project?

The main objective of this study is to design a new vision test for glaucoma patients using letters.

Why have I been invited?

You have expressed interest in taking part in our study, and you meet the inclusion criteria.

What would I have to do?

The task is to recognize letters presented randomly on a computer screen. You will be asked to make a decision verbally and the examiner will do the input using a computer keyboard. The experimenter will provide you with exact information prior to the experiment. Next, you will perform a visual field test. This test has similar task to the previous one but with detection of spots of light rather than recognition of letters. You will use a button box to respond to seen spots of light. After that we will take an image of the back of the eye (retina) using an imaging device. All the tests are non-invasive.

During the sessions you will be asked to wear some tracking devices (eye tracker which is very similar to the glasses and a heart rate sensor with recording watch).

After the full experiment, you will be asked few questions regarding your subjective impression during the experiment.

Will any expenses be paid?

The participation on this project is voluntary and there is no provision for reimbursement for your time or travel expenses.

Do I have to take part?

The participation in this study is completely voluntary. If you decide to participate in the study, you may withdraw at any time without giving a reason, and this will not be held against you. This study will not affect any current or future treatment, or any insurance policies you may have. If you would like to participate, we will ask you to complete a consent form. We will also provide you with a copy to keep for your own records.

Will my records be confidential?

All information collected will be kept strictly confidential and in accordance with the Data Protection Act (1998). Your records may be reviewed by the research team involved in the study, as well as the monitoring or audit team who are approved by the University of Plymouth. The information collected for the study will be stored electronically on a password protected device. Any contact information such as your name, address and contact details will be stored separately, on a password protected device, so that you cannot be identified. All data will be anonymized. However, you may decide that you do not want your data used in this research. If you would like your data removed from the study and permanently deleted, please **contact us within one week from your participation date, otherwise it will be impossible to identify your data as all data will be encrypted after one week from your participation date.**

Please do not disclose research procedures and/or hypotheses to anyone who might participate in this study in the future as this could affect the results of the study.

What are the potential risks/benefits of taking part?

All the tests in this study are standard clinical tests which are regularly conducted for detection and monitoring of Glaucoma. Only new test is the experimental vision test using letters. There are no risks associated with this study. The full experiment has two sessions. Each session will take around 40 minutes and will be designed to provide you with sufficient breaks. In addition, the nature of these experiments will allow you to take breaks at any time.

Who has reviewed this study?

This study has been approved by the Research Ethics and Integrity Committee of the Faculty of Health and Human Sciences.

Who is involved in this study?

Hatem S. H. Barhoom will be present and will provide information about the experiment.

Will my participation be COVID safe?

As discussed in our first phone call. Upon your arrival to the Wellbeing Centre at the University of Plymouth (where the research will take place) the principal investigator will take your temperature at the main gate. Then, you will be asked to wear personal protective equipment (surgical mask or visor, apron and surgical gloves) (if you consent to do so in our first phone call you will be invited to participate). The examiner will be wearing personal protective equipment as well: surgical mask, apron and surgical gloves. We will be examining one participant per day. The principal investigator will make sure that the disinfection for the research room, chairs, tables, devices and PC will be carried out before your arrival and after your departure.

WHAT IF I HAVE FURTHER QUESTIONS OR REQUIRE FURTHER INFORMATION ABOUT THIS STUDY?

IF YOU HAVE ANY QUESTIONS OR WOULD LIKE MORE INFORMATION, NOW OR IN THE FUTURE, PLEASE CONTACT ME BY TELEPHONE, EMAIL OR IN WRITING, USING THE DETAILS BELOW.

Name: Hatem S.H. Barhoom

Address: FF01, School of Health Professions, University of Plymouth, Peninsula Allied Health Centre, Derriford Road, Plymouth, PL6 8BH.

Tel: 07956308255

Email: hatem.barhoom@plymouth.ac.uk

What if I have a complaint about the study?

Should you feel you have any complaints during the study, please contact the Research Administrator to the Faculty Research Ethics and Integrity Committee (lhsethics@plymouth.ac.uk).

School of Optometry

Faculty of Health and Human Sciences.

Plymouth University

Peninsula Allied Health Centre

Derriford Road

Plymouth – Devon PL6 8BH

Tel +44 (0)1752 600600

Fax +44 (0)1752 600600

<http://www.plymouth.ac.uk>



Consent Form

**Home-Based Central Visual Field Test for Monitoring of Macular Damage in
Glaucoma: A New Design.**

DECLARATION OF INFORMED CONSENT

<p>I have read the above description prior to deciding to participate in this study. I have had an opportunity to ask questions and have received acceptable answers. I have had a complete eye examination within the last two years.</p>	
<p>I am aware that this study has been reviewed and received ethics clearance through the Faculty Research Ethics and Integrity Committee, and that if I have any concerns or questions about my participation in this study, I may contact Hatem S. H. Barhoom, Principal Investigator of this study Tel: 07956308255, Email: hatem.barhoom@plymouth.ac.uk</p>	
<p>I am aware that I may withdraw from the study at any time without affecting my relationship with the University of Plymouth or the School of Health Professions. I am aware that the investigator reserves the right to discontinue my participation from the study at any time, either in regard to the research or the health of my eyes.</p>	
<p>I am aware that my participation in this study is voluntary, but that following study procedures is important to the success of the research. With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.</p>	
<p>I am aware that I will receive a copy of this information and consent letter. I am aware that by signing this form I do not waive my legal rights or release the investigator(s) and /or involved institution(s) from their legal and professional responsibilities.</p>	
<p>I am aware that if I would like my data removed from the study and permanently deleted, I can contact the principal investigator within one week from my participation date, otherwise it will be impossible to identify my data as all data will be encrypted after one week from my participation date.</p>	
<p>I am aware that I cannot disclose research procedures and/or hypotheses to anyone who might participate in this study in the future as this could affect the results of the study.</p>	
<p>I consent to participate in this research.</p>	

(Please insert your initials in the boxes, if you agree)

_____ Signature of participant	Date
_____ Printed name of participant	_____
_____ Signature of person explaining consent	Date
_____ Printed name of person explaining consent	_____



Debriefing Form for Participation in a Research Study (V.1)

Thank you for your participation in our study! Your participation is greatly appreciated.

Purpose of the Study:

We previously informed you that the purpose of the study was to design a new vision test for glaucoma patients. The goal of our research is to investigate whether this new design is better (more sensitive) than the conventional tests in detecting glaucoma progression.

Confidentiality:

You may decide that you do not want your data used in this research. If you would like your data removed from the study and permanently deleted, please contact us within one week from your participation date, otherwise it will be impossible to identify your data as all data will be encrypted after one week from your participation date.

Please do not disclose research procedures and/or hypotheses to anyone who might participate in this study in the future as this could affect the results of the study.

Final Report:

If you would like to receive a copy of the final report of this study (or a summary of the findings) when it is completed, please feel free to contact us.

Useful Contact Information:

If you have any questions, receive a copy of the final report or concerns regarding this study, its purpose or procedures, or if you have a research-related problem, please feel free to contact the researcher:

Name: Hatem S.H. Barhoom

Address: FF01, School of Health Professions, University of Plymouth, Peninsula Allied Health Centre, Derriford Road, Plymouth, PL6 8BH.

Tel: 07956308255

Email: hatem.barhoom@plymouth.ac.uk

If you have any questions concerning your rights as a research subject, you may contact the Research Administrator to the Faculty Research Ethics and Integrity Committee (hhsethics@plymouth.ac.uk).

Please keep a copy of this form for your future reference. Once again, thank you for your participation in this study!

QUESTIONNAIRE

Date		Occupation	
Name		Tested eye	
Age		BCVA in the tested eye, Rx	
Gender		Ocular and general History	
Test group		Coded name	

Thank you for your participation in this experiment. Kindly complete the following questionnaire. The purpose of this questionnaire is to evaluate your experience in this experiment. The data from this questionnaire will be used to help us improving the current experiment.

Q1) from a scale of 0 to 10, how do you rate the length of presentation time of the stimulus (0 = extremely short, 10 = extremely long)?

Q2) how did you find the nature of the task (the recognition of letters at the periphery):

1. Very difficult
2. Moderately difficult
3. Normal
4. Moderately easy
5. Intuitive and easy

Q3) how did you find the response method (verbal response):

1. Extremely exhausting
2. Moderately exhausting
3. Neutral
4. Moderately easy
5. Intuitive and easy

Q4) did you prefer to have feedback after each response? Yes / no

If yes, why?

Q5) how would you rate your experience in this test? Using a scale from 0 to 10 (0 = extremely difficult, 10 = piece of cake!)

Q6) according to your experience in the test and if it is shorter, would you sit for this experiment again? Yes /no

If no, why?

Q7) according to your experience in the test and if it is shorter, would you sit for this experiment weekly? Yes /no

If no, why?

Q8) according to your experience in the test and if it is shorter, would you recommend this experiment for a friend? Yes /no

If no, why?

Q9) what is/are the most reason/s that you think affected negatively your performance in this experiment?

Give a measure from a scale of 0 to 10 for each item (0 = no effect, 10 = huge negative effect) on your performance. (negative effect could be, boring, exhausting, stressful, frustrating test etc...)

- Presentation time of the letters
- The nature of the task (letter recognition)
- No feedback after each trial
- The presence of very small letters
- The need to guess for unseen letter rather than just saying 'not seen'.
- The limited letters (only 10)
- The similarity among certain letters.
- The verbal response method
- Occluding one eye during the test
- Maintain fixation.

Q10) which one of the following was the most uncomfortable to you during the experiment?

- Very small letter when you need to guess (no useful details)
- Very large letter when you are very certain.
- The in-between letter sizes where you are uncertain of what you are looking at.

Supporting Documents

Publications and conferences posters



The effect of response biases on resolution thresholds of Sloan letters in central and paracentral vision

Hatem Barhoom^{*}, Mahesh R. Joshi, Gunnar Schmidtman

Eye and Vision Research Group, School of Health Professions, University of Plymouth, Plymouth, UK

ARTICLE INFO

Keywords:

Visual acuity
Letter acuity
Resolution threshold
Response bias
Luce's choice model
Sloan letters

ABSTRACT

Sloan letters are one of the most commonly used optotypes in clinical practice. Sloan letters have different relative legibility which could be due to three factors: perceivability, response bias, and similarity. Similarities between Sloan letters are known to be the major source of errors in threshold determination. However, little is known about the effect of response biases on the resolution thresholds. The aim of the present study was to investigate the effect of response bias and similarity on resolution thresholds of Sloan letters in central and paracentral vision.

Eight subjects with normal ocular health participated in this study. Using the method of constant stimuli, we measured resolution thresholds for the Sloan letters set at 0° (central) and ± 3° eccentricity along the vertical meridian of the visual field. We calculated thresholds from data pooled across the 10 Sloan letters (pooled threshold). For further analysis we also calculated thresholds for each of the 10 Sloan letters (individual threshold). Response biases and letter similarities were determined using Luce's choice model.

Results showed statistically significant differences between the mean individual thresholds of Sloan letters at the central and the upper visual field, but not at the lower visual field. For equally-sized letters at pooled threshold, unlike letter similarity, response biases showed statistically significant correlations to the differences in individual thresholds at the central, upper and lower visual field locations. For equally legible letters at individual thresholds, response biases and similarities showed no significant correlations to the differences in individual thresholds at the central, the upper and the lower visual field locations.

These results suggest that, for equally-sized letters at pooled threshold, the response biases may lead to an underestimation of the pooled threshold, i.e. an overestimation of visual acuity measurements when using Sloan letters.

1. Introduction

Visual acuity is the ability of the visual system to discern the smallest details of an object, typically measured as the minimum angle of resolution (i.e. detection/resolution threshold) and is of high clinical importance. Clinically, different stimuli or optotypes, such as the Tumbling E, Landolt C and alphanumeric characters have been employed to measure visual acuity (Knietstedt & Stamper, 2003). Although the Landolt C is internationally regarded as the reference optotype (Sloan, 1959; Treacy et al., 2015), letters are used in many visual acuity charts, because they are intuitive and easy to use in clinical settings. Furthermore, employing a variety of letters (e.g. Sloan letters) reduces the guessing rate associated with forced-choice tests (Pelli & Robson, 1991).

Letter identification accuracy is influenced by three factors: (i) perceivability, (ii) response bias and (iii) similarity (Mueller & Weidemann, 2012). Perceivability is a measure of how legible the letter is depending solely on the characteristics of the letters, such as the letter size, contrast or shape. The response bias is defined as the tendency of favouring one response over the other alternatives (Macmillan & Creelman, 1990). Similarity is defined as the confusion in letter perception which arises among certain letters. In other words, letter recognition, i.e. the letter detection/resolution threshold, could be affected by changing the amount or the type of the sensory input, e.g. size and contrast (perceivability), the bias towards certain letters in case of uncertainty (response biases), and the confusion between similar letters such as C and O (similarity). Note that from these definitions it is well understood that response biases, unlike perceivability and letter similarities, are

^{*} Corresponding author.

E-mail address: hatem.barhoom@plymouth.ac.uk (H. Barhoom).

<https://doi.org/10.1016/j.visres.2021.06.002>

Received 4 September 2020; Received in revised form 30 May 2021; Accepted 8 June 2021

Available online 9 July 2021

0042-6989/© 2021 Elsevier Ltd. This article is made available under the Elsevier license (<http://www.elsevier.com/open-access/userlicense/1.0/>).

independent of the sensory inputs of the stimulus.

It has been demonstrated that letters have different legibilities at the fovea (Grimm, Rassow, Wesemann, Saur, & Hilz, 1994; Alexander, Xie, & Derlacki, 1997; Reich & Bedell, 2000; Shah, Dakin, & Anderson, 2012; Hamm, Yeoman, Anstice, & Dakin, 2018; Ludvigh, 1941; Strasburger, Rentschler, & Juttner, 2011; Hairol et al., 2015) and peripheral visual field locations (Ludvigh, 1941; Strasburger et al., 2011; Hairol et al., 2015; Anderson & Thibos, 2004; Shah, Dakin, Redmond, & Anderson, 2011; Shah et al., 2012). The focus of most previous research was to investigate the relative legibility of letters in order to determine the most equally legible letters to optimise the design of letter optotypes used in visual acuity charts. The acceptable differences in legibility between letters is determined by the International Standard ISO 8597. According to this standard, the visual acuity measured by a full set of letters should not deviate by more than 0.05 log units from the visual acuity measured with the Landolt C chart. However, Grimm et al. (1994) recommended that the resolution thresholds of each individual letter should be within 0.05 log units from the mean resolution threshold of the letters set (Grimm et al., 1994). This is of particular importance in the case of letter-by-letter visual acuity measurements. It has been shown that further improvement of the precision of the visual acuity measurements using letters can be obtained by weighting the responses to the letters according to their individual legibility and similarity (Grimm et al., 1994; McMonnies & Ho, 1996; McMonnies & Ho, 2000).

Unlike other common letter stimuli, Sloan letters have been adopted in the design of various letter charts (e.g. Early Treatment Diabetic Retinopathy Study; ETDRS chart), because their average legibility, determined by the letter identification accuracy, is similar to the Landolt C (Sloan, 1959; Treacy et al., 2015). It has been shown that Sloan letters have different relative legibility at the fovea where the letter similarities are the major source of errors in threshold determination (McMonnies & Ho, 1996; Reich & Bedell, 2000; Hamm et al., 2018).

However, little is known about the effect of response biases on the resolution thresholds (or legibility) of individual Sloan letters. In this study, we aim to investigate the relationship between the response biases, similarity and resolution thresholds of individual Sloan letters in central and paracentral locations since the pattern of differences in letter thresholds has been found to be different at the central and paracentral locations (Ludvigh, 1941; Strasburger et al., 2011; Hairol et al., 2015).

2. Methods

2.1. Participants

Eight naïve subjects (six females, mean age 22.60 ± 3.70 (SD), age range: 19–28 years) with normal ocular health participated in this study. The mean best corrected visual acuity and the mean refractive error (spherical equivalent) were -0.04 ± 0.05 logMAR and -2.30 ± 2.63 DS respectively. All tests were conducted monocularly (left or right eye, chosen at random), where the fellow eye was occluded using an opaque eye patch. Written informed consent was obtained from all observers, and the study was approved by the University of Plymouth Ethics committee. All experiments were conducted in accordance with the Declaration of Helsinki.

2.2. Apparatus

Stimuli were generated using MATLAB R2016b (MathWorks, Natick, Massachusetts, USA). Routines from the Psychtoolbox-3 were used to present the stimuli (Brainard, 1997; Pelli, 1997; Kleiner, Brainard, Pelli, Ingling, Murray, & Broussard, 2007). The stimuli were presented on a gamma-corrected DELL, P2317H LCD monitor (1920×1080) with a frame rate of 60 Hz. Monitor linearization was achieved by adjusting colour look-up tables, resulting in 150 approximately equally spaced grey levels. Observers viewed the targets at a viewing distance of 350 cm, while sitting on chair without using a chin or forehead rest. The

examiner monitored viewing distance by regular checks. At this viewing distance one pixel subtended 0.258 min of arc ($^{\circ}$) of visual angle. Experiments were carried out under a room illumination of 160 lx. The observer responded by calling out the responses which were entered by the experimenter via a standard computer keyboard. This method minimised errors caused by mistyping and improved fixation compliance.

2.3. Stimuli

High contrast Sloan letters were used in the experiments (black letters of 2.2 cd/m^2 on a white background of 215 cd/m^2 , resulting in 99% Weber contrast). The variables were letter size, expressed in minutes of arc, and the location of presentation. The letters were presented centrally and at paracentral locations along the vertical meridian at an eccentricity of 3° in the upper and lower visual field (Fig. 1a). Ten standard Sloan letters (C, D, H, K, N, O, R, S, V, Z) were used. Sloan letters are designed so that their height is equal to their width and five times the stroke width (Fig. 1b). We conducted multiple pilot experiments to establish appropriate stimulus levels (letter sizes) to cover the whole range of responses (from chance (10%) to certain decision (100%)). Six different letter sizes (spaced logarithmically) were tested; $0.3'$, $0.44'$, $0.64'$, $0.94'$, $1.37'$, and $2'$ for central presentations and $0.5'$, $0.79'$, $1.26'$, $1.99'$, $3.15'$ and $5'$ for paracentral presentations.

2.4. Procedure

The method of constant stimuli was used in all experiments in this study. The Sloan letters were presented randomly across three locations. Each subject completed 1800 trials for the full experiment (six letter sizes \times three locations \times 10 Sloan letters \times 10 presentations per letter). All conditions were interleaved. The presentation time was 250 ms and presentations were accompanied by an auditory signal. The task for the observer was to recognise the presented letter and to report it verbally. During the experiment, the subjects were asked to fixate on a fixation cross (dimensions: length/width $1.55'$, stroke width $0.036'$) presented at the centre of the screen. The fixation cross disappeared for the duration of the central presentations and reappeared for the paracentral presentations. Subjects were encouraged to guess when uncertain about the letter. Only choices of the 10 Sloan letters were accepted. To familiarise the participants with the Sloan letters, the experimenter demonstrated the Sloan letters in the beginning of the session. The observers showed excellent compliance to answer from the Sloan letter set (on average not more than 30 mistakes per subject). In the rare case where observers responded with a non-Sloan letters, the experimenter prompted for a second response. If the observer failed the second attempt, a reminder of the Sloan letter set was provided (this occurred very rarely, on average

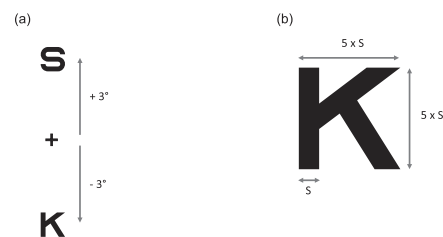


Fig. 1. (a) Sloan letters presented centrally, or along the vertical meridian at an eccentricity of 3° in the upper and lower visual field. The Sloan letters S and K are shown for illustration purposes (not to scale). (b) shows the dimensions of the Sloan letters, exemplified by the letter K. The stroke width $S = 1/5$ of the letter's height. The height and the width of the letter are equal.

not more than once per subject).

2.5. Analysis

Routines from the Palamedes Toolbox (Prins & Kingdom, 2018) were employed to fit individual psychometric functions. Gumbel (Log-Weibull) functions (Eq. (1)) were fit to determine thresholds.

$$P_{correct} = (\gamma + (1 - \gamma - \lambda)) (1 - \exp(-10^{(\lambda x - \alpha)})) \tag{1}$$

where γ is the guessing rate (10 letters, $\gamma = 0.1$), λ is the lapse rate ($\lambda = 0.02$, naïve subjects), x is the letter size (log visual angle), α is the threshold and σ is the slope of the function. The threshold α was defined as x yielding 65.6% correct responses, according to the following equation.

$$P_{correct} = 0.1 + (1 - 0.1 - 0.02) (1 - \exp(-10^{(\sigma(\alpha - x))})) \approx 0.656 \tag{2}$$

Thresholds were calculated from data pooled across the 10 Sloan letters. We refer to these as pooled thresholds. For further analysis we also calculated thresholds for each of the 10 Sloan letters. These are referred to as individual thresholds.

3. Results

The mean pooled thresholds across subjects (\pm SD) were 0.04 ± 0.05 , 0.40 ± 0.11 and 0.42 ± 0.12 log visual angle at the central, upper visual field (3°) and lower visual field (3°) locations respectively. Paired sample t -test revealed no statistically significant differences between the mean pooled thresholds at upper and lower visual field locations ($t(7) = -0.77$, $p = .47$). However, the mean pooled thresholds at the upper visual field ($t(7) = 6.12$, $p < .001$) and lower visual field ($t(7) = 6.90$, $p < .001$) were significantly higher compared to the central location. The mean individual thresholds across subjects at each location are shown in Table 1. One-way ANOVA tests showed statistically significant differences between the mean individual thresholds at the central location ($F(9, 70) = 4.93$, $p < .001$) (with highest and lowest thresholds for O = 0.10 ± 0.06 and V = -0.10 ± 0.05 respectively) and upper visual field ($F(9, 70) = 2.0$, $p < .05$) (with highest and lowest thresholds for S = 0.46 ± 0.18 and H = 0.25 ± 0.14 respectively). However, there was no statistically significant difference between the mean individual thresholds at the lower visual field ($F(9, 70) = 1.7$, $p = .057$) (with highest and lowest thresholds for O = 0.52 ± 0.20 and V = 0.29 ± 0.13 respectively).

In order to investigate the pattern of the individual threshold differences, we calculated the relative thresholds for the letters as the difference between the pooled thresholds and the individual thresholds (Fig. 2). The analysis revealed statistically significant correlations of relative thresholds between the lower visual field and the central location ($r = 0.83$, $n = 10$, $p < .05$) and also between the lower visual field and upper visual field ($r = 0.72$, $n = 10$, $p < .05$). This suggests that the pattern of differences between individual thresholds are consistent at the central location, and the upper and lower visual field (Fig. 3).

For the following analyses, confusion matrices (presented vs.

Table 1

The table shows the mean of individual thresholds (Mean \pm SD in log visual angle) at central and paracentral locations.

	Central	Upper field (3°)	Lower field (3°)
C	0.06 \pm 0.11	0.377 \pm 0.08	0.45 \pm 0.14
D	0.04 \pm 0.09	0.40 \pm 0.13	0.42 \pm 0.12
H	0.02 \pm 0.11	0.25 \pm 0.14	0.34 \pm 0.10
K	0.06 \pm 0.07	0.40 \pm 0.09	0.43 \pm 0.07
N	-0.07 \pm 0.08	0.32 \pm 0.14	0.36 \pm 0.16
O	0.10 \pm 0.06	0.44 \pm 0.17	0.52 \pm 0.20
R	-0.02 \pm 0.06	0.33 \pm 0.17	0.38 \pm 0.14
S	0.06 \pm 0.10	0.46 \pm 0.18	0.39 \pm 0.17
V	-0.10 \pm 0.05	0.27 \pm 0.07	0.29 \pm 0.13
Z	-0.01 \pm 0.02	0.35 \pm 0.11	0.32 \pm 0.10

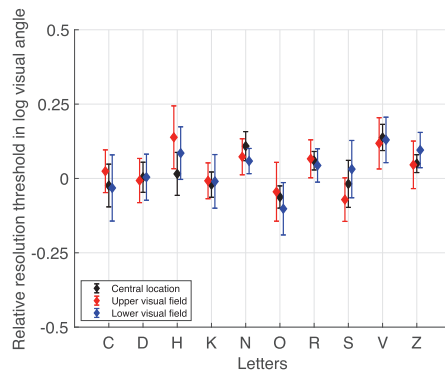


Fig. 2. shows the mean relative thresholds of letters at central, upper (3°) and lower visual field (3°). The error bars here and throughout represent 95% confidence intervals.

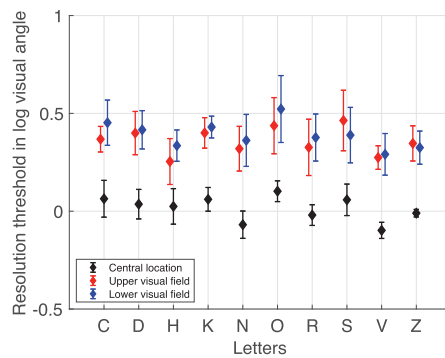


Fig. 3. shows the mean individual thresholds at central, upper and lower visual field locations.

responses of letters) were created for individual observers at each location. Note that the data in the current study were not collected at the pooled threshold or the individual threshold sizes. We therefore extrapolated the expected confusion matrices from the available data (see Appendix for details).

The confusion matrices were calculated for two conditions. In the first condition, the expected confusion matrices at each subject's pooled threshold size were calculated. As a consequence, all presented letters in this condition had equal stroke sizes (referred to as ESS) which was equal to the size of the pooled threshold. In the second condition, the expected confusion matrices at each letter's individual threshold sizes were calculated. In this case all letters had equal legibility sizes (referred to as ELS), regardless of the stroke size of the letter (equal performances; $P_{correct} \approx 0.656$). Fig. 4 shows the mean (\pm SD) of the expected confusion matrices for the two conditions (ESS and ELS) at central, upper and lower visual field.

3.1. Model

The difference in individual thresholds between letters can be caused

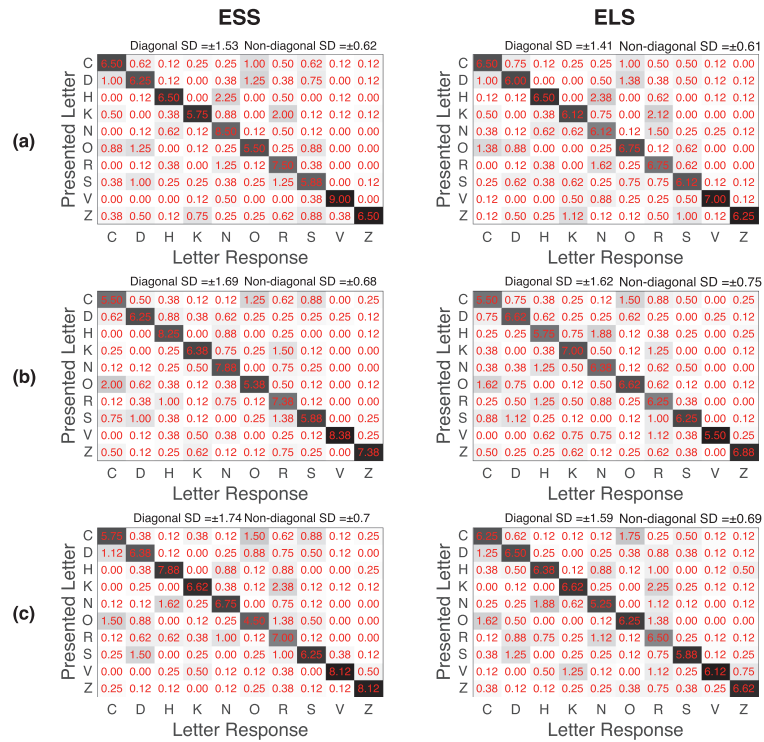


Fig. 4. shows the mean ± SD of expected confusion matrices for the two conditions (ESS and ELS) at the (a) central, (b) upper and (c) lower visual field. The greyscale illustrates the frequency of letter responses, where darker cells show higher frequencies. The diagonal cells represent correct responses, whereas the non-diagonal cells represent incorrect responses. The SD is calculated as the square root of the average of the variances of the cells for diagonal and non-diagonal values separately.

by the difference in the relative legibility of the letters, response biases and/or letter similarities. To investigate the potential effect of these three factors on the letter detection thresholds, response biases and letter similarities were computed using Luce's choice model (Luce, 1963). This model attempts to disentangle the response factor that is sensory-independent (i.e. response biases towards some letters) from the sensory-dependent response factor (i.e. similarities between certain letters). Luce's choice model was used to estimate the response biases (expressed in response bias vector) (Eq. (3)) and letter similarities (Eq. (4)). The model predications are presented as similarity matrices capturing the similarity between each pair of letters parameters. These parameters were calculated from the matrices of the maximum likelihood estimates which resulted from the model fit (see Appendix). According to the Luce's model,

$$\beta_j = \frac{1}{\sum_{k=1}^N \sqrt{\frac{\hat{p}_{kj} \times \hat{p}_{jk}}{\hat{p}_{ij} \times \hat{p}_{ij}}}} \quad (3)$$

and

$$\eta_{ij} = \sqrt{\frac{\hat{p}_{ij} \times \hat{p}_{ji}}{\hat{p}_{ii} \times \hat{p}_{jj}}} \quad (4)$$

where the variable β in Eq. (3) denotes the response bias parameter for the letter j . N is the number of letters (10 letters). η is the similarity parameter of each cell between the letter i and the letter j . \hat{p} is the expected value in each cell obtained from the maximum likelihood estimates matrix.

3.2. Response bias

Table 2 shows the mean (±SD) response biases; β parameter in Eq. (3). Values of 0.1 would imply guessing (i.e. no-bias value). Letters with values higher than 0.1 are considered to be relatively biased on the expense of those with values lower than 0.1. Fig. 5 shows the response biases (mean with 95% CI) for individual letters at the two conditions, ESS and ELS, depicted for central, upper and lower visual field locations. For letters presented at ESS, the β values for the letters C, K, O, V, Z at central, upper and lower, the letter D at central and lower, and the letter S at upper and lower visual field locations were significantly lower than the guessing rate, indicating that the response biases were towards the

Table 2
shows the mean of individual bias parameters (Mean \pm SD) for letters for the two conditions (ESS and ELS) at central and paracentral locations.

	β for letters presented at ESS			β for letters presented at ELS		
	Central	Upper field (3°)	Lower field (3°)	Central	Upper field (3°)	Lower field (3°)
C	0.037 \pm 0.020	0.047 \pm 0.036	0.027 \pm 0.032	0.118 \pm 0.005	0.105 \pm 0.087	0.120 \pm 0.098
D	0.049 \pm 0.036	0.070 \pm 0.074	0.056 \pm 0.036	0.099 \pm 0.011	0.125 \pm 0.100	0.130 \pm 0.103
H	0.121 \pm 0.075	0.151 \pm 0.084	0.118 \pm 0.078	0.084 \pm 0.052	0.101 \pm 0.068	0.078 \pm 0.043
K	0.043 \pm 0.044	0.056 \pm 0.050	0.047 \pm 0.038	0.090 \pm 0.069	0.140 \pm 0.078	0.074 \pm 0.070
N	0.234 \pm 0.085	0.115 \pm 0.081	0.169 \pm 0.142	0.150 \pm 0.071	0.121 \pm 0.062	0.079 \pm 0.081
O	0.031 \pm 0.029	0.020 \pm 0.015	0.032 \pm 0.032	0.104 \pm 0.032	0.102 \pm 0.069	0.072 \pm 0.038
R	0.165 \pm 0.063	0.110 \pm 0.075	0.133 \pm 0.079	0.146 \pm 0.028	0.134 \pm 0.077	0.153 \pm 0.034
S	0.084 \pm 0.076	0.036 \pm 0.038	0.047 \pm 0.028	0.090 \pm 0.060	0.064 \pm 0.049	0.076 \pm 0.057
V	0.028 \pm 0.010	0.015 \pm 0.011	0.048 \pm 0.042	0.038 \pm 0.036	0.017 \pm 0.011	0.028 \pm 0.027
Z	0.026 \pm 0.023	0.041 \pm 0.042	0.046 \pm 0.029	0.017 \pm 0.018	0.011 \pm 0.005	0.094 \pm 0.083

remaining letters (mainly H, N, R). The pattern of the response biases in the central location was found to be significantly correlated with the pattern of response biases at the upper ($r = 0.77, n = 10, p < .01$) and lower ($r = 0.96, n = 10, p < .0001$) visual field locations. The patterns of the response biases at upper and lower visual field locations were also significantly correlated ($r = 0.85, n = 10, p < .01$). Results show that the biased letters (H, N, R) were associated with letters with low resolution thresholds (or high legibility) at central, upper and lower visual field locations. These three letters were among the five letters with low resolution thresholds. Unlike the other low resolution thresholds letters (H, N, R), the letters V and Z were associated with low bias parameters. After excluding the letters V and Z, it was found that the correlations between the response biases parameters and the individual thresholds were significantly and negatively correlated at central ($r = -0.95, n = 8, p < .001$), upper ($r = -0.94, n = 8, p < .001$) and lower ($r = -0.76, n = 8, p < .05$) visual field locations.

For the ELS condition, at least seven letters showed β values distributed around 0.1 (guessing; Fig. 5) at central, upper and lower visual field locations. The response biases were found to be mainly towards the letter R at these locations on the expense of the letter V at central, upper and lower and on the expense of the letter Z at central and upper visual field locations. The pattern of the response biases of central location was found to be significantly correlated with the pattern of response biases at the upper ($r = 0.75, n = 10, p < .05$) and at the lower

($r = 0.70, n = 10, p < .05$) visual field locations. The patterns of the response biases at upper and lower visual field locations were also significantly correlated ($r = 0.70, n = 10, p < .05$). After excluding the letters V and Z, (which showed almost similar behaviour as in ESS condition) it was found that the correlations between the β values and the individual thresholds were not significant at central ($r = -0.60, n = 8, p = .12$), upper ($r = -0.25, n = 8, p = .55$) and lower ($r = -0.31, n = 8, p = .45$) visual field locations.

These results suggest that the differences in individual thresholds (i.e., relative legibility) induced response biases towards the letters with low resolution thresholds (i.e. high legibility, such as H, N, R) and induced biases against the letters with high resolution thresholds (i.e. low legibility, such as C, D, K, O) in the ESS condition.

3.3. Similarity

The mean (\pm SD) of letter similarities for letter pairs (η parameters in Eq. (4)) at the central and paracentral locations are shown in Fig. 6 as similarity matrices for the two conditions (ESS and ELS). η values of 1 represents completely identical letters (not shown in Fig. 6). For the ESS condition, the most confused pairs (i.e. with η parameter closest to 1) were found to be O-D at central, O-C at upper and S-D at lower visual field locations. For further analysis, the confusability of each letter with the remaining letters was calculated as the average of the η parameter (Fig. 7). The letter D at the central and the lower and C at the upper visual field locations showed the highest confusability for the ESS condition, whereas the letters V at the central and Z at the upper and the lower visual field location showed the lowest confusability (Table 3). The pattern of the confusability of individual letters of the central locations was significantly correlated with the pattern of confusability at the upper ($r = 0.89, n = 10, p < .001$) and lower ($r = 0.90, n = 10, p < .001$) visual fields locations. The pattern of confusability at the upper and lower visual fields were also correlated ($r = 0.88, n = 10, p < .001$). Despite some preference towards the confusion among curved letters (O, D, C, S), the confusability did not show significant correlations with the individual thresholds at central ($r = 0.60, n = 10, p = .066$), upper ($r = 0.55, n = 10, p = .10$) and lower ($r = 0.60, n = 8, p = .08$) visual field locations. However, for the letters presented at ELS condition, the most confused pairs were found to be R-N at central, N-H at upper and lower visual field locations. The letters N at the central and the upper and R at the lower visual field location showed the highest confusability. Additionally, the letter V at the central, the upper and the lower visual field locations showed the lowest confusability (Table 3). The pattern of the confusability of individual letters of the central location was significantly correlated with the pattern of confusability at the upper ($r = 0.92, n = 10, p < .001$) and lower ($r = 0.70, n = 10, p < .05$) visual fields locations. The pattern of confusability at the upper and lower visual fields were also significantly correlated ($r = 0.75, n = 10, p < .05$). The confusability did not show significant correlations (or association) with

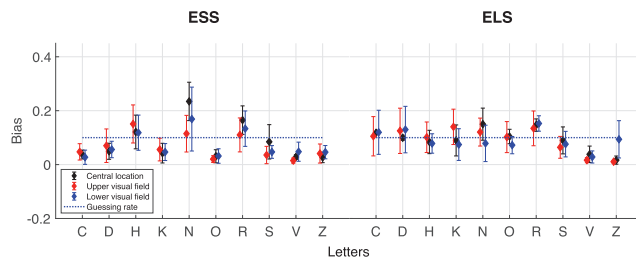


Fig. 5. shows the mean response biases for individual letters calculated by Luce choice model for the two conditions (ESS and ELS) at the central, upper (3°) and lower (3°) visual field. The horizontal dot line is the β parameters of the guessing rate or no-bias (0.1).

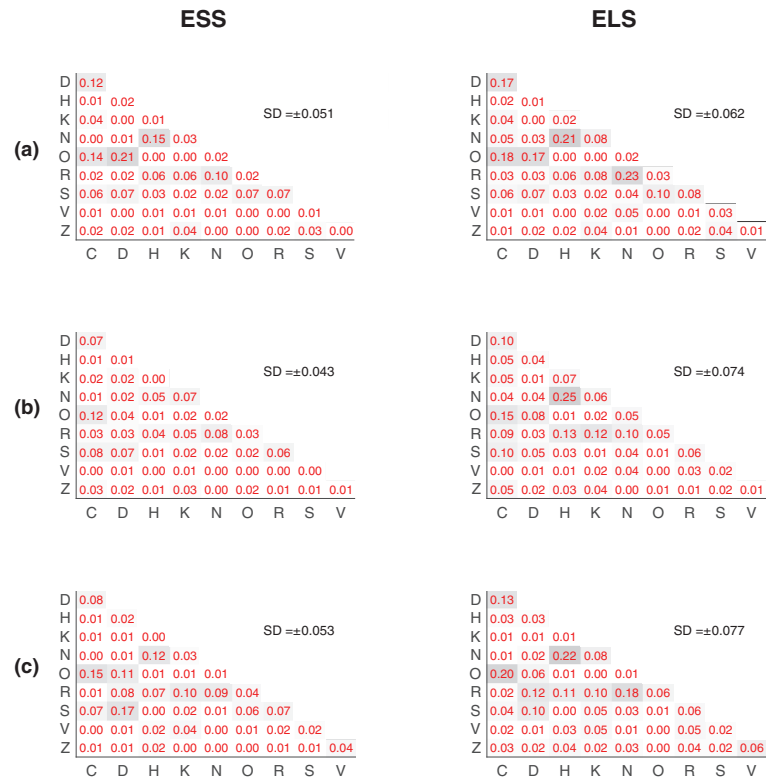


Fig. 6. shows the mean (\pm SD) of letter similarities (η parameters in Eq. (4)) for letter pairs for the two conditions (ESS and ELS) at the (a) central location, (b) upper (3°) and (c) lower (3°) visual field. SD is calculated as the square root of the average of η parameters' variances. The greyscale illustrates η parameters, where darker cells show higher η parameters.

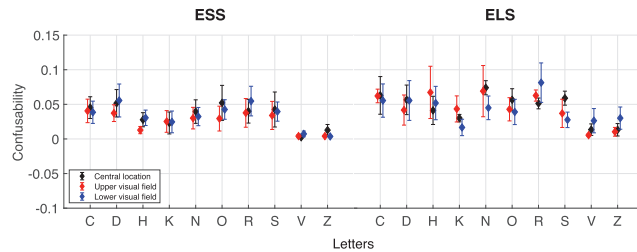


Fig. 7. shows the mean of the confusability of each individual letter for the two conditions (ESS and ELS) at the central, upper (3°) and lower (3°) visual field.

the individual thresholds at central ($r = 0.18, n = 10, p = .61$), upper ($r = 0.04, n = 10, p = .91$) and lower ($r = 0.16, n = 10, p = .65$) visual field locations.

In summary, these results suggest that differences in individual

thresholds (i.e., relative legibility) had a small effect on the similarity preferences. The letters with high resolution thresholds (i.e., low legibility) might induce a preference towards the confusion among the curved letters (O, D, C, S). This preference has disappeared in the ELS

Table 3
shows the mean of individual confusability (Mean \pm SD) for the two conditions (ESS and ELS) for letters at central and paracentral locations.

	The confusability of letters presented at ESS			The confusability of letters presented at ELS		
	Central	Upper field (3°)	Lower field (3°)	Central	Upper field (3°)	Lower field (3°)
C	0.045 \pm 0.019	0.041 \pm 0.020	0.039 \pm 0.019	0.063 \pm 0.033	0.062 \pm 0.012	0.055 \pm 0.029
D	0.052 \pm 0.024	0.037 \pm 0.014	0.056 \pm 0.028	0.056 \pm 0.026	0.042 \pm 0.026	0.056 \pm 0.034
H	0.028 \pm 0.012	0.013 \pm 0.006	0.031 \pm 0.013	0.041 \pm 0.024	0.067 \pm 0.045	0.052 \pm 0.029
K	0.023 \pm 0.019	0.025 \pm 0.019	0.025 \pm 0.019	0.030 \pm 0.006	0.043 \pm 0.023	0.017 \pm 0.014
N	0.040 \pm 0.020	0.030 \pm 0.019	0.032 \pm 0.016	0.074 \pm 0.012	0.069 \pm 0.044	0.045 \pm 0.020
O	0.052 \pm 0.030	0.029 \pm 0.021	0.043 \pm 0.017	0.057 \pm 0.019	0.043 \pm 0.020	0.039 \pm 0.022
R	0.040 \pm 0.021	0.038 \pm 0.025	0.055 \pm 0.026	0.051 \pm 0.009	0.063 \pm 0.010	0.081 \pm 0.034
S	0.043 \pm 0.030	0.034 \pm 0.024	0.040 \pm 0.017	0.059 \pm 0.012	0.037 \pm 0.024	0.028 \pm 0.013
V	0.002 \pm 0.002	0.004 \pm 0.004	0.007 \pm 0.003	0.013 \pm 0.010	0.005 \pm 0.002	0.026 \pm 0.021
Z	0.013 \pm 0.010	0.004 \pm 0.002	0.004 \pm 0.001	0.013 \pm 0.011	0.010 \pm 0.008	0.030 \pm 0.019

condition.

4. Discussion

The aim of this study was to investigate the relationship between the response biases, similarity, and the individual thresholds of Sloan letters. Results showed that pooled thresholds were significantly different between the central and the paracentral visual field locations with no significant difference between the upper (3°) and lower (3°) locations. Fig. 8 shows the slope of the regression line of pooled threshold as a function of eccentricity for the current (blue) and two previous studies (green: Ludvigh, 1941; red: Hairol et al., 2015). A direct comparison of the regression lines shows a similar slope, but higher pooled thresholds measured in the current study (Fig. 8). The higher pooled thresholds in the current study could be the result of differences in study design. Hairol et al. (2015) used Sheridan Gardiner letters and unlimited

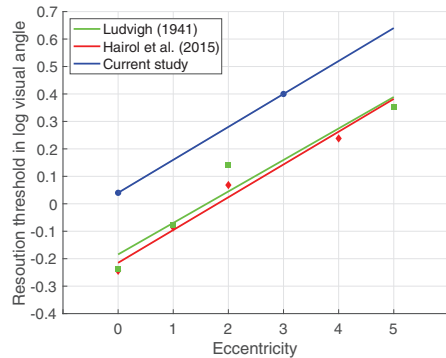


Fig. 8. shows the pooled thresholds as a function of eccentricity for two previous studies (green and red) and for the current study (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

viewing time, and Ludvigh (1941) used F, E, C, L, T letters and did not specify the viewing time.

Similar to previous studies, our results show significant differences (and similar pattern) of individual thresholds (Alexander et al., 1997; Reich & Bedell, 2000; Shah et al., 2012; Hamm et al., 2018). Fig. 9 shows the individual thresholds (central) compared to four previous studies (Alexander et al., 1997; Reich & Bedell, 2000; Shah et al., 2012; Hamm et al., 2018). Our results were generally similar to Alexander et al. (1997), Reich and Bedell (2000) and Shah et al. (2012) results, but different from Hamm's (2018) results, who measured lower individual thresholds for all letters. The lower individual thresholds in Hamm's (2018) study could be the consequence of using white optotypes on a black background, using interleaved QUEST staircase procedure, and a potential learning effect which could arise from the long and multiple experimental sessions (10 h per subject) (Westheimer, 2003; Hamm et al., 2018). However, despite using different psychophysical procedures¹, the pattern of the differences of individual thresholds were similar in all studies. This suggests that using different psychophysical procedures to measure the resolution thresholds has no influence on the pattern of the differences of Sloan letter individual thresholds. The current study also showed similar patterns of the differences in individual thresholds at the fovea, upper, and lower visual field locations, but with higher variability in individual thresholds in the upper and lower visual fields locations. This was consistent to what has been reported by Hairol et al. (2015).

Here we used Luce's choice model (Luce, 1963) to estimate letter similarity and response bias parameters to investigate the relationship of these parameters with the individual thresholds of Sloan letters. Our results are consistent with previous studies. Specifically, curved letters (such as C, D, O, S) were confused more frequently with each other than with the straight vertical/oblique letter (such as H, K, N, R) and vice versa. The majority of confusion pairs at the central location were found to be similar to those at paracentral locations (but with different values and ranks of η parameters). The top five similarly perceived letter pairs in term of η parameter (e.g. C and O is a similarly perceived letters pair) were compared to results reported previously (Table 4). Similar to

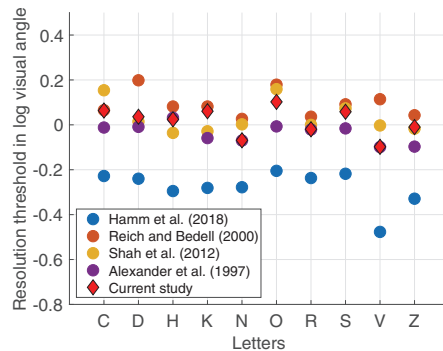


Fig. 9. shows the individual thresholds for the current and four previous studies (Alexander et al., 1997; Reich & Bedell, 2000; Shah et al., 2012; Hamm et al., 2018).

¹ Alexander et al. (1997), Reich and Bedell (2000) and the current study employed the method of constant stimuli, Shah et al. (2012) the method of limits, and Hamm et al. (2018) an interleaved QUEST staircase

previous studies, in the ESS condition, the preference was mostly towards confusion among curved letters (Shah et al., 2012; Reich & Bedell, 2000). Four out of five (three curved letters pairs and one straight letters pairs) in the fovea and two out of five (curved letters pairs) at the upper paracentral visual field were found to be similar to what was reported by Reich and Bedell (2000). These differences could also be due to differences in study design. In contrast to the current study, Reich and Bedell (2000) measured thresholds at 10° superior visual field location and used separate sessions for the fovea and periphery. Reich and Bedell (2000) used 25 letters which would result in different combinations of confusion pairs. Additionally, in the current study, subjects were “forced” to respond only from Sloan letters set. In this case subjects created different combinations or at least different strengths of the letter similarities (Carkeet, 2001).

In the ELS condition, the similarity parameters (and the confusability) were generally increased compared to the ESS condition. The increase was higher for letter pairs with straight vertical/oblique features such as R, N, H. In addition, the variabilities increased at three locations for similarity and confusability which implies that the confusion between letters became more random across subjects with no preference towards either of the letters’ confusion sets. These results are similar to Hamm et al. (2018) who calculated the similarity from data collected near individual threshold of Sloan letters (i.e. ELS condition in the current study). Three out of five similarity pairs (two pairs from curved letters and one pair from straight vertical/oblique lines letters) at the fovea were found to be similar to the results of Hamm et al. (2018), but with different η parameters. As mentioned earlier, this could be the results of using reversed contrast optotypes (white optotypes on a black background) and/or employing different methods to determine the individual thresholds.

We further aimed to investigate the relationship between the letter similarities and the differences in individual thresholds. The similarity (expressed as confusability) did not show significant correlations with the individual thresholds at the central and paracentral locations. Hence, the similarity was unlikely the cause of the differences in individual thresholds. However, our results suggest that differences in individual threshold (i.e., relative legibility) had a small effect in the similarity preferences. The letters with high thresholds (i.e. low legibility) might induce the preference towards the confusion among the curved letters (O, D, C, S) in the ESS condition. This preference was absent in the ELS condition. Nevertheless, the increase of letter similarity and variability in the ELS condition suggest that the letter similarity are a major source of “non-random” errors in the estimation of individual thresholds. These findings are consistent with the previous studies (Erdei & Fulep, 2019; Grimm et al., 1994; McMonnies & Ho, 1996; McMonnies & Ho, 2000; Hamm et al., 2018).

Note that for the letters V and Z, Reich and Bedell (2000) found that the letter V was confused with the letters W and Y at central and the upper visual field (10°), and the letter Z was confused with the letter T at the upper visual field (10°) with no confusion with any letter at the fovea. However, in the current study we limited the responses to Sloan letters, which might explain the very low confusability of the letters V

and Z in the current study and Hamm et al. (2018).

In the ESS condition, the response biases were consistently towards the letters H, N, R across the subjects. The β parameters negatively correlated with the individual thresholds at central, upper and lower visual field locations (after excluding the letters V and Z). In the ELS condition, on the other hand, and similar to the results of Hamm et al. (2018) (Fig. 10), the response biases for at least seven letters at central, upper and lower visual field locations were at guessing rate (0.1) (except for the letter R mainly on the expense of the letters V and Z). Moreover, the correlations between the β values and the individual thresholds were no longer significant at central, upper and lower visual field locations. These findings suggest that the differences of individual thresholds substantially induced the response biases towards the letters with low resolution thresholds (i.e. high legibility) in the ESS condition.

In the current study, the letters V and Z showed low response biases compared to the other highly legible letters such as N, H and R. This could be due to the high legibility associated with the very low confusability of the letters V and Z. The subjects might be reluctant to call these two letters in the case of uncertainty because the letters were distinctive (highly legible with no similar letters in the Sloan letters set). This was possible especially when presenting the letters at different sizes when using the method of constant stimuli. In this case the subjects would not call these two letters unless they were very certain as these letters remained distinct at different sizes. On the other hand, for the letter R for example, which showed high legibility and confusability compared to V and Z, the subjects had a higher chance to call the letter when using the method of constant stimuli, either as correctly perceived letter or wrongly confused with other similar letters (such as N and K). This would lead subjects to overcall this letter compared to V and Z in the case of uncertainty and would consequently cause a higher response

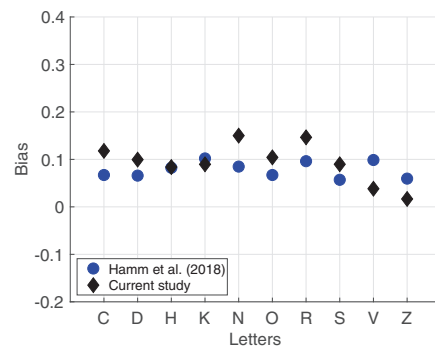


Fig. 10. shows the β parameters of individual Sloan letters at the fovea for the current study and Hamm et al. (2018).

Table 4 shows the top five similarly perceived letters pairs in the current and two previous studies (Hamm et al., 2018; Reich & Bedell, 2000).

Central location		Hamm et al. (2018)					Reich and Bedell (2000)					Upper visual field					
Current experiment		ESS					ESS					Current experiment (3°)					
ELS		ESS					ESS					Current experiment (3°)					
Letters pairs	η	Letters pairs	η	Letters pairs	η	Letters pairs	η	Letters pairs	η	Letters pairs	η	Letters pairs	η	Letters pairs	η	Letters pairs	η
R-N	0.23	O-D	0.21	N-H	0.14	C-O	0.17	N-H	0.25	C-O	0.12	G-O	0.24				
N-H	0.21	N-H	0.15	O-D	0.12	O-D	0.16	C-O	0.15	S-C	0.08	N-H	0.20				
C-O	0.18	C-O	0.14	R-K	0.11	N-H	0.12	R-H	0.13	R-N	0.08	R-D	0.16				
O-D	0.17	C-D	0.12	C-O	0.09	R-K	0.10	R-K	0.12	C-D	0.07	R-K	0.13				
C-D	0.17	R-N	0.10	S-C	0.05	C-D	0.09	D-C	0.10	S-D	0.07	S-C	0.13				

* % conf. = The probability of confusion.

bias.

Our experiment clearly demonstrates that the response bias is substantially influenced by the relative legibility of the individual letters when letters are presented at the size of the pooled threshold (i.e. ESS). The response biases occur for the letters with high legibility. This might be because of the difference in the rate of errors committed for each letter. In experiments using the method of constant stimuli, each letter has the same number of presentations. In this case, high legibility letters have lower error rates than letters with low legibility when presented at the size of the pooled threshold (i.e. ESS). Hence, in case of uncertainty, there will be a lower chance to call the low legibility letters when presenting high legibility letters. On the other hand, when the low legibility letters are presented, in case of uncertainty there will be a higher chance to call the high legibility letters because of the higher error rates associated with the low legibility letters. This assumption is only valid when associated with some level of confusability. In the absence of confusability, the highly legible letters will be very distinctive and will not be called when uncertain, hence low response biases were observed for the letters V and Z. This assumption is also valid for the response biases calculated at individual thresholds (i.e. ELS). In this case, all letters have equal legibility (regardless of the size), hence equal performances lead to equal error rates for all letters. Therefore, all letters have the same chance to be called when uncertain, which causes no or minimal response biases. Note that, in current study, over-calling the highly legible letters did not seem to have a significant effect on the response biases. However, this might be the cause of the response bias observed mainly for the letter R in the ELS condition.

Our results suggest that the response biases could have significant impact on individual and pooled thresholds. Over-calling of the high legible letters increases the chance of the correct guessing compared to low legibility letters. This would result in an underestimation of the individual thresholds of high legible letters and would consequently result in lower pooled thresholds.

In conclusion, the results of the current study emphasize the importance of adopting equally legible letters to minimise the response biases and hence to avoid potential underestimation of pooled thresholds, i.e. an overestimation of visual acuity.

CRediT authorship contribution statement

Hatem Barhoom: Conceptualization, Methodology, Software, Writing – original draft. **Mahesh R. Joshi:** Supervision, Data curation, Writing – review & editing. **Gunnar Schmidtman:** Visualization, Investigation, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Luce's choice model

The fitting algorithm was originally presented by Smith (1982) and consists of two steps (Smith, 1982; Coates, 2015). The first step is to find the maximum likelihood estimate of the model. The iterative proportional fitting is used to converge the raw (confusion) matrix to the maximum likelihood estimate of the model. The starting matrix values are all ones. In order to perform the first iteration, adjustment for rows, columns and similarities are carried out successively. The adjustment for rows is performed by dividing the value of each cell by the sum of the corresponding row values of the starting matrix (ones for the first iteration) then multiplied by the marginal sum of the corresponding row

values of the raw (confusion) matrix, followed by the adjustment for columns and for similarities. The resulting matrix will be the starting matrix for the second iteration. These iterations are repeated until there is no significant change in the estimated values. The resulting matrix is the maximum likelihood estimate of the model. The second step is to compute the parameters of the response bias vector and the similarity matrix from the maximum likelihood estimate of the model, according to Eqs. (3) and (4) respectively.

The MATLAB implementations to compute the response bias and similarity parameters using Luce's choice model can be downloaded from here: <https://github.com/HBarhoom/Codes>

The expected confusion matrix

The expected confusion matrix at the pooled thresholds and the individual thresholds were calculated according to the following method. If a letter's threshold falls between two particular letter sizes, we used the confusion matrices of these two sizes to calculate the expected confusion matrix of the letter. First, we determined the ratio of the letter's threshold to the two enclosing letter sizes (e.g., if the ratio is 0.5, we assumed that the letter's threshold is halfway between the two enclosing letter sizes). Next, we calculated the differences between the corresponding cells of the two confusion matrices of the enclosing letter sizes. Finally, we used the calculated ratio to extrapolate the expected confusion matrix at the letter's threshold level from these differences. We assumed that the central part of the psychometric function between the upper and lower asymptotes (i.e., at the threshold) is approximately linear.

References

- Alexander, K. R., Xie, W., & Derlacki, D. J. (1997). Visual acuity and contrast sensitivity for individual Sloan letters. *Vision Research*, 37(6), 813–819.
- Anderson, R., & Thibos, L. (2004). The filtered Fourier difference spectrum predicts psychophysical letter discrimination in the peripheral retina. *Spatial Vision*, 17(1–2), 5–15.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4), 433–436.
- Carkeet, A. (2001). Modeling logMAR visual acuity scores: Effects of termination rules and alternative forced-choice options. *Optometry and Vision Science*, 78(7), 529–538.
- Coates, D. R. (2015). *Quantifying crowded and uncrowded letter recognition* (Doctoral dissertation. Berkeley): University of California.
- Erdei, G., & Fulep, C. (2019). U.S. Patent Application No. 16/394,388.
- Grimm, W., Rassow, B., Wesemann, W., Saur, K., & Hiltz, R. (1994). Correlation of optotypes with the Landolt ring – A fresh look at the comparability of optotypes. *Optometry and Vision Science*, 71(1), 6–13.
- Hairol, M. I., Abd-Lattif, N. A., Low, P., Lim, W. P., Aik, J. Y., & Kaur, S. (2015). Effects of foveal and eccentric viewing on the resolution and contrast thresholds of individual letters. *Psychology & Neuroscience*, 8(2), 183–192.
- Hamm, L. M., Yeoman, J. P., Anstice, N., & Dakin, S. C. (2018). The Auckland Optotypes: An open-access pictogram set for measuring recognition acuity. *Journal of Vision*, 18(3), 13.
- Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, 36(14), 1–16.
- Kniestedt, C., & Stamper, R. (2003). Visual acuity and its measurement. *Ophthalmology Clinics of North America*, 16(2), 155–170.
- "Detection and recognition". *Handbook of mathematical psychology* (Vol. 1., (1963), 103–189.
- Ludvig, E. (1941). Extrafoveal visual acuity as measured with snellen test-letters. *American Journal of Ophthalmology*, 24(3), 303–310.
- Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, 107(3), 401–413.
- McMonnies, C. W., & Ho, A. (1996). Analysis of errors in letter acuity measurements. *Clinical and Experimental Optometry*, 79(4), 144–151.
- McMonnies, C. W., & Ho, A. (2000). Letter legibility and chart equivalence. *Ophthalmic and Physiological Optics*, 20(2), 142–152.
- Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceptibility, similarity, and bias. *Acta Psychol (Amst)*, 139(1), 19–37.
- Pelli, D., & Robson, J. (1991). Are letters better than gratings? *Clinical Vision Sciences*, 6(5), 409–411.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4), 437–442.
- Prins, N., & Kingdom, F. A. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the palamedes toolbox. *Frontiers in Psychology*, 9(1250).

- Reich, L. N., & Bedell, H. E. (2000). Relative legibility and confusions of letter acuity targets in the peripheral and central retina. *Optometry and Vision Science*, *77*(5), 270–275.
- Shah, N., Dakin, S. C., Redmond, T., & Anderson, R. S. (2011). Vanishing optotype letter acuity: Repeatability and effect of the number of alternatives. *Ophthalmic and Physiological Optics*, *31*, 17–22.
- Shah, N., Dakin, S. C., & Anderson, R. S. (2012). Effect of optical defocus on detection and recognition of vanishing optotype letters in the fovea and periphery. *Investigative ophthalmology & visual science*, *53*(11), 7063–7070.
- Sloan, L. L. (1959). New test Charts for the Measurement of Visual Acuity at far and Near Distances*. *American Journal of Ophthalmology*, *48*(6), 807–813.
- Smith, J. E. K. (1982). Recognition models evaluated: A commentary on Keren and Baggett. *Perception & Psychophysics*, *31*(2), 183–189.
- Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, *11*(5), 13.
- Treacy, M. P., Hurst, T. P., Conway, M., Duignan, E. S., Dimitrov, B. D., Brennan, N., & Cassidy, L. (2015). The early treatment in diabetic retinopathy study chart compared with the tumbling-E and Landolt-C. *Ophthalmology*, *122*(5), 1062–1063 e1061.
- Westheimer, G. (2003). Visual acuity with reversed-contrast charts: I. Theoretical and psychophysical investigations. *Optometry and Vision Science*, *80*(11), 745–748.



Revealing the influence of bias in a letter acuity identification task: A noisy template model

Mark A. Georgeson^{a,b,*}, Hatem Barhoom^{b,c}, Mahesh R. Joshi^b, Paul H. Artes^b, Gunnar Schmidtman^b

^a School of Life & Health Sciences, Aston University, B4 7ET, UK

^b Eye & Vision Research Group, School of Health Professions, University of Plymouth, PL4 8AA, UK

^c Islamic University of Gaza, P.O. Box 108, Gaza, Palestine

ARTICLE INFO

Keywords:

Visual acuity
Letter recognition
Bias
Noisy template model
Sloan letters

ABSTRACT

In clinical testing of visual acuity, it is often assumed that performance reflects sensory abilities and observers do not exhibit strong biases for or against specific letters, but this assumption has not been extensively tested. We re-analyzed single-letter identification data as a function of letter size, spanning the resolution threshold, for 10 Sloan letters at central and paracentral visual field locations. Individual observers showed consistent letter biases across letter sizes. Preferred letters were named much more often and others less often than expected (group averages ranged from 4% to 20% across letters, where the unbiased rate was 10%). In the framework of signal detection theory, we devised a noisy template model to distinguish biases from differences in sensitivity. When bias varied across letter templates the model fitted very well - much better than when sensitivity varied without bias. The best model combined both, having substantial biases and small variations in sensitivity across letters. The over- and under-calling decreased at larger letter sizes, but this was well-predicted by template responses that had the same additive bias for all letter sizes: with stronger inputs (larger letters) there was less opportunity for bias to influence which template gave the biggest response. The neural basis for such letter bias is not known, but a plausible candidate is the letter-recognition machinery of the left temporal lobe. Future work could assess whether such biases affect clinical measures of visual performance. Our analyses so far suggest very small effects in most settings.

1. Introduction

Observers in sensory and perceptual tasks typically make decisions about the presence, absence or identity of target stimuli presented under conditions of uncertainty. The aim of such experiments is often to characterize the limits of performance of a human sensory system, such as vision or hearing, by finding the weakest, quietest or smallest stimulus that can still be reliably detected, or the smallest difference between two similar stimuli that can be reliably resolved. In optometric practice, clinicians routinely measure visual acuity, i.e. the ability of a patient to discriminate the smallest details of a stimulus (optotype). Thus, the assessment of vision in clinical optometry is essentially a psychophysical experiment. The conceptual framework for psychophysical experiments of this kind was revolutionized in the 1950s and early 1960s by the application of Signal Detection Theory (SDT) to the interpretation of human sensory performance (Green & Swets, 1966; Swets, 1961; Swets,

Tanner, & Birdsall, 1961; Tanner, 1956; Tanner & Swets, 1954). One of SDT's lasting contributions was to enable a formal distinction between two key components of performance - sensitivity and bias. Despite this conceptual revolution, SDT has not often been used explicitly to interpret letter acuity performance, or to examine the role of bias in clinical letter acuity measurements. In this paper, we apply SDT using a fairly simple template model for letter recognition, and we quantify the degree of bias shown by individual observers and in group performance.

Many previous studies on the role of bias in letter identification tasks have used versions of Luce's choice model (Luce, 1963), but these studies did not interpret or investigate the role of bias in visual acuity measurements, and often assumed that bias had little effect on the measured visual acuity using letters as optotypes (Candy et al., 2011; Coates, 2015; Hamm et al., 2018; Barhoom et al., 2021). Before SDT, the dominant idea ('high-threshold theory') was that the observer correctly detected the target or target difference if it was above threshold and

* Corresponding author at: School of Life & Health Sciences, Aston University, B4 7ET, UK.
E-mail address: m.a.georgeson@aston.ac.uk (M.A. Georgeson).

<https://doi.org/10.1016/j.visres.2023.108233>

Received 12 August 2022; Received in revised form 13 March 2023; Accepted 15 March 2023

Available online 2 May 2023

0042-6989/© 2023 Elsevier Ltd. All rights reserved.

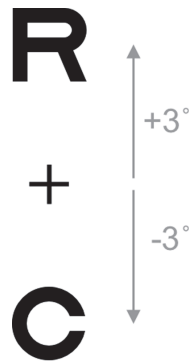


Fig. 1. Sloan letters presented singly, either centrally or along the vertical meridian at an eccentricity of 3 in the upper (+) or lower (-) visual field. Sloan letters R and C are shown for illustration purposes (not to scale).

therefore 'seen', and if not seen then the observer guessed. This may be intuitively plausible, but according to SDT it is fundamentally incorrect. With the arrival of SDT, the observer was held to make observations from one or more noisy mechanisms that responded to the stimulus, and then to employ a decision rule that mapped these observations onto a choice about which external stimulus event was most likely to have occurred. This transition from a focus on subjective events ('seen' vs 'not seen') to objective events ('which stimulus occurred?') is of crucial importance. Because it excludes direct appeals to consciousness, SDT is equally applicable to information-processing tasks carried out by humans, other animals, or machines. It is a theoretical framework, growing out of information theory, within which more specific models of detection, discrimination and recognition can be constructed. A key contribution made by SDT was to enable a distinction between discriminability (determined by the signal to noise ratio in the system) and bias (which varied with the decision rule adopted). It is often assumed that bias in such tasks is due to cognitive decision bias or behavioural response bias, also referred to as criterion shift – but it is likely that biases can also arise earlier within the sensory/perceptual mechanisms themselves. For example, perceptual aftereffects such as the tilt aftereffect and motion aftereffect can be viewed as biases temporarily induced within early coding mechanisms (e.g. Morgan, 2014; Storrs, 2015).

Here we re-examine a dataset on letter acuity (Barhoom et al., 2021; with two additional participants) to reveal *prima facie* evidence for systematic biases in letter judgements, and then formulate a simple model of letter recognition that allows us to quantify the degree of bias shown by individual observers, and by the average observer. We refer to this model as the *noisy template model*, and we test the idea that bias in letter choice arises from shifts in the baseline response level of letter-detecting mechanisms (templates). An upward (or downward) shift in the baseline for a given template makes it more (or less) likely to signal the presence of its preferred letter.

2. Methods

2.1. Participants

Ten naïve subjects (seven females; mean age 23.8 ± 4.4 (SD), age range: 19–32 years) with normal ocular health participated in the study. The mean best-corrected visual acuity and the mean refractive error (spherical equivalent) were -0.05 ± 0.06 logMAR and -2.5 ± 2.3 DS

respectively. All experiments were conducted monocularly (left or right eye, chosen at random). The fellow eye was occluded using an opaque eye patch. Written informed consent was obtained from all observers, and the study was approved by the University of Plymouth Ethics committee. All experiments were conducted in accordance with the Declaration of Helsinki.

2.2. Apparatus

The stimuli used in the experiment were generated using *Matlab* R2016b (MathWorks, Natick, Massachusetts, USA). Functions from the Psychtoolbox-3 were used to present the stimuli (Brainard, 1997; Kleiner et al., 2007). Stimuli were presented on a gamma-corrected Dell P2317H LCD monitor (1920 × 1080 pixels) with a frame rate of 60 Hz. Room illumination was 160 lx and viewing distance was 350 cm. At this distance one pixel subtended 0.258 min of arc ('). The observer called out his/her responses which were then entered by the experimenter via a computer keyboard. This method minimised mistyping and improved fixation compliance.

2.3. Stimuli

Stimuli in the experiment were Sloan letters (C, D, H, K, N, O, R, S, V, Z; black letters of 2.2 cd/m^2 on a white background of 215 cd/m^2 , resulting in 99% Weber contrast). The letters were presented centrally and at paracentral locations along the vertical meridian at an eccentricity of 3' in the upper (+) and lower (-) visual field (Fig. 1). Multiple pilot experiments were conducted to establish appropriate stimulus levels (letter sizes) to cover the whole range of responses from guessing (10% correct) to certain decision (100% correct). Six different letter sizes (always defined by their stroke width, and spaced logarithmically) were tested; 0.3', 0.44', 0.64', 0.94', 1.37', and 2' stroke width for central presentations and 0.5', 0.79', 1.26', 1.99', 3.15' and 5' for paracentral presentations.

2.4. Procedure

We used the method of constant stimuli in all experiments, with a single letter per trial. The letter location (1 of 3) and letter identity (1 of 10) were chosen randomly on each trial. Each subject completed 1800 trials for the full experiment (six letter sizes × three locations × 10 Sloan letters × 10 trials per letter). All conditions were interleaved. On each trial the stimulus was presented for 250 ms, accompanied by an auditory signal. The task was to recognise the presented letter and to report it verbally. Subjects were asked to fixate on a fixation cross (dimensions: length/width 1.55', stroke width 0.516') presented at the centre of the screen. The fixation cross was presented only in trials that tested the paracentral locations. Only choices from the 10-letter set were accepted. In rare cases where observers responded with other letters, the experimenter prompted for a second response. If the observer failed the second attempt, a reminder of the Sloan letter set was provided (this occurred very rarely, on average not more than once per subject). To familiarise the participants with the Sloan letters, the experimenter demonstrated the Sloan letters at the beginning of the session. Observers showed excellent compliance in responding from the Sloan letter set (<30 errors per subject in 1800 trials).

3. The noisy template model

3.1. Letter usage

If observers were equally sensitive to all letters in the 10-letter Sloan set and exhibited no biases to favour some letters more than others, then all 10 letters would be chosen as a response (correctly or not) with equal expected frequency, namely 10% or 0.1. We shall refer to the relative frequency of letter choice (averaged over the six letter sizes) as *letter*

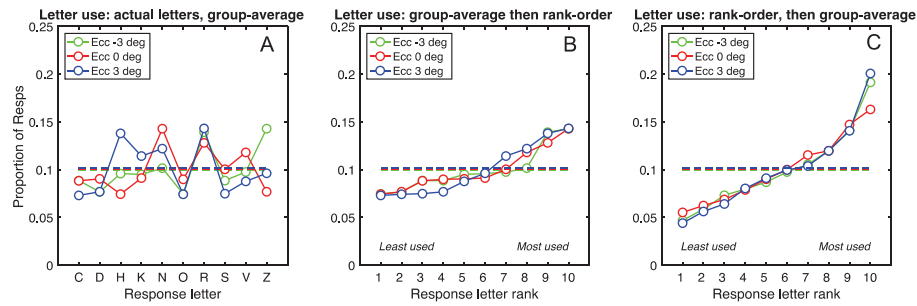


Fig. 2. Data overview: letter usage. A) Proportion of trials in the experiment on which each of the 10 test letters was reported (correctly or not), averaged over the 6 letter sizes and 10 observers. Each point is based on 600 trials. Colours indicate the 3 test locations. B) Same data as A, but rank-ordered by group-average frequency of use, separately for the 3 locations. Because the ordering of letters was different for the 3 locations, we must label the x-axis in terms of letter rank (1–10) rather than letter identity. C) Similar to B, but the usage frequencies were rank-ordered separately for each observer, then averaged over observers.

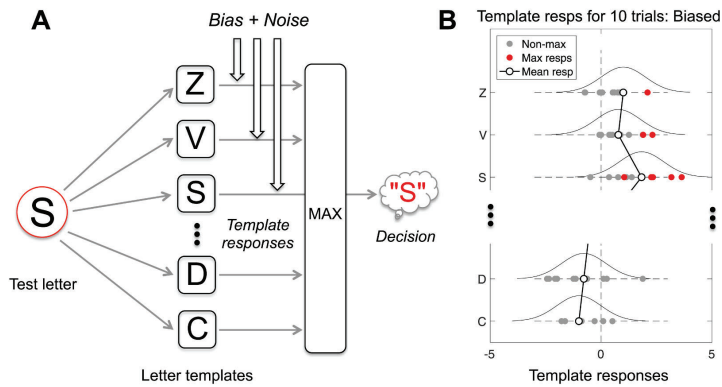


Fig. 3. A) Schematic view of the noisy template model. The most active template on a given trial determines the letter choice made. B) Because of noise, the most active template over trials (red dots) may be the correct one (e.g. S) or an incorrect one (e.g. V). Positive bias (top 3 rows) increases the chance of incorrect responses (red dots for Z or V) and correct responses (S). Negative biases (bottom two rows) decrease the chance of these letters being called, correctly or not. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

usage, and we expect the letter usages to deviate from 0.1 when differences in bias or sensitivity exist across the set of letters. However, several serious issues need to be resolved before any inferences about bias can be made, as Fig. 2 illustrates.

Fig. 2A shows the pattern of letter usage averaged across observers, for each of the three test locations in the experiment. In central vision (red) N, R, V appear to be over-used, while at -3° eccentricity (green) it was R and Z, and at $+3^\circ$ eccentricity (blue) it was H, N, R. Fig. 2B shows the same data, but rank-ordered in terms of group-average usage, from least to most. These variations in response proportions might reflect different biases for different letters. But if observers have different patterns of bias across letters, then averaging over observers before rank-ordering is likely to underestimate the range of individual biases – washing out the very effect we are aiming to capture. A potential answer to this problem is to rank-order the usage data of individuals first, and then average the ranked proportions across observers, as shown in Fig. 2C. We used this form of averaging in all our analyses. Note how the range of variation around the unbiased mean (0.1) has approximately

doubled from Fig. 2B to 2C, and how similar the trends are for the three test locations.

A second problem, however, applies to both forms of averaging: the risk of treating noise in the data as signal. For example, suppose that in Fig. 2A the true usages were all 0.1 and the observed variation was all due to random effects (sampling noise) within and between observers. If that were so, then the trends created by re-ordering in Fig. 2B or C would be artefactual – creating signal out of noise. By modelling the process of response generation, we found that the false appearance of bias arose only when the number of trials was small, and the generating model had no biases. See Supplementary file, Section 2, Figs. S2, S3, for details. We conclude from many such simulations that sampling noise is not a major concern, for three main reasons. It does not generally imitate the effects of bias; its effects are small even with just 10 trials per condition (Fig. S4 A, B); and its effects are minimized when real biases are present (Fig. S6 A, B; Fig. S7 A, B).

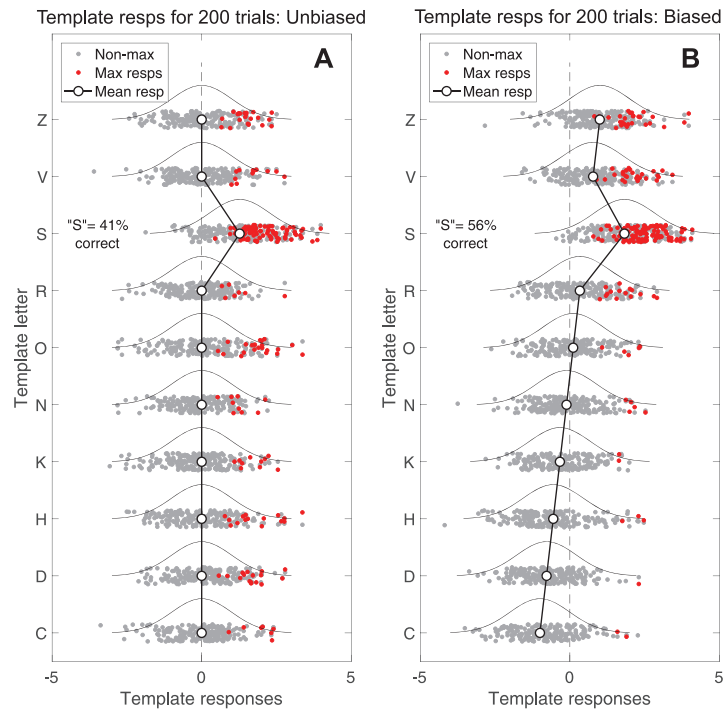


Fig. 4. How the noisy template model works, either without bias (A), or with different biases on each template (B). Bias in panel B is ordered from negative through to positive with a mean of zero. We call this the 'bias gradient', *B*. Letter identity for each template here is arbitrary, for illustration only. Presenting a test letter (e.g. S) increases the mean response for the S-template but leaves others unchanged. The letter decision on a given trial is made by choosing the template with the largest response on that trial (the 'MAX operator', Fig. 3A). Choices vary from trial to trial because of noise in each template channel (indicated by the Gaussian distribution curve). For illustration, red points represent those trials on which a given template gave the max response; grey points are activations that were lower, hence not chosen. Notice how positive bias (e.g. Z), and presentation of the preferred letter (e.g. S) increase the likelihood of choosing certain letters, sometimes correctly, sometimes not. We also tested for the presence of a sensitivity gradient (*S'*) across templates, in either the same or opposite direction to the bias gradient. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Outline of the model

We devised a model that allowed us to focus on the influence of bias in letter identification. We call it the 'noisy template model' (Fig. 3), and we make two simplifying assumptions about the templates. The first is that there are as many letter templates as there are test letters in the experiment (i.e. 10 templates), and secondly, the templates are assumed to give a response only to their own preferred letter, with no response to other letters (Fig. 4A); that is, the templates are orthogonal. Macmillan & Creelman (2005, p.246) noted that the assumption of orthogonality between detectors is 'optimistic' but appealing because of the simplicity it confers on the modelling. Two key features in our model are that the output of each template is perturbed by additive Gaussian noise and is subject to bias: the mean output level may be shifted up or down by a constant amount that is the same across trials and is the same whether the template's preferred letter is present or not. Crucially however, the bias may vary between templates (Fig. 4B). Our goal in this paper is to use the model to discover whether such bias exists in letter acuity tasks, and if so to determine how strong it is, and whether it varies with letter

size. One possibility is that bias might be present for all letter sizes; another is that it could be restricted to small letter sizes where subjective uncertainty is high. Another, as we shall see, is that both may be true, depending on precisely how the term 'bias' is interpreted - as a feature of the data, or a feature of the model.

The mean activity level of a template is raised not only by positive bias, but also when the preferred letter is present. But unlike bias, the letter-evoked response goes away when the letter is not the preferred one. For simplicity, the model does not further specify the nature of the template-matching process, but a detailed model for that process, incorporating optical and neural filtering, along with the limitations imposed by spatial sampling and noise, was developed by Watson & Ahumada (2015). They noted that all image-based models of letter identification to date are based on the template concept. For a brief review of letter recognition models in cognitive psychology and neuroscience, see Grainger, Rey, & Dufau (2008), and in the broader context of word recognition and reading, see Dehaene et al. (2005).

3.3. Model structure & equations

Suppose that the templates can be rank-ordered from least- (most-negative) bias to most-positive bias, indexed by $i = 1$ to m (where $m = 10$). We make the strong but simplifying assumption that bias values B_i are a linear function of i , ranging from $B_1 = -B'$ to $B_m = B'$. With no prior information, this linear gradient seems a good place to start because it greatly reduces the number of free parameters per subject [one bias parameter (B') per location instead of nine], and we shall see that the group-average data support this linear assumption quite closely.

It follows that bias B_i for the i^{th} template is:

$$B_i = \frac{B'(2i - m - 1)}{m - 1} \quad (1)$$

where B' is a free parameter. When $B' = 0$ the system is unbiased. In similar vein, we allow for the possibility that template sensitivity differs between templates. We take this sensitivity to be formed from a baseline value S_0 , and a linear variation around that value, ranging from $S_0(1 - S')$ to $S_0(1 + S')$. Thus, the sensitivity of the i^{th} template to the j^{th} test letter is:

$$S_{ij} = S_0 \left(1 + \frac{S'(2i - m - 1)}{m - 1} \right) \text{ if } i = j, \text{ else } 0 \quad (2)$$

We shall refer to B' and S' as the *bias gradient* and *sensitivity gradient* respectively. When fitted to data, B' and/or S' were free parameters for individual subjects, and if not fitted then $B' = 0$ and/or $S' = 0$. S_0 was also a fitted parameter for each subject and retinal location, controlling the overall level of performance.

The most fundamental fact in acuity testing is that identification performance improves with increasing letter size. Our simple model describes this by supposing that a template's mean response μ_{ij} is an increasing function of the sensitivity S , bias B , and test letter size t (expressed as letter stroke width, in min arc):

$$\mu_{ij} = B_i + (S_{ij}t)^p \quad (3)$$

where p is a constant exponent of a power function relation between mean response μ and letter size t . The exponent p controls the slope of the psychometric function (proportion of correct trials vs test letter size), and from initial explorations of model and data we set $p = 2.5$ for central vision, and $p = 2$ at $\pm 3^\circ$ eccentricity. When $p = 2$ the signal (μ) underlying correct performance increases with the square of the letter size. We might interpret this as arising from physiological nonlinearities (e.g. the half-squaring model of V1 cell responses; Heeger, 1992), or because the template area that collects contrast information from the retinal image increases as the square of the (1-D) letter size. Note that for unstimulated templates (where $i \neq j$), $S_{ij} = 0$ (the orthogonality assumption) and so eq. 3 then reduces to $\mu_{ij} = B_i$: template responses to non-preferred letters depend only on bias and noise. We adopted the standard SDT assumption that the i^{th} template output is perturbed by additive zero-mean Gaussian noise with variance σ_i^2 , and that $\sigma_i = 1$ for all i .

3.4. Letter identification & the MAX operator

We assume that the observer's decision about which letter was shown on a given trial is determined by whichever template is most active (Fig. 2A). This winner-take-all process, often known as the 'MAX operator', has a long history in psychophysics and pattern recognition. Watson & Ahumada, (2015) used the MAX over templates as the decision rule for letter identification in their Neural Image Classifier model but did not address possible effects of bias. When all the templates are equally sensitive ($S' = 0$) and unbiased ($B' = 0$), the MAX rule is the optimal decision rule (Kingdom & Prins, 2010, p.172-3; Wickens, 2002, p.106-8). We think it remains a reasonable assumption even in the face of some bias and unequal sensitivity across templates. This amounts to

assuming that the decision-making apparatus has no information about variation in S_{ij} or B_i , and so cannot devise a better decision rule than the MAX rule. See DeCarlo (2012) for more information on additive bias terms (B_i , eq. 3) in the SDT analysis of *m*AFC (m alternative forced choice) experiments, and Ma, Shen, Dziugaite, & van den Berg (2015) for a wide-ranging critical discussion of the MAX rule in the context of different tasks.

To determine the effect of the MAX operator on letter decisions, we need to compute (for all i, j) the probability P_{ij} that the i^{th} template delivers a response to the j^{th} test letter that is larger than that of all the other templates. The probability of a *correct* response to the j^{th} letter is P_{jj} . This involves integrals that are difficult or impossible to solve analytically (Wickens 2002, p.108), hence we used numerical integration with a function written in *Matlab*, elaborated from equations and *R* code by Nadarajah & Kotz (2008). Our *Matlab* function (*M_stats_maxN.m*) is freely available from the *Journal of Vision* as [supplementary material](https://doi.org/10.1167/14.13.24) to a paper by Zhou, Georgeson, & Hess (2014) at <https://doi.org/10.1167/14.13.24>. For a given set of parameters (p, S_0, S', B'), the model that uses this function returns a matrix of stimulus-response probabilities for each letter size, from which proportions correct and letter usages were easily computed.

3.5. Model fitting

The model was fitted to data (proportion correct identification for each of 6 letter sizes and 10 letter identities) separately for each observer and each test eccentricity, using maximum likelihood - adjusting parameter values S_0, S', B' to maximize the log likelihood (*LL*) of the parameters given the data. Dropping the subscripts for brevity, we have

$$LL = \sum \{n \cdot \log(P_c) + (t - n) \cdot \log(1 - P_c)\} \quad (4)$$

where for each condition n is the observed number of correct responses in t trials, P_c is the model probability of being correct, the variables n, t, P_c range over the 10 test letters and 6 letter sizes, and the summation takes place over those 60 conditions. One must take care to avoid asking for the log of zero, and this was assisted by including a small lapse rate ($P_{\text{lapse}} = 0.01$). We define P_{lapse} as the small proportion of trials on which lapses of attention or memory, eye blinks, etc. cause the observer to gather no sensory data about the letter identity. On these lapse trials the observer is taken to make an unbiased guess with probability $1/m$ of being correct. In most trials (no-lapse; proportion $1 - P_{\text{lapse}}$) the model probability of being correct is P_c . Hence for a given condition the expected performance over no-lapse and lapse trials is the combined probability $P_c^* = P_c(1 - P_{\text{lapse}}) + (1/m)P_{\text{lapse}}$, and P_c^* replaces P_c in Eqn. (4).

The fitting was done in two phases: firstly, assuming no bias ($B' = 0$) and no variations in template sensitivity ($S' = 0$) we adjusted overall sensitivity S_0 to find the best-fitting value that maximized *LL*. Using the best S_0 , we then ran a finely sampled grid search to find the best values of the gradients B' and/or S' for each observer and test location. In this second run, the model templates were indexed in order of bias (from $-B'$ to B') and so it was essential that the experimental data for the 10 test letters had a corresponding order. Our best estimate of this correspondence for a given observer was the rank order of letter usage, averaged over letter sizes discussed above. Importantly, this ordering was *the same for all letter sizes*, did not depend on the correctness of the trial responses, and did not depend on any model parameters and so did not vary during the fitting.

We fitted four versions of the template model, in which the fitted parameters represented both bias and sensitivity gradients, denoted as model (B1 S1), or bias gradient only (B1 S0), sensitivity gradient only (B0 S1), or neither (B0 S0).

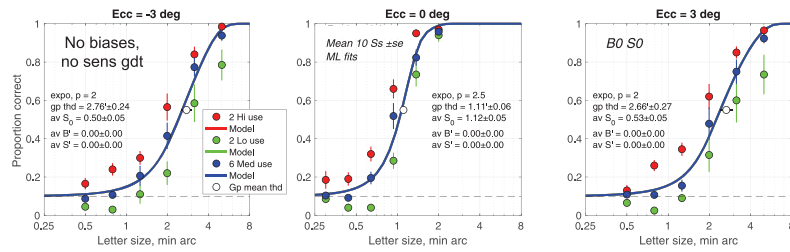


Fig. 5. Data overview: psychometric functions. Some letters are used as a response more than others, across all letter sizes. Red symbols plot the proportion of correct trials averaged over the two most-used letters, identified separately for each observer, as in Fig. 2C, then averaged over observers. Green symbols similarly show the results averaged for the two least-used letters. Blue symbols are averages over the 6 intermediate letters, and the blue curve is a fit of the default model (B0 S0). With no letter biases and no variation in sensitivity between letters, this model does not capture the large differences in performance between preferred and non-preferred letters (red vs green symbols). *Note:* in this paper, letter size is always defined by the letter stroke width, in min arc. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

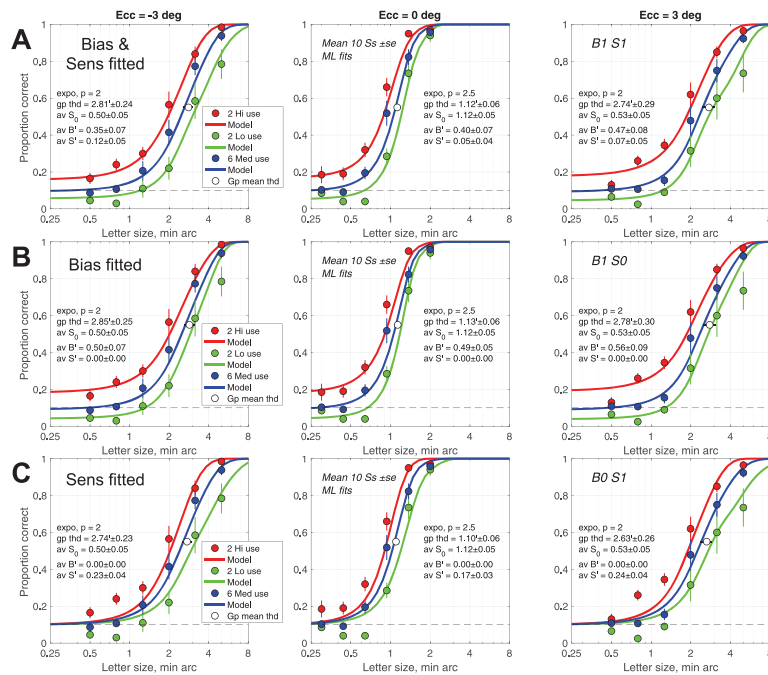


Fig. 6. The noisy template model was fitted to the whole dataset (row A) by choosing the best-fitting gradients of bias & sensitivity across the letter templates, or (B) by fitting the bias gradient while sensitivity was the same for all letters, or (C) fitting the sensitivity gradient, while bias and bias-gradient were zero. In each case the model was fitted to data of individual subjects; these graphs show the group mean data (\pm s.e.) and model curves averaged over observers. The results support the view that these acuity data reflect two factors in letter recognition: (i) observers have consistent biases both for and against certain letters, and (ii) they are a little more sensitive to some letters than others. See Tables 1 & 2 for supporting statistical analysis.

Table 1
Comparison of 4 models via AIC analysis, for the group of 10 Ss.

Location	Model	LL	K	n	AICc	DiffAIC	Ak. Wt.
-3°	B0 S0	-2549.44	11	600	5121.34	300.449	0
	B0 S1	-2416.46	21	600	4876.53	55.639	0
	B1 S0	-2412.81	21	600	4869.22	48.332	0
	B1 S1	-2377.70	31	600	4820.89	0	1
0°	B0 S0	-2455.85	11	600	4934.15	245.369	0
	B0 S1	-2358.50	21	600	4760.60	71.827	0
	B1 S0	-2338.19	21	600	4719.98	31.204	0
	B1 S1	-2311.64	31	600	4688.78	0	1
+3°	B0 S0	-2579.69	11	600	5181.83	310.365	0
	B0 S1	-2452.50	21	600	4948.60	77.139	0
	B1 S0	-2423.98	21	600	4891.56	20.104	0
	B1 S1	-2402.98	31	600	4871.46	0	1

Notation: LL, total log likelihood; K, total no. of free parameters across the group; n, total no. of data points.

Table 2
Group Analysis of Deviance, 10 Ss.

	ChiSq	df	P value
<i>(i) Model B0 S1 vs B1 S1</i>			
-3°	77.53	10	< 0.000001
0°	93.72	10	< 0.000001
+3°	99.03	10	< 0.000001
<i>(ii) Model B1 S0 vs B1 S1</i>			
-3°	70.23	10	< 0.000001
0°	53.10	10	< 0.000001
+3°	42.00	10	0.000008

4. Results

4.1. Psychometric functions

Across 10 subjects, 3 locations, and 10 test letters, we could in principle generate 300 psychometric functions, each with 6 test sizes and 10 trials per size per letter - from a total of 18,000 trials. Data reduction is needed. To summarize the evidence for letter biases in the experiment, we used the letter usage data to split the performance data for each subject into three groups: the top 2 (most used) letters, the bottom 2 (least used), and the remaining 6 letters with intermediate usage. Data were then averaged over subjects for each of these groupings and the results are shown by the red, green and blue symbols respectively in Fig. 5. Left, middle and right panels are for the three locations, as indicated. We emphasize that the actual letters contributing to each grouping (e.g. the red symbols) vary a good deal across observers; it is their usage ranks that correspond.

Fig. 5 shows that performance (defined by proportion correct) was strikingly and consistently higher for the high-usage letters (red symbols) than the low-usage (green), and this was true across all letter sizes and test locations. At small sizes, the least-used letters (green symbols) fell consistently below the chance level of 0.1. Note that below-chance performance is an inevitable consequence of bias, because favoring some letters with positive bias necessarily biases choices away from other letters. More surprising perhaps is that these differences existed over the entire psychometric function, even up to high performance levels from 80 to 100 % correct. Proportions correct were typically higher by 0.2 to 0.3 for the most-used than the least-used letters, except where performance for both saturated at 100 % correct (Fig. 5, center panel).

Predictions of the default model (B0 S0) that assumed no bias and no variation in sensitivity between letters are shown by the blue curves in Fig. 5. This model predicts the same psychometric curve for all 10 letters

(hence red & green curves are hidden behind the blue). It cannot describe the consistent difference in proportion of correct responses for high-usage and low-usage letters.

On the other hand, Fig. 6A shows that the full model (B1 S1) captured the pattern and magnitude of the differences in performance very closely, including the below-chance performance at all three locations. The ten model psychometric functions (one for each level of bias B_i) were grouped by bias values [$i = 1:2, 3:8, 9:10$] corresponding to the way the empirical data were grouped by usage [low, medium, high], and then averaged within each grouping and over the ten model observers to give the smooth curves seen in Fig. 6A. Unlike the default model (Fig. 5), these model curves gave an extremely close account of the data.

4.2. Comparison of models

In Fig. 6A we examined the model where both bias and sensitivity gradients were fitted. Both gradients emerged as positive (see text in each panel). This is reasonable, since both higher bias and higher sensitivity for (say) letter S will raise the mean response of the S template, and lead to more observed "S" responses. The difference is that bias induces more "Ss" when the letter is S and when it is not; higher sensitivity of the S template also gives more correct "S" responses but does not affect incorrect "S" responses, since (with the orthogonality assumption) an unbiased S template is silent (noise only) when the letter is not S. Fig. 4 may help the reader to visualize these relationships.

Which, then, is the greater influence in our data: bias or sensitivity variation? We addressed this key question in two ways: firstly, by fitting the four models (above), with and without the influence of B' or S' , and secondly, by fitting the full model (B1 S1) and then setting B' or S' to zero to find out which was the more influential factor in the full model.

Fig. 6B shows the result when only the bias gradient was fitted (B1 S0). The data are the same as Fig. 6A, and the model curves appear to fit almost equally well. The best-fitting bias gradient was steeper by about 30% at all 3 locations when sensitivity gradient was removed (see inset values), but more importantly bias gradient alone was able to capture most of the effects in the data. Closer examination reveals some deviation of the low-use curves from the data (green) at just two points, the highest performance levels in the -3° and +3° locations.

In strong contrast, when bias gradient was not fitted (B0 S1; Fig. 6C), the quality of fit was substantially worse: notably poor for the 3 smallest letters at all 3 locations, but about the same as model (B1 S1) for the 3 largest letters. The reason for the poor fit is clear: variation in sensitivity shifted the psychometric curves laterally on a log scale (because in model Eqn. 3 a change in sensitivity is equivalent to re-scaling the letter size) but it did not show the 'fingerprint' of bias - the above-chance and below-chance asymptotes at small sizes shown by the biased model, and by the data.

At the smallest letter sizes, data at all 3 locations appear to converge back towards the chance level (0.1). This (fairly small) transition from bias to no-bias is not predicted by our current versions of the template model (Fig. 6).

We compared the four models, using AIC (Akaike Information Criterion; Akaike, 1974), either at the level of individual subjects or at the group level. In brief, AIC allows one to evaluate the relative strength of evidence across a set of candidate models fitted to the same dataset, no matter whether the models are nested or not; see Burnham & Anderson (2002) for a full account. At the individual level, no clear result emerged from AIC, most likely because for individual subjects there were too few trials, hence too little power. However, at the group level, Akaike weights (Burnham & Anderson, 2002, chapter 2; Wagenmakers & Farrell, 2004) were 1 for the (B1 S1) model and 0 for the other three, at all three test locations (Table 1). This means that the strength of evidence (Akaike weights) exclusively favoured the two-factor model (B1 S1) over the other three.

To confirm this conclusion, we also conducted an analysis of deviance for nested models (i) B0 S1 vs B1 S1, and (ii) B1 S0 vs B1 S1. This

Table 3
Comparison of 3 models via AIC analysis, for the group of 10 Ss.

Location	Model	LL	K	n	AICc	DiffAIC	Ak. Wt.
-3°	B0 S0	-2549.44	11	600	5121.34	252.117	0
	B0 S1	-2416.46	21	600	4876.53	7.307	0.025
	B1 S0	-2412.81	21	600	4869.22	0	0.975
0°	B0 S0	-2455.85	11	600	4934.15	214.165	0
	B0 S1	-2358.50	21	600	4760.60	40.623	0
	B1 S0	-2338.19	21	600	4719.98	0	1
+3°	B0 S0	-2579.69	11	600	5181.83	290.262	0
	B0 S1	-2452.50	21	600	4948.60	57.035	0
	B1 S0	-2423.98	21	600	4891.56	0	1

tests whether the extra free parameter (in B1 S1) yields a significant improvement in the fit of the model, thereby justifying its inclusion. If $D1$ is the deviance for the nested (more restricted) model, and $D2$ the deviance for the more general model (B1 S1), then for each subject $ChiSq = D1 - D2$ with $df = 1$ (Collett, 2003, pp. 65-73). Table 2 applies this analysis to the group by summing $ChiSq$ and df over subjects. In both comparisons, at all 3 locations, the outcome was highly significant, confirming the AIC analysis, and implying that gradients of bias and sensitivity across letter templates were both influential in the fit of the two-factor model (B1 S1).

The finding that bias and sensitivity gradients both play a significant role naturally prompts the question whether one of them is more important than the other? It seems clear by eye that fitting only the bias

gradient (Fig. 6B) still gave a good account of the data, while fitting the sensitivity gradient alone did not (Fig. 6C). We tested these 1-factor models against each other using AIC again (because these two models are not nested) and removing the two-factor model (B1 S1). Table 3 shows a clear outcome: the model with bias gradient (B1 S0) had Akaike weights at or close to 1, and so was exclusively favored over the two models that had no bias gradient (B0 S0, or B0 S1).

In summary, differences in both bias and sensitivity were found to play a role in determining how often different letters were correctly reported in this acuity/identification task (Tables 1 & 2). But, taken separately, letter biases accounted for the data more closely than differences in letter visibility did (Table 3). These two findings suggest that when both gradients are present in the best-fitting model (B1 S1) bias will play a bigger part than sensitivity variation. Clear support for this is seen in Fig. 7, where we took the best-fitting model (panel A) and compared the effects of removing the sensitivity gradient ($S'=0$) or the bias gradient ($B'=0$) without re-fitting any other parameters. The goodness of fit was clearly much better when the bias gradient alone was present ($S'=0$; Fig. 7B) than when only the sensitivity gradient was present ($B'=0$; Fig. 7C), showing that bias contributed a good deal more than sensitivity variation to the original two-factor fit (Fig. 7A).

To visualize the main features of the best-fitting model, Fig. 8 (top row) plots the linear trend of bias B_i across templates, for the group average and for individual observers (thin lines). The agreement across observers is reasonable; 29 of 30 model slope estimates (B') were

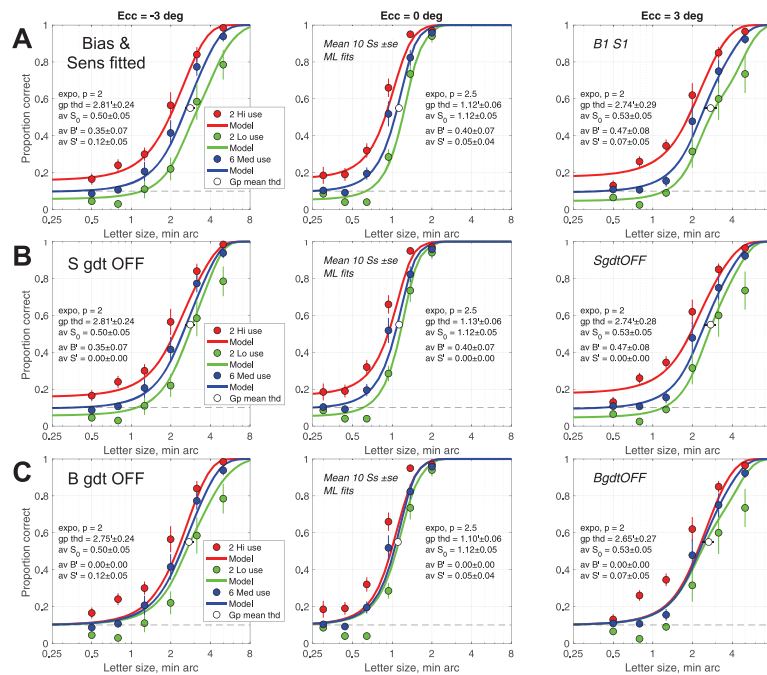


Fig. 7. How important were the B gradient and S gradient in the best-fitting two-factor model? (A) Fit of the two-factor model, copied from Fig. 6A. (B) The model was re-computed with B gradient unchanged, but the S gradient set to 0 (i.e. same template sensitivity S_0 for all letters). (C) The model was re-computed with S gradient unchanged, but the B gradient set to 0. Comparison of row B with C implies that bias (in row B) was the major contributing factor, while variation in sensitivity across letters (row C) played a much smaller role, mainly at the larger letter sizes in the peripheral viewing conditions.

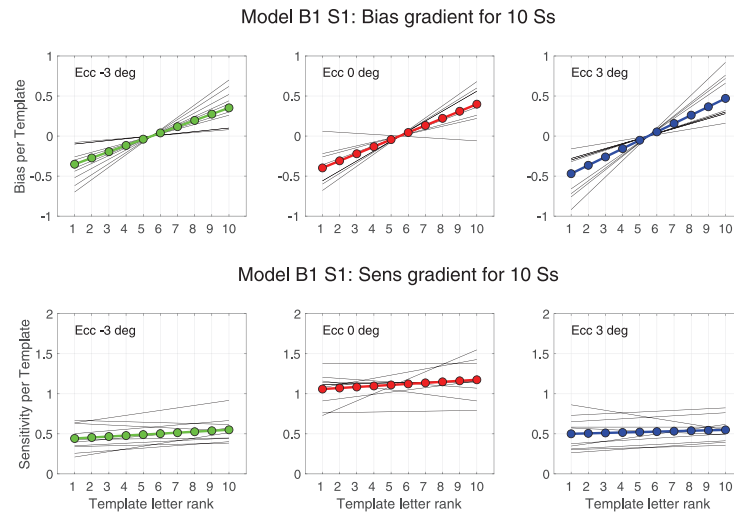


Fig. 8. Variation in bias and sensitivity across letters in the full model (B1 S1). Top row: Average linear trend in modelled bias (coloured symbols) was similar across the 3 test locations. Ordinate plots bias values B_i , for $i = 1$ to 10, ranging from about -0.4σ to $+0.4 \sigma$, where $\sigma = 1$ is the template noise standard deviation. Thin black lines are fitted model slopes for individual subjects. Lower row: analogous plots for template sensitivity S_{ij} rather than bias.

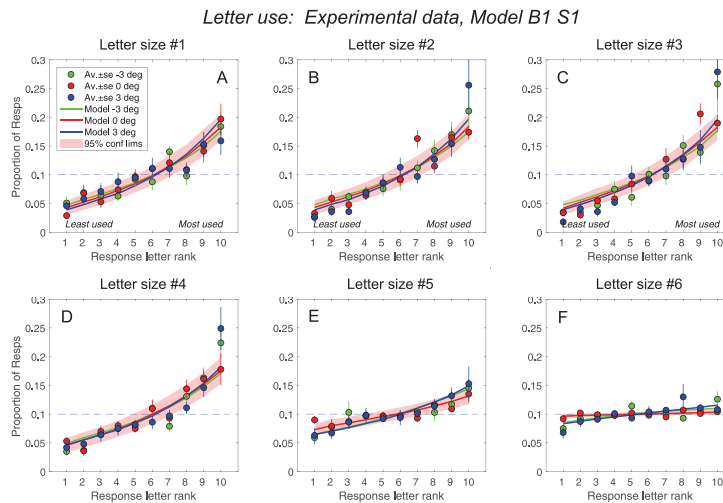


Fig. 9. Data & model compared via letter usage, separated out according to letter size. (A - F) Each panel relates to one of the 6 test letter sizes, from smallest to largest, as marked. Ordinate plots proportion of trials a given letter was named, whether correct or not. Symbols are group-average data ± 1 s.e. Smooth curves are the average of 200 independent Monte Carlo, trial-by-trial simulations, with 10 Ss and 10 trials per letter, as in the real experiment. Pink band represents 95% confidence region (± 2 SD) on model values for the 0° eccentricity condition (red curve). For all six sizes this model confidence band nicely embraces the means and scatter of experimental data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Letter recognition thresholds for 4 models and 3 eccentricities.

Model	Ecc = -3 deg	Ecc = 0 deg	Ecc = +3 deg
B0 S0	2.76 ± 0.24	1.11 ± 0.06	2.66 ± 0.27
B0 S1	2.74 ± 0.23	1.10 ± 0.06	2.63 ± 0.26
B1 S0	2.85 ± 0.25	1.13 ± 0.06	2.78 ± 0.30
B1 S1	2.81 ± 0.24	1.12 ± 0.06	2.74 ± 0.29

Table legend: Table entries are group mean letter-size thresholds (letter stroke widths, in min arc ± s.e., N = 10 observers). For each model fitted to the data, model psychometric functions for each subject and letter identity were averaged over letter identity, and the threshold size for letter recognition was interpolated at 55 % correct (halfway between chance (10 %) and perfect (100 %) performance). These thresholds and model psychometric functions were then averaged over subjects (see supplementary Fig. S1). The group-mean threshold values in this table are also shown as white symbols in Figs. 5 and 6.

positive. Similarly, Fig. 8 (lower row) plots sensitivity S_{ii} (Eqn. (2)) for each template. Mean sensitivity (S_0) was about twice as high for the central location (red), in line with the superior acuity centrally, while the average sensitivity gradient was positive but shallow at all three locations. Relative sensitivity (expressed as S_{ii}/S_0) varied rather little across templates, by ± 11%, ± 5.2% and ± 4.7% at the -3, 0 and + 3° locations. In the model, the impact of varying sensitivity is equivalent to varying letter size, and so these shallow sensitivity gradients are equivalent to only ± 5 to 10% variation in letter size. This reflects the weak (but not negligible) contribution that sensitivity variation made to the full model's description of the psychometric function data.

4.3. Letter usage decomposed across letter sizes

The rank-order of response letter usage (used as a proxy for the template ordering) was derived from the overall usage of each letter (pooled across all sizes, as in Fig. 2C). Here (in Fig. 9) we examine letter usages separated according to letter size, and we ask how well the model (B1 S1) predicts these usage data. For a given subject the rank-order of response letters was derived from overall usage, and so was the same for all 6 sizes, but the order of actual letters would differ across subjects according to their usage. Predictions of letter usage grouped by letter size (curves in Fig. 9) were derived from the model fitted to psychometric data (as in Fig. 6A). But the usage results here (symbols in Fig. 9) are not a trivial replotting of the proportion-correct data for two important reasons. Firstly, the dependent measures (usage data vs psychometric data) are different. Usage data are concerned with what responses were made but not with their correctness, while the psychometric data are concerned with the identity and correctness of responses. Secondly, the predictions here for usage were generated from a trial-by-trial Monte Carlo version of the model, specifically so that the combined effects of both bias and sampling variation could be observed. Fig. 9 shows that the biased template model predicts very well the variation in usage across letters, for each size, even though data in this form were not used to fit the model. It also describes the decrease in slope of that trend at the two largest letter sizes. As letter size increases, the data and model inevitably converge onto a constant (0.1) level of usage for each letter (bottom right panel). That happens because in the limit, at 100 % correct, there is no opportunity for 'over-calling' any letter. The response letters always match the stimulus letters, and those are presented equally often, thus enforcing a usage rate of 0.1 per letter. Fig. 9 reveals that these large variations in usage with letter size, and the upward curvature of the usage profiles, are (perhaps surprisingly) consistent with two important simplifying assumptions that (i) the bias on each model template has the same fixed value for all letter sizes and (ii) the bias terms vary in a linear fashion across the templates (Fig. 8).

Is a linear gradient of bias values across templates (Fig. 8) supported by data at the level of individual observers? We reasoned that if the model with a linear gradient of bias captures an observer's usage data with sufficient accuracy, then the linear assumption is supported for that

observer. We plotted the usage profiles of all individuals and compared them with the model fits. This analysis is described in the [Supplementary File](#), Sec. 4, Figs. S8, S9. In summary, for 24/30 cases (80%) a linear bias gradient was sufficient to account for the observed usage profiles, while in the remaining 6/30 cases (20%) deviations between model and data implied that additional factors came into play. We identified a possible source of additional bias for 3 of the 6 deviant cases (Fig. S9), but these cases are relatively rare and, until more is known, we favour the simplicity and parsimony of the linear assumption.

4.4. Does letter bias affect estimates of letter acuity?

Table 4 shows group mean estimates of acuity aggregated over the 10 different letters. It reveals that acuity thresholds for pooled-letter performance showed very small differences when estimated by four different psychometric models that took account of letter biases (B1) or did not (B0).

Overall, the noisy, biased template model for the 10AFC letter recognition task has given strong evidence that substantial letter biases, consistent within-subject but different between-subjects, do occur in letter acuity tasks. But taking account of the biases, or ignoring them, made almost no difference to acuity values derived from the aggregate performance data. Hence, in clinical assessment of acuity, the analysis in Table 4 implies that such biases should not greatly affect estimates of acuity obtained from pooled-letter performance.

5. Discussion

5.1. Modelling letter bias in acuity

In this paper we developed a fairly simple model, within the framework of SDT, that allowed us to infer the extent and impact of letter biases on letter recognition performance in a single-interval 10AFC task. The essence of the model is that

- there is a dedicated detector (*template*) for each of the 10 possible letters;
- the template's response is noisy but its mean increases with letter size;
- each template's mean response is also shifted, up or down, by a fixed, additive, bias term that varies across letters but does not vary with letter size;
- the observer's decision rule on each trial is to report the letter whose template gave the largest output on that trial (cf. DeCarlo, 2012).

When a linear range of biases was imposed over the templates, this form of model gave an excellent account of the data, both when the data were expressed as the proportion of correct responses (Fig. 6A, B), and as the proportion of trials on which different letters were used as a response, whether correct or not (Fig. 9). Bias was the key factor: the model fitted poorly when model biases were absent (Fig. 6C, 7C, S4, S5).

To simplify analysis, we adopted two key assumptions. First, for each observer the rank-order of letters, from least- to most-biased, was given by *letter usage* - the frequency with which each letter was used as a response (summed over all letter sizes at a given test location, and rank-ordered from least- to most-used). Second, we assumed that the bias term for each letter was a linear function of its usage rank. The slope of this function (the *bias gradient*) was a free parameter allowing for the possibility that bias was zero, or even negative (meaning that bias would decrease as letter usage increased). Both assumptions were well supported by the data. The linear bias gradient was seen to result in (i) an upward curvature of the letter-usage profiles that fitted the data very well, and (ii) a decrease in the slope of these profiles with increasing letter size (Fig. 9). Thus, the template biases were the same for all letter sizes, but they had decreasing impact on the data as letter size and visibility increased.

5.2. Template bias order parallels letter usage?

We estimated bias order for *individual* observers to avoid under-estimation of bias that would arise by pooling data across observers whose bias orders were different. But these individual usage profiles might introduce the opposite risk of over-estimating bias by converting random variations in usage into (apparently) systematic variation in bias. We used the model as a tool to address this risk, by comparing Monte Carlo simulations (which simulated this sampling-noise artefact) with noise-free calculations (that avoided the artefact). We concluded from several detailed analyses (see [Supplementary file](#)) that an artefact of this kind can be induced in the rank-ordered data but is generally small with the number of trials that we used (10 per condition) and diminished even further when actual biases were introduced. With the bias gradients that accounted for our data, the artefact was negligible.

5.3. Are letter biases consistent across observers and/or eccentricities?

It is natural – but not essential to our model – to ask (i) whether different observers showed similar patterns of bias for or against particular letters, and (ii) whether patterns of bias were consistent across test locations. Appendix A addresses these questions quantitatively, and we summarize the conclusions here:

(i) There is a statistically significant but mostly rather weak similarity shown by pairs of observers (mean rank correlation 0.21) in the patterns of letter choice, and hence in the likely biases ([Fig. A.1](#)).

(ii) Patterns of letter bias (as reflected in patterns of letter usage) can be highly reliable across test locations for some observers. But for most observers such consistency was weaker, and sometimes absent ([Fig. A.2](#)).

Given this variability across individuals, it is hard to draw general conclusions about theoretical issues such as the stability of template bias, or whether the main source of bias resides at the template level. However, we should recall that in developing the biased template model we envisaged that observers would show different patterns of bias – which we have confirmed here – and we devised the method of ranking letter usage for each observer to circumvent this issue and to allow averaging of data across observers in a way that preserves individual biases and does not average them out.

5.4. Does bias influence acuity estimation?

In visual acuity tests researchers and clinicians are usually interested in estimating the resolution threshold. However, the estimated resolution threshold is potentially contaminated by biases towards or against some of the letters ([Yeshurun et al., 2008](#); [Sridharan et al., 2014](#); [Jogan & Stocker 2014](#)). The impact of such bias could be to alter estimates of sensitivity if the bias is not properly accounted for. Nevertheless, bias has often been assumed to have a minimal effect on estimated thresholds using Sloan letters, but without directly investigating its effect on the task ([Alexander, Xie & Derlacki, 1997](#); [Hamm et al., 2018](#); [Barhoom et al., 2021](#)). Our analysis ([Table 4](#) and [Fig. S1](#)) showed that when data were aggregated across all letter identities, it made little difference whether the threshold estimation procedure took account of bias or not. [Fig. 6A](#) and [6B](#) strongly imply that for models and data the substantial effects of positive and negative biases (plotted in red and green) will largely cancel when performance is aggregated over letters. Hence the presence or absence of bias should make little difference to aggregate performance either in experiments (as in [Table 4](#)) or in clinical testing. The biases are revealed by fitting models to individual letter performance, and by combining across subjects in a way ([Fig. 2C](#)) that prevents the bias effects from cancelling out in the group averages.

5.5. Cerebral origin of letter bias

Whether the bias is of a decisional or perceptual nature is unknown.

Because our model places the bias at the template level, *before* the MAX operator that makes a perceptual decision, we err towards an early perceptual level as the site of bias. Current models of letter and word recognition propose, from behavioural and brain-imaging evidence, a hierarchy of stages in the left prefrontal cortex and inferotemporal lobe by which letter fragments, then letter shapes, then bigrams (letter pairs), and then fragments of single words are represented ([Dehaene et al., 2005](#)). The letter templates in psychophysical models like ours might correspond to some intermediate letter-level in this sequence. Given the ubiquity of feedback from higher to lower areas in the brain, we speculate that the biasing signals could arise at higher levels of the hierarchy, perhaps influenced by context, expectation and learning, and then be fed back to the templates themselves. Such bias would effectively be a combination of both decisional and perceptual biases ([Linares et al., 2019](#); [Rahnev, 2021](#)).

5.6. Conclusion

Our strong conclusion from this re-analysis and modelling of an extensive dataset ([Barhoom et al., 2021](#)) is that a gradient of biases across letter templates accounts strikingly well for the variation in letters that people choose, and for the pattern of variation in correctness with which they choose them.

CRedit authorship contribution statement

Mark A. Georgeson: Conceptualization, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Hatem Barhoom:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Mahesh R. Joshi:** Conceptualization, Methodology, Writing – review & editing. **Paul H. Artes:** Conceptualization, Methodology, Writing – review & editing. **Gunnar Schmidtmann:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We thank an anonymous reviewer for prompting us to consider the data of individual subjects in more detail ([Appendix A & Supplementary file, Sec. 4](#)).
Declarations of interest: None.

Appendix A

A.1. Are patterns of letter bias similar across observers?

In this paper, we have treated variations in letter usage as an empirical indicator of bias, so this question amounts to asking whether different observers had similar rank orders of letter usage when actual letter identities were considered.

Following our definition of *usage*, response counts for a given letter were pooled over all 6 letter sizes, without regard to correctness of the responses, and then the set of 10 letters was ranked to give the usage ranks shown for each test eccentricity and observer in the upper half of [Table A.1](#). For each observer, rank 1 denotes the letter least-used as a

response; rank 10 denotes the most-used letter.

Judging the similarity of these ranking patterns (Table A.1) by eye is impossible. To quantify similarity, we therefore calculated the rank correlations between pairs of observers, as follows. When each of the 10 observers was paired with each of the 9 others, this resulted in 45 distinct pairings. For a given test location (e.g. -3 deg), the similarity of letter response usage between any pair of observers (e.g. S1, S2) was given by the correlation of their two sets of ranks (e.g. [5 3 7 1 6 8 10 4 2 9] vs [2 4 3 8 5 1 10 6 7 9]). In this example, the rank correlation was low ($R = 0.0667$). In a second example (S5, S6) the correlation was fairly high ($R = 0.503$). The distributions of these correlation values are plotted in Fig. A.1.

Table A.1 shows the letter ranks in full and summarizes the rank correlations found for pairs of observers at each location. Mean and median correlations were significantly > 0 at each location ($P < 0.002$ for means, $P < 0.003$ for medians; see lower part of Table A.1). Mean correlations were low around 0.2, and medians lay between 0.2 and 0.3, but with a fairly large spread over positive and negative correlation values (as seen in Fig. A1). We conclude that there is a statistically significant but overall rather weak similarity in the patterns of letter choice, and hence in the likely biases, shown by different observers.

A.2. Are patterns of letter bias similar across locations?

Our second question was whether individual observers showed similar patterns of letter bias across the different test locations. The analysis was analogous to that above and used the same ranking data (upper part of Table A.1). For each observer, we computed the correlation of usage ranks but this time across pairs of test locations, for the 3 possible pairings [-3,0; 0,+3; -3,+3]. The boxplot in Fig. A.2 summarizes the distribution of these correlation values across the subject group, and shows the individual values as filled symbols.

For all three pairs of locations, the group mean and median correlations were significantly greater than zero (see Table A.2). Almost all (28/30) correlations were positive; 17/30 were > 0.5 . These findings imply a significant degree of similarity in patterns of letter usage across the three test locations. Fig. A.2 (right) suggests that the greatest similarity might be between the two eccentric locations (-3,+3), but the median for this case was not significantly different from the other two pairings (signed-ranks test for equal medians, $P = 0.125$, $P = 0.232$), and in a 1-way repeated-measures ANOVA (performed using the Matlab *ranova* function), the possible upward trend in Fig. A.2 (the main effect of the 'location pairing' factor) was not significant [$F(2,18) = 2.179$, P

Table A1
Usage Rank listed in original (alphabetic) letter order [CDHKNORSVZ].

	Ecc. = -3 deg	Ecc. = 0 deg	Ecc. = +3 deg
Response letter:	CDHKNORSVZ	CDHKNORSVZ	CDHKNORSVZ
Observer			
S1	5 3 7 1 6 8 10 4 2 9	4 2 6 6 9 7 8 1 10 3	7 1 7 3 8 4 10 5 2 9
S2	2 4 3 8 5 1 10 6 7 9	5 1 3 2 8 4 6 10 9 7	1 3 9 6 8 2 10 5 7 4
S3	2 1 7 6 3 5 8 10 9 4	4 3 8 2 6 7 9 10 6 1	1 3 10 7 8 6 9 4 5 2
S4	9 4 5 8 3 1 6 2 7 10	9 10 3 5 7 4 6 2 8 1	9 4 10 6 8 1 7 2 5 3
S5	3 2 4 8 10 1 9 7 6 5	5 2 3 8 10 1 9 8 6 4	6 2 5 10 9 1 8 4 3 7
S6	1 3 8 4 9 5 10 2 7 6	1 2 6 5 9 8 10 4 7 3	1 3 8 5 9 7 10 2 6 4
S7	10 9 7 5 2 3 5 1 6 8	6 7 3 2 9 10 4 8 5 1	8 9 5 2 3 6 4 1 7 10
S8	2 5 6 8 3 1 9 4 7 10	2 10 1 5 8 3 9 6 7 4	1 8 4 10 6 2 7 3 5 9
S9	9 1 3 4 6 8 10 7 6 2	7 2 4 6 10 3 9 5 8 2	4 3 7 9 6 1 10 5 8 2
S10	2 4 6 3 10 5 8 7 2 9	3 4 3 9 10 5 8 6 1 7	3 2 6 5 10 9 5 7 1 8
Summary of rank correlations between pairs of observers ($N = 45$ distinct pairs)			
Mean correl \pm S.D.	0.159 \pm 0.345	0.248 \pm 0.254	0.214 \pm 0.373
Mean is > 0 ? Yes	$t = 3.09$, $P = 0.0017$	$t = 6.54$, $P < 0.00001$	$t = 3.86$, $P = 0.00019$
Median	0.191	0.248	0.295
Median > 0 ? Yes	$Z = 2.74$, $P = 0.0030$	$Z = 4.85$, $P < 0.00001$	$Z = 3.27$, $P = 0.0005$
Min, Max	-0.528, 0.879	-0.277, 0.816	-0.588, 0.879

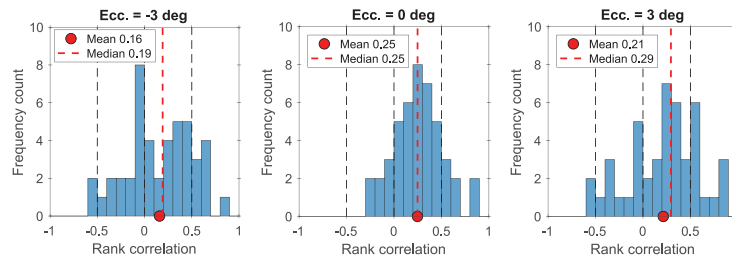


Fig. A1. For each of the three test locations, histograms show the distributions of between-subject similarity in usage (rank correlation values) for the 45 pairings of 10 observers. Red symbol shows the mean correlation, dashed red line the median. Dashed black lines mark correlation values of -0.5, 0 and +0.5. Means and medians were significantly above zero (Table A.1), but the spread of values was wide, especially at the eccentric locations (± 3 deg).

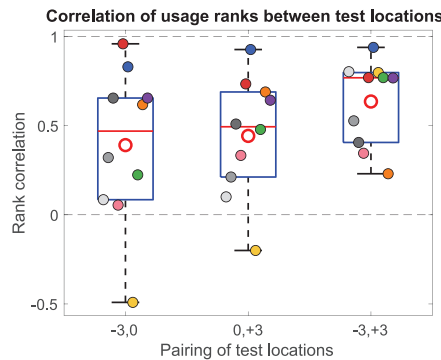


Fig. A2. For each observer, we computed the similarity in letter usage (rank correlation) across a given pair of test locations. These individual-subject values are plotted as filled symbols, with a different colour for each observer. To avoid overlap, observers S1 to S10 are displaced from left to right across each boxplot. Upper and lower edges of the boxplot mark the inter-quartile range; red line is the median value. Red open circle is the group mean value. Dashed vertical 'whiskers' mark the highest and lowest values for each location pairing.

Table A2

Summary of rank correlations in letter usage computed between pairs of test locations for individual subjects (N = 10 Ss).

Location pairing	[-3, 0]	[0, +3]	[-3, +3]
Mean correl \pm S.D.	0.391 \pm 0.438	0.443 \pm 0.337	0.636 \pm 0.239
Mean is > 0? Yes	t = 2.83, df = 9, P = 0.01	t = 4.16, df = 9, P = 0.001	t = 8.42, df = 9, P < 0.00001
Median	0.470	0.494	0.769
Median > 0? Yes	Wilcoxon, P = 0.0098	Wilcoxon, P = 0.0029	Wilcoxon, P = 0.00098
Min, Max	-0.491, 0.959	-0.200, 0.927	0.230, 0.939

= 0.142; P = 0.169 with Greenhouse-Geisser correction]. Two observers (S5, S6; red, blue in Fig. A.2) showed strikingly high correlations (mean 0.86, range 0.73 to 0.96). A third observer (S10, purple in Fig. A.2) was similarly consistent across the three location pairings, but with somewhat lower correlations (mean 0.69). We conclude from these analyses that patterns of letter bias (as reflected in patterns of letter usage) can be highly reliable across test locations for some observers. But for most observers (Fig. A.2) such consistency was weaker, and sometimes absent.

Appendix B. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.visres.2023.108233>.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Xplore: IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Alexander, K. R., Xie, W., & Derlacki, D. J. (1997). Visual acuity and contrast sensitivity for individual Sloan letters. *Vision Research*, 37(6), 813–819.
- Barboom, H., Joshi, M. R., & Schmidtmann, G. (2021). The effect of response biases on resolution thresholds of Sloan letters in central and paracentral vision. *Vision Research*, 187, 110–119.
- Brainard, D. H. (1997). Psychophysics software for use with MATLAB. *Spatial Vision*, 10(4), 433–436.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). New York: Springer.
- Candy, T. R., Mishoulam, S. R., Nosofsky, R. M., & Dobson, V. (2011). Adult discrimination performance for pediatric acuity test optotypes. *Investigative ophthalmology & visual science*, 52(7), 4307–4313.

- Coates, D. R. (2015). *Quantifying crowded and uncrowded letter recognition*. Berkeley: University of California. Ph.D. thesis.
- Collett, D. (2003). *Modelling Binary Data* ((2nd ed.)). Boca Raton & London: Chapman & Hall/CRC.
- DeCarlo, L. T. (2012). On a signal detection approach to *m*-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology*, 56(3), 196–207. <https://doi.org/10.1016/j.jmp.2012.02.004>
- Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, 9(7), 335–341. <https://doi.org/10.1016/j.tics.2005.05.004>
- Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in Cognitive Sciences*, 12(10), 381–387. <https://doi.org/10.1016/j.tics.2008.06.006>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Hamm, L. M., Yeoman, J. P., Anstice, N., & Dakin, S. C. (2018). The Auckland Optotypes: An open-access pictogram set for measuring recognition acuity. *Journal of Vision*, 18(3):13, 1–15.
- Heeger, D. J. (1992). Half-squaring in responses of cat striate cells. *Visual Neuroscience*, 9, 427–443.
- Jogan, M., & Stocker, A. A. (2014). A new two-alternative forced choice method for the unbiased characterization of perceptual bias and discriminability. *Journal of Vision*, 14(3):20, 1–18.
- Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A practical introduction* (1st ed.). London: Academic Press.
- Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3? *Perception*, 36(14), 1–16.
- Linares, D., Aguilar-Lleyda, D., & López-Moliner, J. (2019). Decoupling sensory from decisional choice biases in perceptual decision making. *Elife*, 8, e43994.
- Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology: 1* (pp. 103–189). New York: Wiley.
- Ma, W. J., Shen, S., Dzingaite, G., & van den Berg, R. (2015). Requiem for the max rule? *Vision Research*, 116, 179–193. <https://doi.org/10.1016/j.visres.2014.12.019>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A user's guide* (2nd ed.). Mahwah, NJ & London: Lawrence Erlbaum Associates.

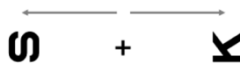
- Morgan, M. J. (2014). A bias-free measure of retinotopic tilt adaptation. *Journal of Vision*, *14*(1):7, 1–9. <http://www.journalofvision.org/content/14/1/7>. <https://doi.org/10.1167/14.1.7>
- Nadarajah, S., & Kotz, S. (2008). Exact Distribution of the Max/Min of Two Gaussian Random Variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *16*(2), 210–212. <https://doi.org/10.1109/tvlsi.2007.912191>
- Rahnev, D. (2021). Response bias reflects individual differences in sensory encoding. *Psychological Science*, *32*(7), 1157–1168.
- Sridharan, D., Steinmetz, N. A., Moore, T., & Knudsen, E. I. (2014). Distinguishing bias from sensitivity effects in multialternative detection tasks. *Journal of Vision*, *14*(9): 16, 1–32.
- Storrs, K. R. (2015). Are high-level aftereffects perceptual? *Frontiers in Psychology*, *6* (Feb), 6–9. <https://doi.org/10.3389/fpsyg.2015.00157>
- Swets, J. A. (1961). Is there a sensory threshold? *Science*, *134*, 168–177.
- Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340.
- Tanner, W. P. (1956). Theory of recognition. *Journal of the Acoustical Society of America*, *28*, 882–888.
- Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 401–409.
- Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.
- Watson, A. B., & Ahumada, A. J. (2015). Letter identification and the neural image classifier. *Journal of Vision*, *15*(2):15, 1–26. <https://doi.org/10.1167/15.2.15>
- Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford & New York: Oxford University Press.
- Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision research*, *48*(17), 1837–1851.
- Zhou, J., Georgeson, M. A., & Hess, R. F. (2014). Linear binocular combination of responses to contrast modulation: Contrast-weighted summation in first- and second-order vision. *Journal of Vision*, *14*(13):24, 1–19. <https://doi.org/10.1167/14.13.24>. doi

BACKGROUND

- Previous studies suggest biases towards certain letters in acuity measurements [1-4]
- Here we measured letter identification performance against letter size, spanning the resolution threshold, for 10 Sloan letters at central and para-central (43° eccentricity) visual field locations
- The aim of the current study was to develop a model to quantify such biases

METHODS

- N=10 (naïve subjects)
- Visual acuity measurements for Sloan letters
- Central and 3° eccentricity in upper and lower visual field (vertical meridian).



- Data of the psychometric functions were split into three groups:
 - The top 2 (most used) letters
 - The bottom 2 (least used)
 - The 6 letters with intermediate usage

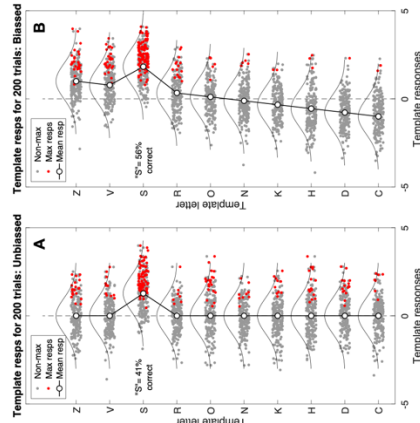
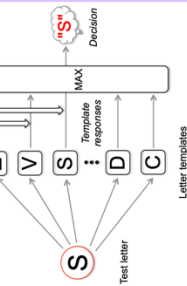
The role of bias in a letter acuity identification task: a noisy template model

Hatem Barhoom¹, Gunnar Schmidtmann¹, Mahesh R. Joshi¹, Paul H. Artes¹, Mark A. Georgeson²
¹University of Plymouth, Plymouth, UK, ²Aston University, Birmingham, UK



MODEL

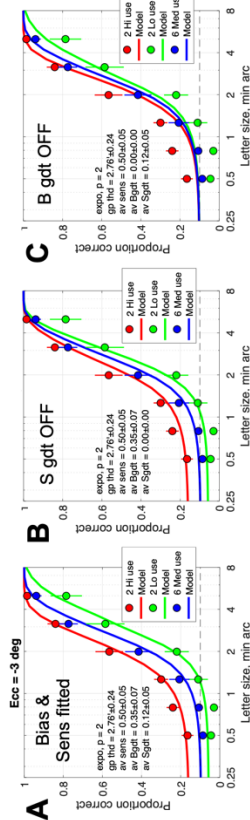
- A "noisy template" model was derived to distinguish biases from differences in visibility



- The Noisy Template Model:
 - A without biases
 - B with biases
- Biases (or sensitivity) are ordered from negative to positive with a mean of zero.
- Free model parameters are: Baseline sensitivity, Bias gradient (Bgd), Sensitivity gradient (Sgdt)

RESULTS

- Results show that observers had different individual letter biases consistent across the whole range of sizes.
- Preferred letters were called more frequently and others less frequently than expected (group averages from 4% to 20% across letters, where the unbiased rate was 10%)
- The over- and under-calling decreased at larger letter sizes, but was well-predicted by templates having fixed additive response bias: with stronger inputs (larger letters) there is less opportunity for bias to influence which template gives the biggest response



- A fit of the two-factor (bias and sensitivity) model
- B the model was re-computed with Sgdt = 0.
- C the model was re-computed with Bgd = 0.
- Comparison of row B with C strongly suggests that:
 - Bias (in row B) was the major contributing factor to the variation in letter usage
 - Variation in sensitivity across letters (row C) played a much smaller role, mainly at the larger letter sizes at the periphery.

CONCLUSION

- Results show that biases are responsible for the observed differences in letter usage in the letter acuity measurements with Sloan letters
- In the future, it will be important to investigate whether the observed response biases are likely to have a meaningful effect on clinical measures of visual performance

REFERENCES

1. Barhoom, H., Joshi, M. R., & Schmidtmann, G. (2021). The effect of response biases on resolution thresholds of Sloan letters in central and para-central vision. *Vision Research*, 187, 110-119.
2. 110-119. In: Moshelashvili, S. R., Neesley, R. M., & Dobson, V. (2011). Adult discrimination performance for pediatric acuity test optotypes. *Investigative ophthalmology & visual science*, 52(7), 4307-4313.
3. Conies, D. R. (2015). *Quantifying crowded and uncrowded letter recognition*. University of California, Berkeley.
4. Hamm, L. M., Neuman, J. P., Ardson, N., & Cohen, S. C. (2016). The Audubon Optotypes: an open-access optogram set for measuring recognition acuity. *Journal of vision*, 16(3), 13-13.



Hatem Barhoom^{1,3}, Gunnar Schmidtman¹, Mahesh R. Joshi¹, Paul H. Artes¹, Mark A. Georgeson²
¹University of Plymouth, Plymouth, UK, ²Aston University, Birmingham, UK, ³Islamic University of Gaza, Palestine

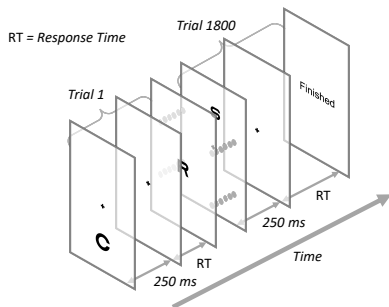


BACKGROUND

- Letters have complex structure, and it is likely that letter bias and similarity between letters are sources of non-random errors in letter identification tasks¹⁻³
- Here we introduce a novel and significant extension to our recently derived model⁴, to reveal the joint effect of bias and similarity in letter identification.

METHODS

- Letter identification task, near the acuity limit; 10 naive Ss
- 10 Sloan letters, presented singly for 250 ms
- Central and 3° eccentricity in upper and lower visual field
- Method of constant stimuli (1800 trials/subject)

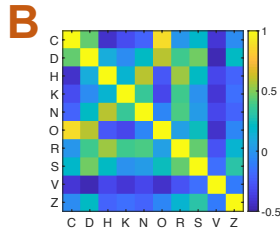
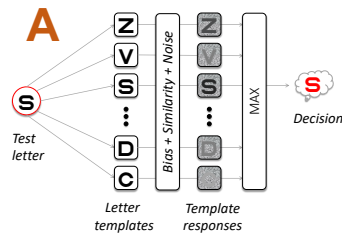


- Data collated as a *stimulus-response matrix*: each cell counts the number of times a given letter was chosen in response to the letter presented.

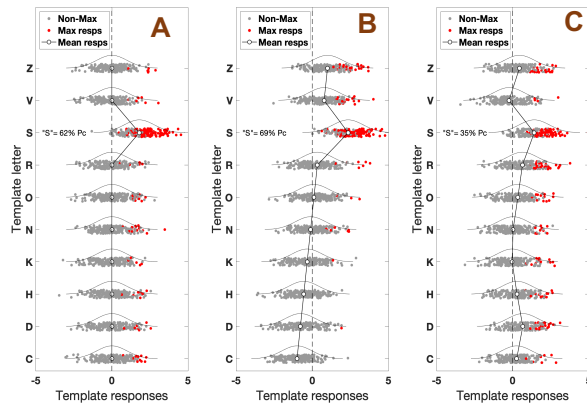
MODEL

(A) The “noisy template” model⁴ was extended to capture letter biases & effects of between-letter similarity in experimental data.

(B) The optotype correlation matrix⁵ was used to model similarity.

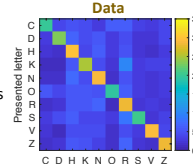


- The Noisy Template Model:
 - A** without biases or similarity.
 - B** with biases but no similarity.
 - C** with similarity but no biases.
- Model parameters are Baseline sensitivity, Bias gradient, Confusion strength.
- The model was fitted to the observed number of responses.



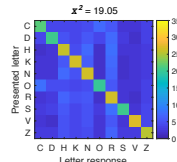
RESULTS

- Data**: stimulus-response matrix (group average).
- Model 1**: prediction with bias and similarity.
- Model 2**: prediction with bias only.
- Model 3**: prediction with similarity only.
- Model comparison using AIC showed that model 1 was favoured over the other two models.
- Chi square (χ^2) scores similarly showed that bias & similarity (model 1) predicted the data more closely (smaller χ^2 value) than bias or similarity alone did.



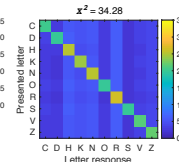
Model 1

$\chi^2 = 19.05$



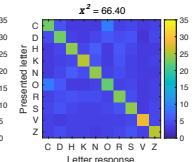
Model 2

$\chi^2 = 34.28$



Model 3

$\chi^2 = 66.40$



CONCLUSIONS

- Letter biases and between-letter similarity together shape the pattern of correct responses (negative diagonal) and errors (off-diagonal) in the letter identification task with Sloan letters near the acuity limit.
- In future work, it will be important to investigate the impact of bias and similarity on the estimated acuity using letters as optotypes.

REFERENCES

- Barhoom, H., Joshi, M. R., & Schmidtman, G. (2021). The effect of response biases on resolution thresholds of Sloan letters in central and para-central vision. *Vision Research*, 187, 110-119.
- McMinnies, C. W. & A. Ho (1996). Analysis of errors in letter acuity measurements. *Clinical and Experimental Optometry* 79(4): 144-151.
- Reich, L. N. and H. E. Bedell (2000). Relative legibility and confusions of letter acuity targets in the peripheral and central retina. *Optometry and Vision Science* 77(5): 270-
- Georgeson MA, Barhoom H, Joshi MR, Artes PH, Schmidtman G (2022). Revealing the influence of bias in a letter acuity identification task: a noisy template model. *Vision Research*, submitted.
- Fülep, C., Kovács, I., Kránitz, K., & Erdei, G. (2017). Correlation-based evaluation of visual performance to reduce the statistical error of visual acuity. *JOSA A*, 34(7), 1255-1264.

The role of bias, sensitivity and similarity in letter identification task: a noisy template model

Hatem Barhoom^{1,3}, Mark A. Georgeson², Mahesh R. Joshi¹, Paul H. Artes¹, Gunnar Schmidtmann¹
¹University of Plymouth, Plymouth, UK, ²Aston University, Birmingham, UK, ³Islamic University of Gaza, Palestine



UNIVERSITY OF PLYMOUTH



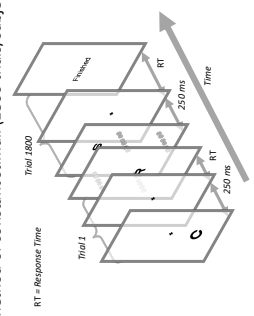
Aston University BIRMINGHAM UK

BACKGROUND

- Letters have complex structure, and it is likely that letter bias, sensitivity differences and similarity between letters are sources of errors in letter identification tasks¹⁻³
- Here we extend our recently derived model⁴ to assess the effect of similarity between letters in addition to bias and sensitivity in letter identification tasks.

METHODS

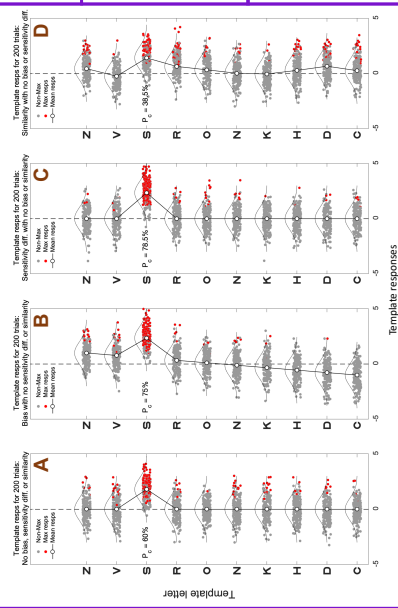
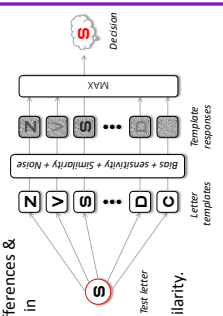
- 10 alternative single interval Letter identification task
- N = 10 (naive)
- 10 Sloan letters
- presentation time = 250 ms
- Central and 3° eccentricity in upper and lower visual field
- Method of constant stimuli (1800 trials/subject)



- Data collated as a *stimulus-response matrix*: each cell counts the number of times a given letter was chosen in response to the letter presented.

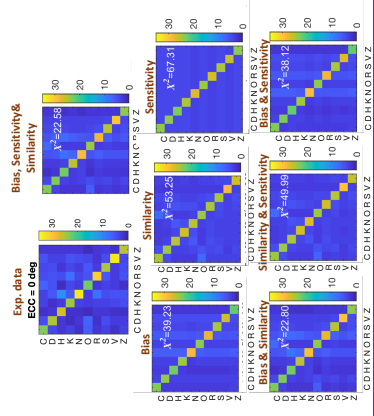
MODEL

- The “noisy template” model⁴ was extended to capture letter biases, sensitivity differences & effects of between-letter similarity in experimental data.
- The Noisy Template Model:
 - A without biases, sensitivity or similarity.
 - B with biases only.
 - C with similarity only.
 - D with similarity only.
- Model parameters are Baseline sensitivity, Bias gradient, Sensitivity gradient, and Confusion strength.
- The model was fitted to the observed number of responses.



RESULTS

- Exp. Data (Fovea): stimulus-response matrix (group average).
- Chi square (χ^2) scores and Model comparison using AIC showed that Bias, sensitivity & similarity model was favoured over the other models' variants.



CONCLUSION

- Letter biases, sensitivity and between-letter similarity together shape the pattern of responses in the letter identification task.
- No significant effect of bias, sensitivity, or similarity on the estimated letter acuity.
- Models that included similarity increases the spread of the psychometric function towards upper asymptote (mainly in the peripheral test locations).

REFERENCES

- Barhoom, H., Joshi, M. R., & Schmidtmann, G. (2021). The effect of response biases on resolution thresholds of Sloan letters in central and peripheral vision. *Vision Research*, 187, 110-119.
- McKinnon, C. W. & A. H. (1996). Analysis of errors in letter acuity measurements. *Clinical and Experimental Optometry* 73(6): 144-151.
- Redd, L. W. & H. E. Barbur (2002). Relative visibility of confusable letter acuity targets in the peripheral visual field. *Optometry* 73(5): 27-31.
- Georgeson, M.A., Barhoom, H., Joshi, M.R., Artes, P.H., & Schmidtmann, G. (2023). Revealing the influence of bias in a letter acuity identification task: a noisy template model. *Vision Research*.

References

- Acton, J. H., & Greenstein, V. C. (2013). Fundus-driven perimetry (microperimetry) compared to conventional static automated perimetry: similarities, differences, and clinical applications. *Canadian Journal of Ophthalmology*, 48(5), 358-363.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Xplore: IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Alexander, K. R., W. Xie and D. J. Derlacki (1997). Visual acuity and contrast sensitivity for individual Sloan letters. *Vision research* 37(6): 813-819.
- Allison, K., Patel, D., & Alabi, O. (2020). Epidemiology of glaucoma: the past, present, and predictions for the future. *Cureus*, 12(11).
- Alluwimi, M. S., Swanson, W. H., & King, B. J. (2018). Identifying glaucomatous damage to the macula. *Optometry and Vision Science*, 95(2), 96.
- Almasieh, M., Wilson, A. M., Morquette, B., Vargas, J. L. C., & Di Polo, A. (2012). The molecular basis of retinal ganglion cell death in glaucoma. *Progress in retinal and eye research*, 31(2), 152-181.
- Anctil, J. L., & Anderson, D. R. (1984). Early foveal involvement and generalized depression of the visual field in glaucoma. *Archives of ophthalmology*, 102(3), 363-370.
- Anderson, A. J. (2003). Utility of a dynamic termination criterion in the ZEST adaptive threshold method. *Vision research*, 43(2), 165-170.
- Anderson, A. J., Bedggood, P. A., Kong, Y. X. G., Martin, K. R., & Vingrys, A. J. (2017). Can home monitoring allow earlier detection of rapid visual field progression in glaucoma? *Ophthalmology*, 124(12), 1735-1742.
- Anderson, R. S. (1996). The selective effect of optical defocus on detection and resolution acuity in peripheral vision. *Current eye research*, 15(3), 351-353.
- Anderson, R. S. (2006). The psychophysics of glaucoma: improving the structure/function relationship. *Progress in retinal and eye research*, 25(1), 79-97.
- Anderson, R. S., & Ennis, F. A. (1999). Foveal and peripheral thresholds for detection and resolution of vanishing optotype tumbling E's. *Vision research*, 39(25), 4141-4144.
- Anderson, R., & Thibos, L. (2004). The filtered Fourier difference spectrum predicts psychophysical letter discrimination in the peripheral retina. *Spatial vision*, 17(1-2), 5-15.
- Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision research*, 14(7), 589-592.
- Arditi, A. (2006). Lapse resistance in the verbal letter reporting task. *Vision research*, 46(8-9), 1327-1330.

- Attneave, F., & Arnoult, M. D. (1956). The quantitative study of shape and pattern perception. *Psychological bulletin*, 53(6), 452.
- Aulhorn, E., & Harms, H. (1967). Early visual field defects in glaucoma. In *Glaucoma* (151-186).
- Bailey, I. L., & Lovie, J. E. (1976). New design principles for visual acuity letter charts. *American journal of optometry and physiological optics*, 53(11), 740-745.
- Barhoom, H., Joshi, M. R., & Schmidtman, G. (2021). The effect of response biases on resolution thresholds of Sloan letters in central and paracentral vision. *Vision Research*, 187, 110-119.
- Barhoom, H., Schmidtman, G., Joshi, M. R., Artes, P. H., & Georgeson, M. A. (2022). The role of similarity and bias in letter acuity measurements: a noisy template model. In *Perception*, 51, 192-193.
- Bartlett, H., Davies, L. N., & Eperjesi, F. (2005). The macular mapping test: a reliability study. *BMC ophthalmology*, 5(1), 1-7.
- Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical methods in medical research*, 8(2), 135-160.
- Bohigian, G. M. (2008). An ancient eye test—using the stars. *Survey of Ophthalmology*, 53(5), 536-539.
- Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, 10(4), 433-436.
- Candy, T. R., Mishoulam, S. R., Nosofsky, R. M., & Dobson, V. (2011). Adult discrimination performance for pediatric acuity test optotypes. *Investigative ophthalmology & visual science*, 52(7), 4307-4313.
- Carkeet, A. (2001). Modelling logMAR visual acuity scores: effects of termination rules and alternative forced-choice options. *Optometry and Vision Science*, 78(7), 529-538.
- Casson, R. J., Chidlow, G., Wood, J. P., Crowston, J. G., & Goldberg, I. (2012). Definition of glaucoma: clinical and experimental concepts. *Clinical & experimental ophthalmology*, 40(4), 341-349.
- Chen, S., McKendrick, A. M., & Turpin, A. (2015). Choosing two points to add to the 24-2 pattern to better describe macular visual field damage due to glaucoma. *British Journal of Ophthalmology*, 99(9), 1236-1239.
- Chhetri, A. P., Wen, F., Wang, Y., & Zhang, K. (2010). Shape discrimination test on handheld devices for patient self-test. In *Proceedings of the 1st ACM International Health Informatics Symposium* (502-506).
- Chong, R. S., & Martin, K. R. (2014). Retinal ganglion cell dendrites and glaucoma: a case of missing the wood for the trees? *Expert Review of Ophthalmology*, 9(3), 149-152.

- Coates, D. R. (2015). Quantifying crowded and uncrowded letter recognition (Doctoral dissertation, University of California, Berkeley).
- Colenbrander, A. (2008). The historical evolution of visual acuity measurement. *Visual impairment research*, 10(2-3), 57-66.
- Collett, D. (2002). *Modelling binary data*. (2nd ed.)
- Crossland, M., & Rubin, G. (2007). The Amsler chart: absence of evidence is not evidence of absence. *British Journal of Ophthalmology*, 91(3), 391-393.
- Curcio, C. A., & Allen, K. A. (1990). Topography of ganglion cells in human retina. *Journal of comparative Neurology*, 300(1), 5-25.
- Curcio, C. A., Sloan Jr, K. R., Packer, O., Hendrickson, A. E., & Kalina, R. E. (1987). Distribution of cones in human and monkey retina: individual variability and radial asymmetry. *Science*, 236(4801), 579-582.
- De Jong, P. T. (2022). A history of visual acuity testing and optotypes. *Eye*, 1-12.
- De Moraes, C. G., Hood, D. C., Thenappan, A., Girkin, C. A., Medeiros, F. A., Weinreb, R. N., ... & Liebmann, J. M. (2017). 24-2 visual fields miss central defects shown on 10-2 tests in glaucoma suspects, ocular hypertensives, and early glaucoma. *Ophthalmology*, 124(10), 1449-1456.
- De Moraes, C. G., Muhammad, H., Kaur, K., Wang, D., Ritch, R., & Hood, D. C. (2018). Interindividual variations in foveal anatomy and artifacts seen on inner retinal probability maps from spectral domain OCT scans of the macula. *Translational Vision Science & Technology*, 7(2), 4-4.
- De Moraes, C. G., Song, C., Liebmann, J. M., Simonson, J. L., Furlanetto, R. L., & Ritch, R. (2014). Defining 10-2 visual field progression criteria: exploratory and confirmatory factor analysis using pointwise linear regression. *Ophthalmology*, 121(3), 741-749.
- De Moraes, C. G., Sun, A., Jarukasetphon, R., Rajshekhar, R., Shi, L., Blumberg, D. M., & Hood, D. C. (2019). Association of macular visual field measurements with glaucoma staging systems. *JAMA ophthalmology*, 137(2), 139-145.
- Demirel, S., & Vingrys, A. J. (1994). Eye movements during perimetry and the effect that fixational instability has on perimetric outcomes. *Journal of glaucoma*, 3(1), 28-35.
- Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8599-8603). IEEE.
- Ennis, F. A., Anderson, A. J., & Johnson, C. A. (2002). A new staircase method, 2A-non-FC, in resolution acuity threshold recording. *Investigative Ophthalmology & Visual Science*, 43(13), 2137-2137.

- Erdei, G., & Fulep, C. (2019). U.S. Patent Application No. 16/394,388
- Evangelho, K., Mogilevskaya, M., Losada-Barragan, M., & Vargas-Sanchez, J. K. (2019). Pathophysiology of primary open-angle glaucoma from a neuroinflammatory and neurotoxicity perspective: a review of the literature. *International ophthalmology*, (39), 259-271.
- Ferris 3rd, F. L., & Bailey, I. (1996). Standardizing the measurement of visual acuity for clinical research studies: Guidelines from the Eye Care Technology Forum. *Ophthalmology*, 103(1), 181-182.
- Ferris III, F. L., Kassoff, A., Bresnick, G. H., & Bailey, I. (1982). New visual acuity charts for clinical research. *American journal of ophthalmology*, 94(1), 91-96.
- Fink, W., & Sadun, A. A. (2004). Three-dimensional computer-automated threshold Amsler grid test. *Journal of biomedical optics*, 9(1), 149-153.
- Fiset, D., Blais, C., Ethier-Majcher, C., Arguin, M., Bub, D., & Gosselin, F. (2008). Features for identification of uppercase and lowercase letters. *Psychological science*, 19(11), 1161-1168.
- Foster, P. J., Buhmann, R., Quigley, H. A., & Johnson, G. J. (2002). The definition and classification of glaucoma in prevalence surveys. *British journal of ophthalmology*, 86(2), 238-242.
- Frisén, L. (1986). Vanishing optotypes: new type of acuity test letters. *Archives of ophthalmology*, 104(8), 1194-1198.
- Frisén, L. (1987). A computer-graphics visual field screener using high-pass spatial frequency resolution targets and multiple feedback devices. In *Seventh International Visual Field Symposium, Amsterdam*, (441-446).
- Frisén, L. (1993). High-pass resolution perimetry: a clinical review. *Documenta ophthalmologica*, (83), 1-25.
- Fülep, C., Kovács, I., Kránitz, K., & Erdei, G. (2017). Correlation-based evaluation of visual performance to reduce the statistical error of visual acuity. *JOSA A*, 34(7), 1255-1264.
- García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision research*, 38(12), 1861-1881.
- Garg, A., Hood, D. C., Pensec, N., Liebmann, J. M., & Blumberg, D. M. (2018). Macular damage, as determined by structure-function staging, is associated with worse vision-related quality of life in early glaucoma. *American journal of ophthalmology*, 194, 88-94.
- Georgeson, M. A., Barhoom, H., Joshi, M. R., Artes, P. H., & Schmidtman, G. (2023). Revealing the influence of bias in a letter acuity identification task: A noisy template model. *Vision Research*, 208, 108-233

- Geyer, L. H. (1971). A two-channel theory of short-term visual storage (Doctoral dissertation, ProQuest Information & Learning).
- Geyer, L. H., & DeWald, C. G. (1973). Feature lists and confusion matrices. *Perception & Psychophysics*, 14(3), 471-482.
- Gibson, E. J. (1969). *Principles of perceptual learning and development*.
- Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, 25(5), 425-431.
- Glen, F. C., Baker, H., & Crabb, D. P. (2014). A qualitative investigation into patients' views on visual field testing for glaucoma monitoring. *BMJ open*, 4(1), e003996.
- Gordon-Bennett, P. S. C., Ioannidis, A. S., Papageorgiou, K., & Andreou, P. S. (2008). A survey of investigations used for the management of glaucoma in hospital service in the United Kingdom. *Eye*, 22(11), 1410-1418.
- Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research*, 41(17), 2261-2271.
- Graham, N. V. S. (1989). *Visual pattern analyzers*. Oxford University Press.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1, pp. 1969-2012). New York: Wiley.
- Grillo, L. M., Wang, D. L., Ramachandran, R., Ehrlich, A. C., De Moraes, C. G., Ritch, R., & Hood, D. C. (2016). The 24-2 visual field test misses central macular damage confirmed by the 10-2 visual field test and optical coherence tomography. *Translational vision science & technology*, 5(2), 15-15.
- Grimm, W., Rassow, B., Wesemann, W., Saur, K., & Hilz, R. (1994). Correlation of optotypes with the Landolt ring - a fresh look at the comparability of optotypes. *Optometry and vision science: official publication of the American Academy of Optometry*, 71(1), 6-13.
- Hahn, G. A., Messias, A., MacKeben, M., Dietz, K., Horwath, K., Hyvärinen, L., & Trauzettel-Klosinski, S. (2009). Parafoveal letter recognition at reduced contrast in normal aging and in patients with risk factors for AMD. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 247, 43-51.
- Hairol, M. I., N. A. Abd-Latif, P. Low, W. P. Lim, J. Y. Aik and S. Kaur (2015). "Effects of foveal and eccentric viewing on the resolution and contrast thresholds of individual letters." *Psychology & Neuroscience* 8(2): 183-192.
- Hamm, L. M., J. P. Yeoman, N. Anstice and S. C. Dakin (2018). "The Auckland Optotypes: An open-access pictogram set for measuring recognition acuity." *Journal of Vision*, 18(3): 13.

- Hamzah, F. B., Lau, C., Nazri, H., Ligot, D. V., Lee, G., Tan, C. L., ... & Chung, M. H. (2020). CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. *Bull World Health Organ*, 1(32), 1-32.
- Hangai, M., Ikeda, H. O., Akagi, T., & Yoshimura, N. (2014). Paracentral scotoma in glaucoma detected by 10-2 but not by 24-2 perimetry. *Japanese journal of ophthalmology*, 58, 188-196.
- Heeger, D. J. (1992). Half-squaring in responses of cat striate cells. *Visual Neuroscience*, 9, 427-443.
- Himmelberg, M. M., Winawer, J., & Carrasco, M. (2023). Polar angle asymmetries in visual perception and neural architecture. *Trends in Neurosciences*.
- Hirasawa, K., Okano, K., Koshiji, R., Funaki, W., & Shoji, N. (2016). Smaller fixation target size is associated with more stable fixation and less variance in threshold sensitivity. *PLoS One*, 11(11), e0165046.
- Hirooka, K., Misaki, K., Nitta, E., Ukegawa, K., Sato, S., & Tsujikawa, A. (2016). Comparison of Macular Integrity Assessment (MAIA™), MP-3, and the Humphrey Field Analyzer in the Evaluation of the Relationship between the Structure and Function of the Macula. *PLoS One*, 11(3), e0151000.
- Holbrook, M. B. (1975). A comparison of methods for measuring the interletter similarity between capital letters. *Perception & Psychophysics*, 17, 532-536.
- Hood, D. C., Nguyen, M., Ehrlich, A. C., Raza, A. S., Sliesoraityte, I., De Moraes, C. G., & Schiefer, U. (2014). A test of a model of glaucomatous damage of the macula with high-density perimetry: implications for the locations of visual field test points. *Translational vision science & technology*, 3(3), 5-5.
- Hood, D. C., Raza, A. S., De Moraes, C. G. V., Liebmann, J. M., & Ritch, R. (2013). Glaucomatous damage of the macula. *Progress in retinal and eye research*, 32, 1-21.
- Hood, D. C., Raza, A. S., De Moraes, C. G. V., Odel, J. G., Greenstein, V. C., Liebmann, J. M., & Ritch, R. (2011). Initial arcuate defects within the central 10 degrees in glaucoma. *Investigative ophthalmology & visual science*, 52(2), 940-946.
- Hood, D. C., Tsamis, E., Bommakanti, N. K., Joiner, D. B., Al-Aswad, L. A., Blumberg, D. M., & De Moraes, C. G. (2019). Structure-function agreement is better than commonly thought in eyes with early glaucoma. *Investigative ophthalmology & visual science*, 60(13), 4241-4248.
- Hou, H., Moghimi, S., Zangwill, L. M., Shoji, T., Ghahari, E., Penteado, R. C., & Weinreb, R. N. (2019). Macula vessel density and thickness in early primary open-angle glaucoma. *American journal of ophthalmology*, 199, 120-132.

- Ishiyama, Y., Murata, H., Mayama, C., & Asaoka, R. (2014). An objective evaluation of gaze tracking in Humphrey perimetry and the relation with the reproducibility of visual fields: a pilot study in glaucoma. *Investigative ophthalmology & visual science*, 55(12), 8149-8152.
- Jones, P. R., Campbell, P., Callaghan, T., Jones, L., Asfaw, D. S., Edgar, D. F., & Crabb, D. P. (2021). Glaucoma home monitoring using a tablet-based visual field test (Eyecatcher): an assessment of accuracy and adherence over 6 months. *American Journal of Ophthalmology*, 223, 42-52.
- Jung, Y., Park, H. Y. L., Jeong, H. J., Choi, S. Y., & Park, C. K. (2015). The ability of 10-2 short-wavelength perimetry in detecting functional loss of the macular area in preperimetric glaucoma patients. *Investigative ophthalmology & visual science*, 56(13), 7708-7714.
- Jung, Y., Park, H. Y. L., Park, Y. R., & Park, C. K. (2017). Usefulness of 10-2 matrix frequency doubling technology perimetry for detecting central visual field defects in preperimetric glaucoma patients. *Scientific reports*, 7(1), 14622.
- Kalesnykas, G., Oglesby, E. N., Zack, D. J., Cone, F. E., Steinhart, M. R., Tian, J., & Quigley, H. A. (2012). Retinal ganglion cell morphology after optic nerve crush and experimental glaucoma. *Investigative ophthalmology & visual science*, 53(7), 3847-3857.
- Kanski, J. J., & Bowling, B. (2011). *Clinical ophthalmology: a systematic approach*. Elsevier Health Sciences.
- Kasten, E., Bunzenthall, U., & Sabel, B. A. (2006). Visual field recovery after vision restoration therapy (VRT) is independent of eye movements: an eye tracker study. *Behavioural brain research*, 175(1), 18-26.
- Khan, S., & Ullah, K. (2017). Visual acuity test for isolated words using speech recognition. *International Conference on Innovations in Electrical Engineering and Computational Technologies (1-6)*. IEEE.
- Khanna, V., Joon, A., Viswanath, S., & Chhabra, K. (2022). Perimetry-Recent Advances. *The Official Scientific Journal of Delhi Ophthalmological Society*, 32(4), 15-24.
- Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A practical introduction (1st ed.)*. London: Academic Press.
- King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation, and practical implementation. *Vision research*, 34(7), 885-912.
- Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?
- Kniestedt, C. and R. Stamper (2003). Visual acuity and its measurement. *Ophthalmology Clinics of North America* 16(2), 155-170.

- Kovalevskaya, M., Milyutkina, S., Fink, W., Belyi, Y., & Tereshchenko, A. (2016). 3D-CTAG testing of functional and structural changes of the macula. *AOVS*, 4(2).
- Lakhani, B. K., Attzs, M. S., Stead, R., & Tambe, K. (2021). The impact of the COVID-19 pandemic on ophthalmology services across the United Kingdom: a brief report on a cross-sectional survey of clinical leads. *Therapeutic Advances in Ophthalmology*, 13.
- Landolt, E. (1899). Nouveaux opto-types pour la détermination de l'acuité visuelle. *Archives d'Ophtalmologie*, 465-471.
- Laughery, K. R. (1970). Computer simulation of short-term memory: A component-decay model. In *Psychology of Learning and Motivation*, 3, 135-200.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.
- Leeprechanon, N., Giaconi, J. A., Manassakorn, A., Hoffman, D., & Caprioli, J. (2007). Frequency doubling perimetry and short-wavelength automated perimetry to detect early glaucoma. *Ophthalmology*, 114(5), 931-937.
- Leung, C. K. S., Weinreb, R. N., Li, Z. W., Liu, S., Lindsey, J. D., Choi, N., & Lam, D. S. C. (2011). Long-term in vivo imaging and measurement of dendritic shrinkage of retinal ganglion cells. *Investigative ophthalmology & visual science*, 52(3), 1539-1547.
- Liu, L., Wang, Y. Z., & Bedell, H. E. (2014). Visual-function tests for self-monitoring of age-related macular degeneration. *Optometry and Vision Science*, 91(8), 956-965.
- Loomis, J. M. (1982). Analysis of tactile and visual confusion matrices. *Perception & Psychophysics*, 31, 41-52.
- Luce, R., Bush, R. R., & Galanter, E. E. (1963). *Handbook of mathematical psychology: I.*, Wiley, (1), (103-189).
- Ludvig, E. (1941). Extrafoveal visual acuity as measured with Snellen test-letters. *Am J Ophthalmol*, 24, 303-310.
- MacKeben, M. (2008). Topographic mapping of residual vision by computer. *Journal of Visual Impairment & Blindness*, 102(10), 649-655.
- Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological bulletin*, 107(3), 401.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: a user's guide* (2nd ed.). Mahwah, NJ & London: Lawrence Erlbaum Associates.
- Majaj, N. J., Pelli, D. G., Kurshan, P., & Palomares, M. (2002). The role of spatial frequency channels in letter identification. *Vision research*, 42(9), 1165-1184.
- Markowitz, S. N., & Reyes, S. V. (2013). Microperimetry and clinical practice: an evidence-based review. *Canadian journal of ophthalmology*, 48(5), 350-357.

- Matsuura, M., Murata, H., Fujino, Y., Hirasawa, K., Yanagisawa, M., & Asaoka, R. (2018). Evaluating the usefulness of MP-3 microperimetry in glaucoma patients. *American journal of ophthalmology*, 187, 1-9.
- McKendrick, A. M., Johnson, C. A., Anderson, A. J., & Fortune, B. (2002). Elevated vernier acuity thresholds in glaucoma. *Investigative ophthalmology & visual science*, 43(5), 1393-1399.
- McMonnies, C. W., & Ho, A. (2000). Letter legibility and chart equivalence. *Ophthalmic and Physiological Optics*, 20(2), 142-152.
- McMonnies, C.W. and A. Ho (1996). "Analysis of errors in letter acuity measurements." *Clinical and Experimental Optometry* 79(4): 144-151.
- Michelessi, M., Lucenteforte, E., Oddone, F., Brazzelli, M., Parravano, M., Franchi, S., & Virgili, G. (2015). Optic nerve head and fibre layer imaging for diagnosing glaucoma. *Cochrane Database of Systematic Reviews*, 11.
- Millodot, M., Johnson, C. A., Lamont, A., & Leibowitz, H. W. (1975). Effect of dioptics on peripheral visual acuity. *Vision Research*, 15(12), 1357-1362.
- Molina-Martín, A., Pérez-Cambrodí, R. J., & Piñero, D. P. (2018). Current clinical application of microperimetry: a review. In *Seminars in Ophthalmology*, 33(5), 620-628.
- Mueller, S. T. and C. T. Weidemann (2012). "Alphabetic letter identification: effects of perceivability, similarity, and bias." *Acta Psychol (Amst)* 139(1): 19-37.
- Nassar, M. K., Badawi, N. M., & Diab, M. M. M. (2015). Resurrection of the Amsler chart in macular diseases. *Menoufia Medical Journal*, 28(1), 174.
- Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7, 19143-19165.
- Navarro, R. (2009). The optical design of the human eye: a critical review. *Journal of Optometry*, 2(1), 3-18.
- Neto, A. M., Victorino, A. C., Fantoni, I., Zampieri, D. E., Ferreira, J. V., & Lima, D. A. (2013). Image processing using Pearson's correlation coefficient: Applications on autonomous robotics. *13th International Conference on Autonomous Robot Systems (1-6)*. IEEE.
- Nguyen, D. T., Fahimi, A., Fink, W., Nazemi, P. P., Kim, J. K., & Sadun, A. A. (2009). Novel 3D computer-automated threshold Amsler grid visual field testing of scotomas in patients with glaucoma. *European journal of ophthalmology*, 19(5), 776-782.
- Nisar, S., Khan, M. A., Algarni, F., Wakeel, A., Uddin, M. I., & Ullah, I. (2022). Speech recognition-based automated visual acuity testing with adaptive mel filter bank. *Comput. Syst. Sci. Eng*, 70, 2991-3004.

- Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23, 94–140.
- Numata, T., Matsumoto, C., Okuyama, S., Tanabe, F., Hashimoto, S., Nomoto, H., & Shimomura, Y. (2016). Detectability of visual field defects in glaucoma with high-resolution perimetry. *Journal of Glaucoma*, 25(10), 847-853.
- O'Leary, N., Chauhan, B. C., & Artes, P. H. (2012). Visual field progression in glaucoma: estimating the overall significance of deterioration with permutation analyses of pointwise linear regression (PoPLR). *Investigative ophthalmology & visual science*, 53(11), 6776-6784.
- Park, H. Y. L., Lee, J., & Park, C. K. (2018). Visual field tests for glaucoma patients with initial macular damage: comparison between frequency-doubling technology and standard automated perimetry using 24-2 or 10-2 visual fields. *Journal of glaucoma*, 27(7), 627-634.
- Park, S. C., Kung, Y., Su, D., Simonson, J. L., Furlanetto, R. L., Liebmann, J. M., & Ritch, R. (2013). Parafoveal scotoma progression in glaucoma: Humphrey 10-2 versus 24-2 visual field analysis. *Ophthalmology*, 120(8), 1546-1550.
- Parrinello, C. M., Grams, M. E., Sang, Y., Couper, D., Wruck, L. M., Li, D., & Coresh, J. (2016). Iterative outlier removal: a method for identifying outliers in laboratory recalibration studies. *Clinical chemistry*, 62(7), 966-972.
- Pelli, D. and J. Robson (1991). "Are letters better than gratings?" *Clinical vision sciences* 6(5): 409-411.
- Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10, 437-442.
- Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision research*, 46(28), 4646-4674.
- Phelps, C. D. (1984). Acuity perimetry and glaucoma. *Transactions of the American Ophthalmological Society*, 82, 753.
- Phipps, J. A., Zele, A. J., Dang, T., & Vingrys, A. J. (2001). Fast psychophysical procedures for clinical testing. *Clinical and Experimental Optometry*, 84(5), 264-269.
- Prea, S. M., Kong, G. Y., Guymer, R. H., & Vingrys, A. J. (2021). Uptake, persistence, and performance of weekly home monitoring of visual field in a large cohort of patients with glaucoma. *American Journal of Ophthalmology*, 223, 286-295.
- Prins, N. (2016). *Psychophysics: a practical introduction*. Academic Press.
- Provis, J. M., Penfold, P. L., Cornish, E. E., Sandercoe, T. M., & Madigan, M. C. (2005). Anatomy and development of the macula: specialisation and the vulnerability to macular degeneration. *Clinical and Experimental Optometry*, 88(5), 269-281.

- Quigley, H. A., & Broman, A. T. (2006). The number of people with glaucoma worldwide in 2010 and 2020. *British journal of ophthalmology*, 90(3), 262-267.
- Racette, L., Fischer, M., Bebie, H., Holló, G., Johnson, C. A., & Matsumoto, C. (2016). Visual field digest. A guide to perimetry and the Ocotpus perimeter, 6.
- Rankin, S. J. (1999). Tumbling E resolution perimetry in glaucoma. In *Perimetry Update 1998/1999: Proceedings of the XIIIth International Perimetric Society Meeting*, (179-186).
- Rao, H. L., Senthil, S., Choudhari, N. S., Mandal, A. K., & Garudadri, C. S. (2013). Behavior of visual field index in advanced glaucoma. *Investigative ophthalmology & visual science*, 54(1), 307-312.
- Reich, L. N. and H. E. Bedell (2000). "Relative legibility and confusions of letter acuity targets in the peripheral and central retina." *Optometry and Vision Science* 77(5): 270-
- Robson, J. G., & Graham, N. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, 21(3), 409–418.
- Rodriguez-Una, I., & Azuara-Blanco, A. (2018). New technologies for glaucoma detection. *The Asia-Pacific Journal of Ophthalmology*, 7(6), 394-404.
- Rosén, R., Lundström, L., & Unsbo, P. (2011). Influence of optical defocus on peripheral vision. *Investigative Ophthalmology & Visual Science*, 52(1), 318-323.
- Rosser, D. A., Cousens, S. N., Murdoch, I. E., Fitzke, F. W., & Laidlaw, D. A. (2003). How sensitive to clinical change are ETDRS logMAR visual acuity measurements? *Investigative ophthalmology & visual science*, 44(8), 3278-3281.
- Rosser, D. A., Murdoch, I. E., & Cousens, S. N. (2004). The effect of optical defocus on the test–retest variability of visual acuity measurements. *Investigative ophthalmology & visual science*, 45(4), 1076-1079.
- Rossi, E. A., & Roorda, A. (2010). The relationship between visual resolution and cone spacing in the human fovea. *Nature neuroscience*, 13(2), 156-157.
- Sabeti, F., Lane, J., Rohan, E. M., Rai, B. B., Essex, R. W., McKone, E., & Maddess, T. (2021). Correlation of central versus peripheral macular structure-function with acuity in age-related macular degeneration. *Translational Vision Science & Technology*, 10(2), 10-10.
- Selvaraaju Murugesan (2022). Speech Recognition, MATLAB Central File Exchange. (<https://www.mathworks.com/matlabcentral/fileexchange/47220-speechrecognition>),
- Shah, N., Dakin, S. C., & Anderson, R. S. (2012). Effect of optical defocus on detection and recognition of vanishing optotype letters in the fovea and periphery. *Investigative ophthalmology & visual science*, 53(11), 7063-7070.

- Shah, N., Dakin, S. C., Redmond, T., & Anderson, R. S. (2011). Vanishing Optotype acuity: repeatability and effect of the number of alternatives. *Ophthalmic and Physiological Optics*, 31(1), 17-22.
- Shah, N., Laidlaw, D. A. H., Shah, S. P., Sivasubramaniam, S., Bunce, C., & Cousens, S. (2011). Computerized repeating and averaging improve the test-retest variability of ETDRS visual acuity measurements: implications for sensitivity and specificity. *Investigative ophthalmology & visual science*, 52(13), 9397-9402.
- Sharma, P., Sample, P. A., Zangwill, L. M., & Schuman, J. S. (2008). Diagnostic tools for glaucoma detection and management. *Survey of ophthalmology*, 53(6), S17-S32.
- Siderov, J., & Tiu, A. L. (1999). Variability of measurements of visual acuity in a large eye clinic. *Acta Ophthalmologica Scandinavica*, 77(6), 673-676.
- Sloan, L. L. (1959). "New test Charts for the Measurement of Visual Acuity at far and Near Distances*." *American Journal of Ophthalmology* 48(6): 807-813.
- Smith, J. E. K. (1992). Alternative biased choice models. *Mathematical Social Sciences*, 23,199–219.
- Smith, J. K. (1982). Recognition models evaluated: A commentary on Keren and Baggen. *Perception & psychophysics*, 31, 183-189.
- Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of vision*, 11(5), 13-13.
- Sung, K. R., Wollstein, G., Kim, N. R., Na, J. H., Nevins, J. E., Kim, C. Y., & Schuman, J. S. (2012). Macular assessment using optical coherence tomography for glaucoma diagnosis. *British journal of ophthalmology*, 96(12), 1452-1455.
- Swanson, W. H., & Birch, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & psychophysics*, 51(5), 409-422.
- Tanna, A. P., & Desai, R. U. (2014). Evaluation of visual field progression in glaucoma. *Current Ophthalmology Reports*, 2, 75-79.
- Thibos, L. N., Still, D. L., & Bradley, A. (1996). Characterization of spatial aliasing and contrast sensitivity in peripheral vision. *Vision research*, 36(2), 249-258.
- Thibos, L. N., Walsh, D. J., & Cheney, F. E. (1987). Vision beyond the resolution limit: aliasing in the periphery. *Vision Research*, 27(12), 2193-2197.
- Tomarek, R. H., Aboud, S. A., Hassan, M., & Mohamed, A. H. (2020). Studying the role of 10-2 visual field test in different stages of glaucoma. *European Journal of Ophthalmology*, 30(4), 706-713.
- Townsend, J. T. (1971). Alphabetic confusion: A test of models for individuals. *Perception & Psychophysics*, 9, 449-454. (b)
- Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 9, 40-50. (a)

- Treacy, M. P., Hurst, T. P., Conway, M., Duignan, E. S., Dimitrov, B. D., Brennan, N., & Cassidy, L. (2015). The early treatment in diabetic retinopathy study chart compared with the tumbling-E and Landolt-C. *Ophthalmology*, 122(5), 1062-1063.
- Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision research*, 35(17), 2503-2522.
- Trevino, R. (2008). Recent progress in macular function self-assessment. *Ophthalmic and Physiological Optics*, 28(3), 183-192.
- Tsamis, E., La Bruna, S., Leshno, A., De Moraes, C. G., & Hood, D. (2022). Detection of early glaucomatous damage: performance of summary statistics from optical coherence tomography and perimetry. *Translational Vision Science & Technology*, 11(3), 36-36.
- Vanden Bosch, M. E., & Wall, M. (1997). Visual acuity scored by the letter-by-letter or probit methods has lower retest variability than the line assignment method. *Eye*, 11(3), 411-417.
- Vesti, E., Johnson, C. A., & Chauhan, B. C. (2003). Comparison of different methods for detecting glaucomatous visual field progression. *Investigative ophthalmology & visual science*, 44(9), 3873-3879.
- Wang, Y. Z., He, Y. G., Mitzel, G., Zhang, S., & Bartlett, M. (2013). Handheld shape discrimination hyperacuity test on a mobile device for remote monitoring of visual function in maculopathy. *Investigative ophthalmology & visual science*, 54(8), 5497-5505.
- Wang, Y. Z., Thibos, L. N., & Bradley, A. (1997). Effects of refractive error on detection acuity and resolution acuity in peripheral vision. *Investigative ophthalmology & visual science*, 38(10), 2134-2143.
- Watson, A. B., & Ahumada, A. J. (2015). Letter identification and the neural image classifier. *Journal of Vision*, 15(2):15, 1–26. <https://doi.org/10.1167/15.2.15>
- Watson, A. B., & Ahumada, A. J. Jr., (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, 5(9), 717–740.
- Wen, Y., Chen, Z., Zuo, C., Yang, Y., Xu, J., Kong, Y., & Yu, M. (2021). Low-Contrast High-Pass Visual Acuity Might Help to Detect Glaucoma Damage: A Structure-Function Analysis. *Frontiers in Medicine*, 8.
- Westheimer, G., Chu, P., Huang, W., Tran, T., & Dister, R. (2003). Visual acuity with reversed-contrast charts: II. Clinical investigation. *Optometry and vision science*, 80(11), 749-752.
- Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8), 1293-1313.
- Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford & New York: Oxford University Press.

- Wild, J. M. (2001). Short wavelength automated perimetry. *Acta Ophthalmologica Scandinavica*, 79(6), 546-559.
- Williams, D. R., Artal, P., Navarro, R., McMahon, M. J., & Brainard, D. H. (1996). Off-axis optical quality and retinal sampling in the human eye. *Vision research*, 36(8), 1103-1114.
- Zhang, C., Tatham, A. J., Weinreb, R. N., Zangwill, L. M., Yang, Z., Zhang, J. Z., & Medeiros, F. A. (2014). Relationship between ganglion cell layer thickness and estimated retinal ganglion cell counts in the glaucomatous macula. *Ophthalmology*, 121(12), 2371-23
- Zhou, J., Georgeson, M. A., & Hess, R. F. (2014). Linear binocular combination of responses to contrast modulation: Contrast-weighted summation in first-and second-order vision. *Journal of vision*, 14(13), 24-24.