Vision Research 208 (2023) 108233

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

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 Schmidtmann [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

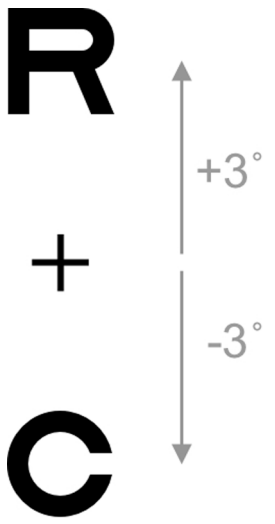

Fig. 1. Sloan letters presented singly, either centrally or along the vertical meridian at an eccentricity of 3° in the upper (+) or lower (-) visual field. Sloan letters R and C are shown for illustration purposes (not to scale).

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*

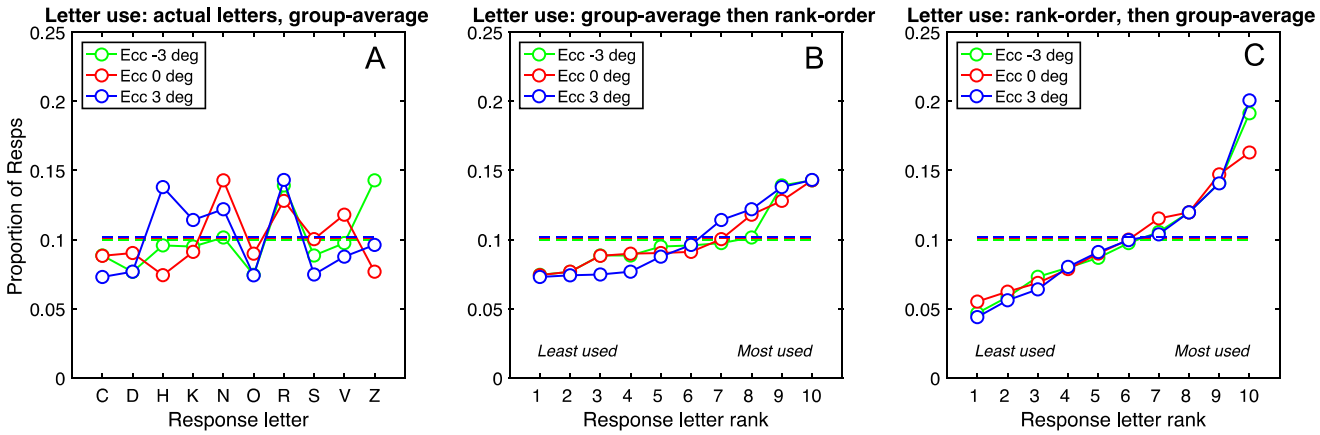

Fig. 2. Data overview: letter usage. A) Proportion of trials in the experiment on which each of the 10 test letters was reported (correctly or not), averaged over the 6 letter sizes and 10 observers. Each point is based on 600 trials. Colours indicate the 3 test locations. B) Same data as A, but rank-ordered 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.

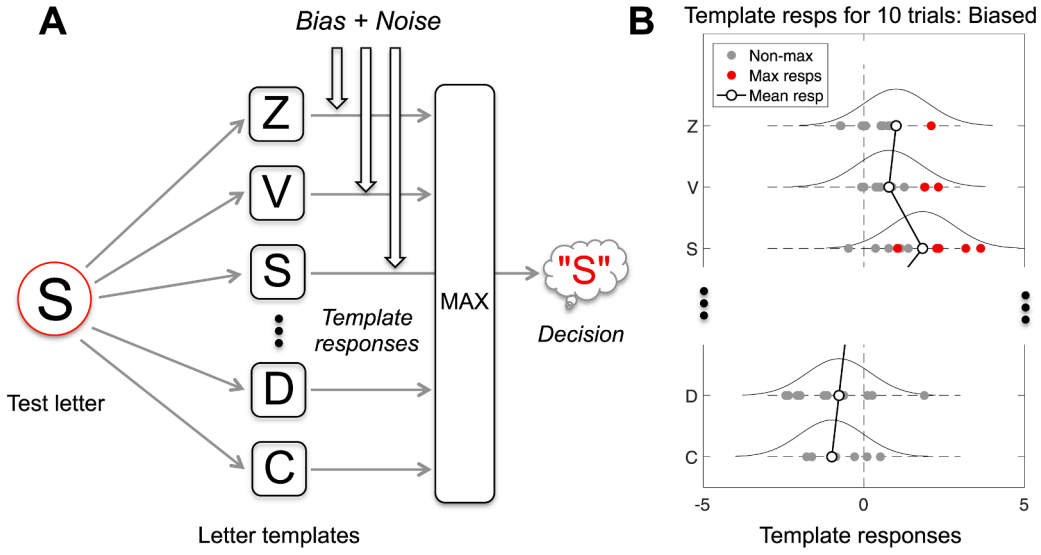

Fig. 3. A) Schematic view of the noisy template model. The most active template on a given trial determines the letter choice made. B) Because of noise, the most active template over trials (red dots) may be the correct one (e.g. S) or an incorrect one (e.g. V). Positive bias (top 3 rows) increases the 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° 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).

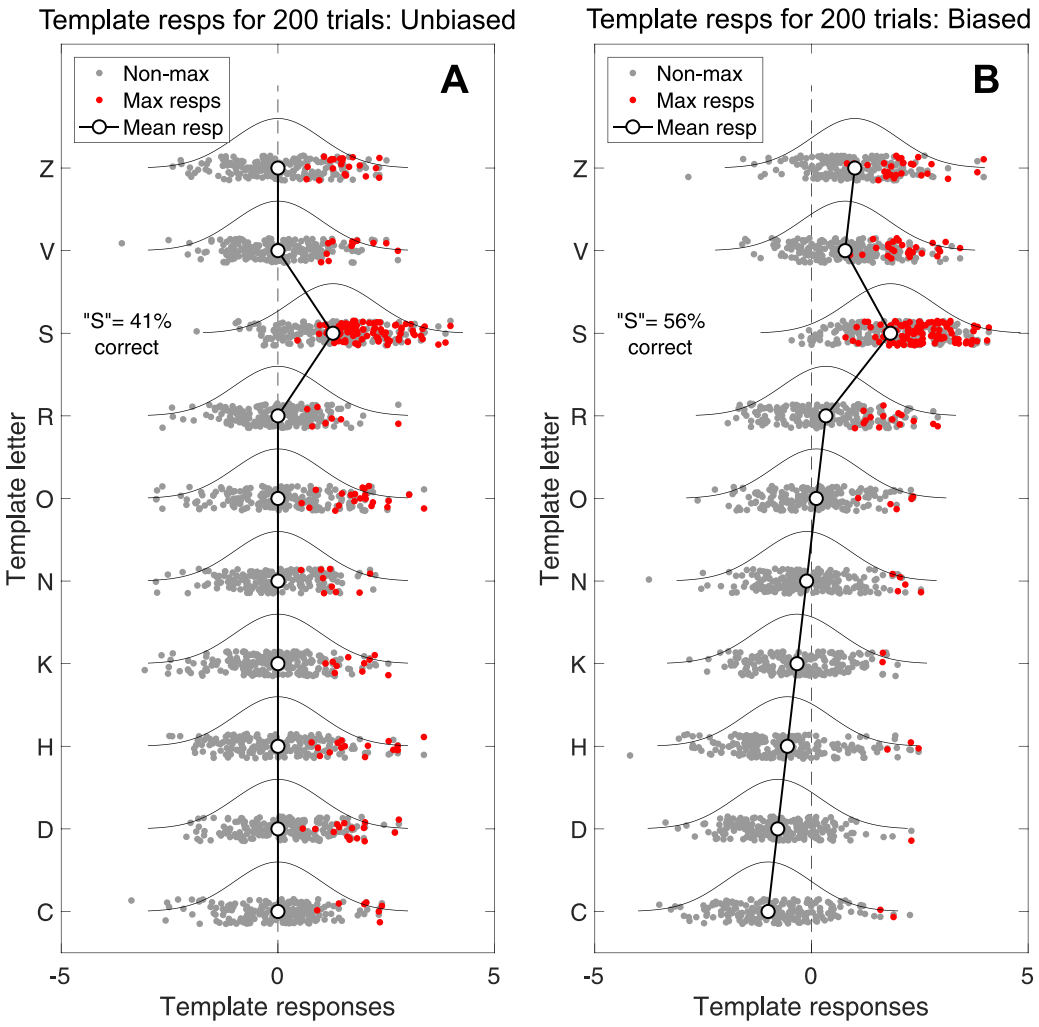

Fig. 4. How the noisy template model works, either without bias (A), or with different biases on each template (B). Bias in panel B is ordered from negative through to positive with a mean of zero. We call this the 'bias gradient', *B'*. Letter identity for each template here is arbitrary, for illustration 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:

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

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:

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

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):

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

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 *mAFC* (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 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

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

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 ($P_{lapse} = 0.01$). 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 $1\text{-}P_{lapse}$) 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 $P_C^* = P_C(1 - P_{lapse}) + (1/m)P_{lapse}$, 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).

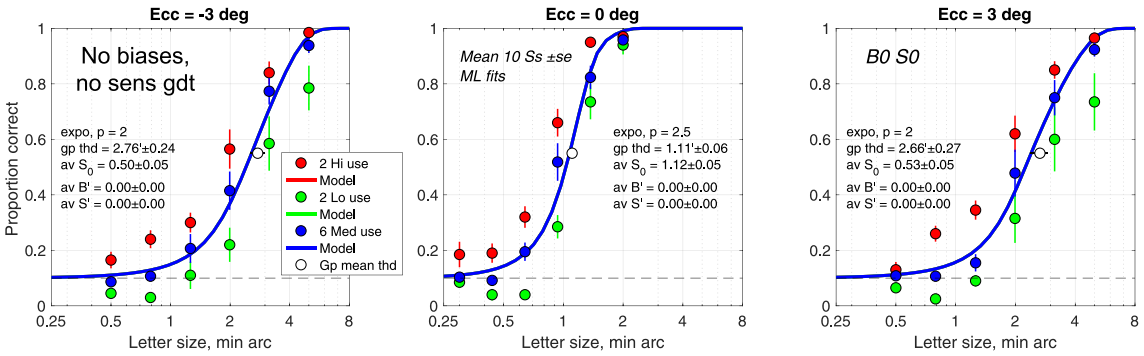

Fig. 5. Data overview: psychometric functions. Some letters are used as a response more than others, across all letter sizes. Red symbols plot the proportion of correct trials averaged over the two most-used letters, identified separately for each observer, as in Fig. 2C, then averaged over observers. 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.)

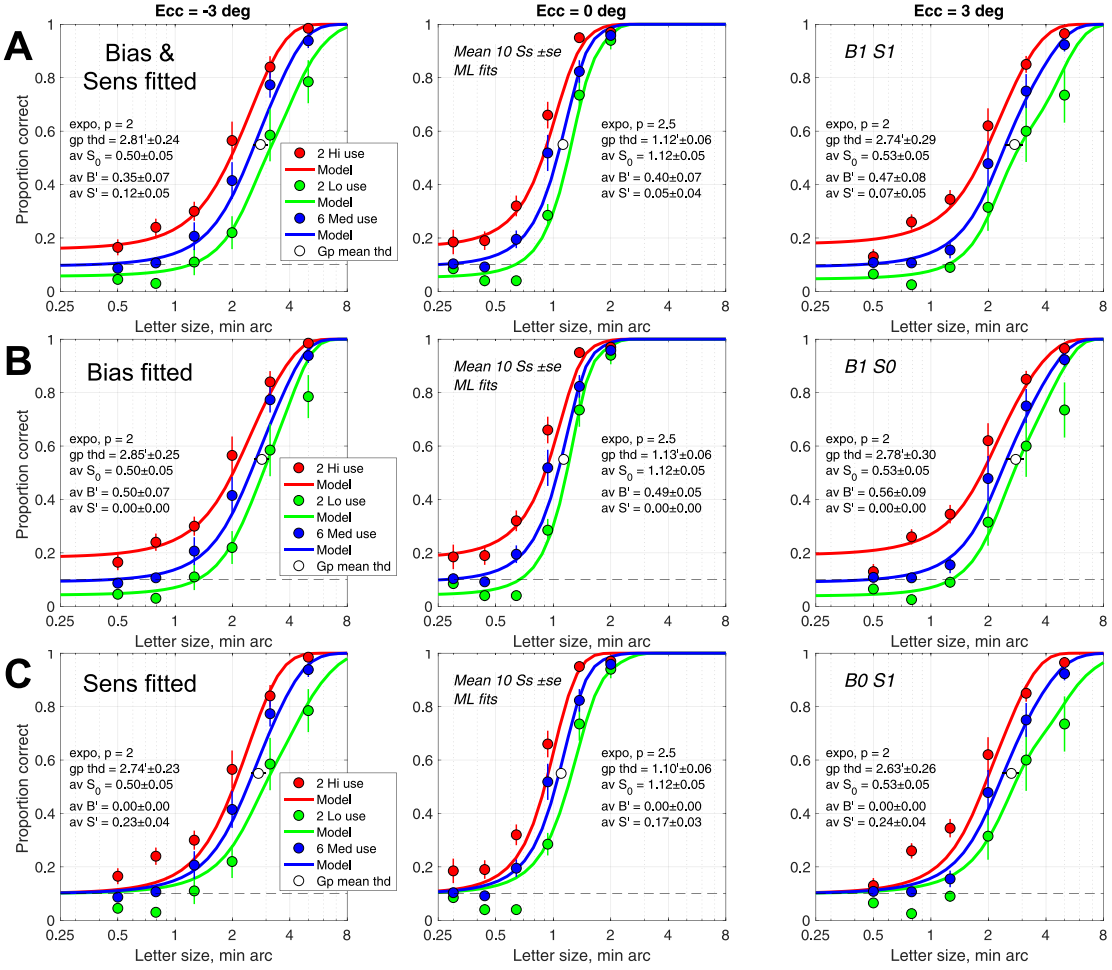

Fig. 6. The noisy template model was fitted to the whole dataset (row A) by choosing the best-fitting gradients of bias & sensitivity across the letter templates, or (B) by fitting the bias gradient while sensitivity was the same for all letters, or (C) fitting the sensitivity gradient, while bias and bias-gradient were zero. In each case the model was fitted to data of individual subjects; these graphs show the group mean data (±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) Model B0 S1 vs B1 S1			
−3°	77.53	10	< 0.000001
0°	93.72	10	< 0.000001
+3°	99.03	10	< 0.000001
(ii) Model B1 S0 vs B1 S1			
−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

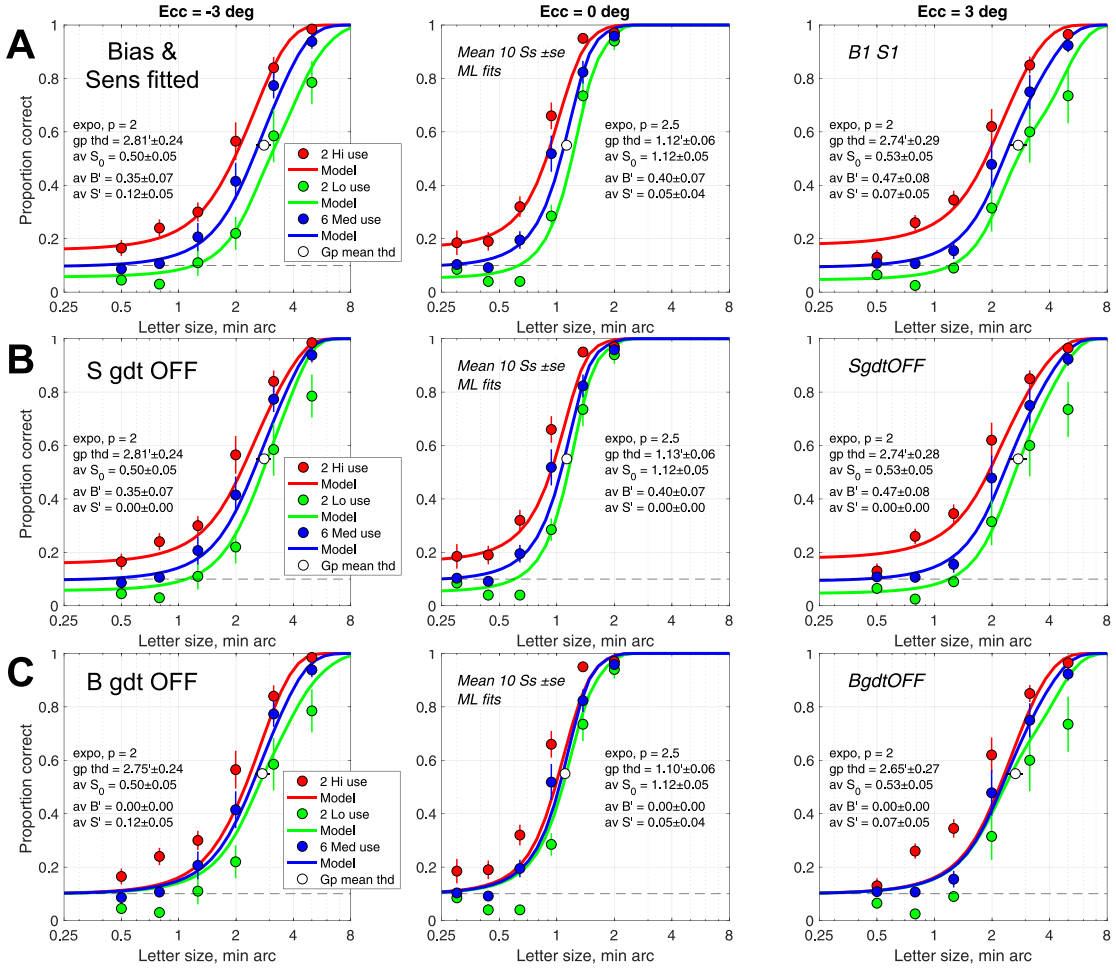

Fig. 7. How important were the *B* gradient and *S* gradient in the best-fitting two-factor model? (A) Fit of the two-factor model, copied from Fig. 6A. (B) The model was re-computed with *B* gradient unchanged, but the *S* gradient set to 0 (ie. same template sensitivity S_0 for all letters). (C) The 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.

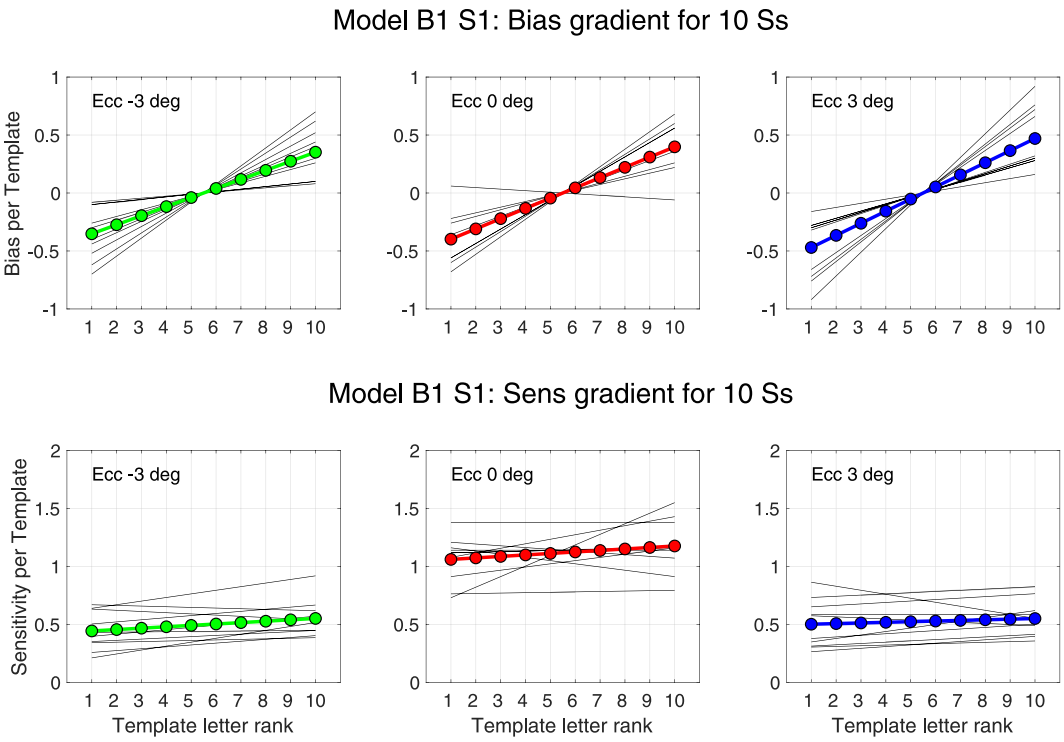

Fig. 8. Variation in bias and sensitivity across letters in the full model (B1 S1). Top row: Average linear trend in modelled bias (coloured symbols) was similar across the 3 test locations. Ordinate plots bias values B_i, for $i = 1$ to 10, ranging from about -0.4 σ to $+ 0.4$ σ, where σ = 1 is the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

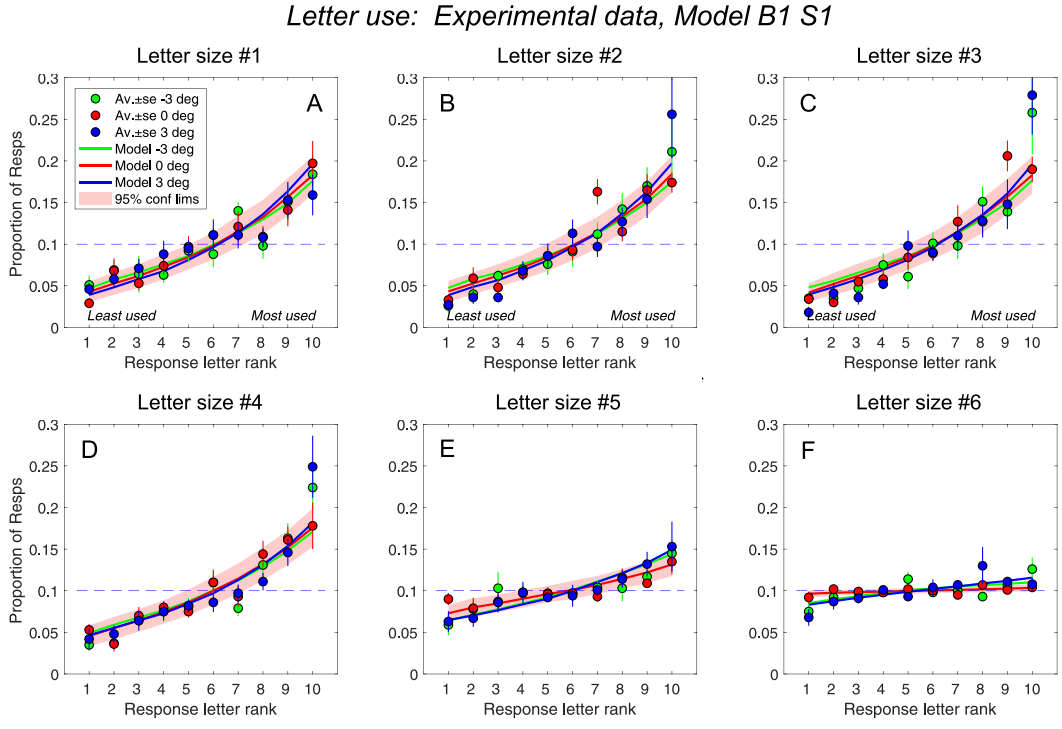

Fig. 9. Data & model compared via letter usage, separated out according to letter size. (A - F) Each panel relates to one of the 6 test letter sizes, from smallest to largest, as marked. Ordinate plots proportion of trials a given letter was named, whether correct or not. Symbols are group-average data ± 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 underestimation 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 prestriate 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:	C D H K N O R S V Z	C D H K N O R S V Z	C D H K N O R S V Z
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 ± S.D.	0.159 ± 0.345	0.248 ± 0.254	0.214 ± 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

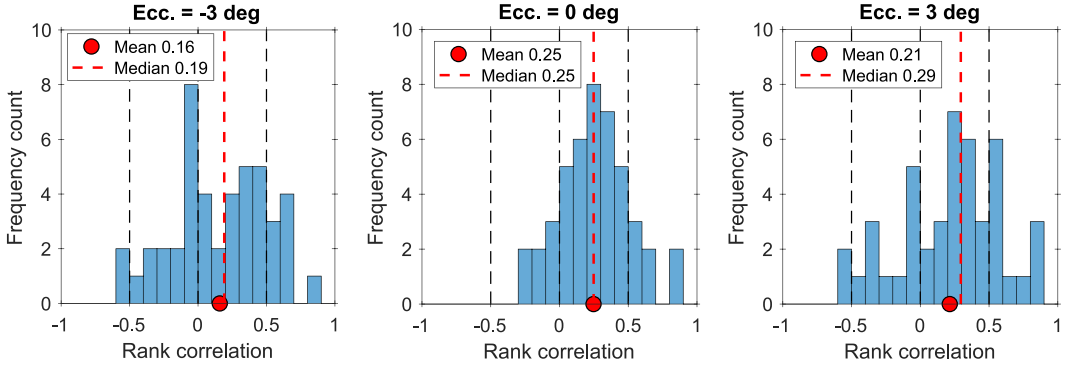

Fig. A1. For each of the three test locations, histograms show the distributions of between-subject similarity in *usage* (rank correlation values) for the 45 pairings of 10 observers. Red symbol shows the mean correlation, dashed red line the median. Dashed black lines mark correlation values of −0.5, 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).

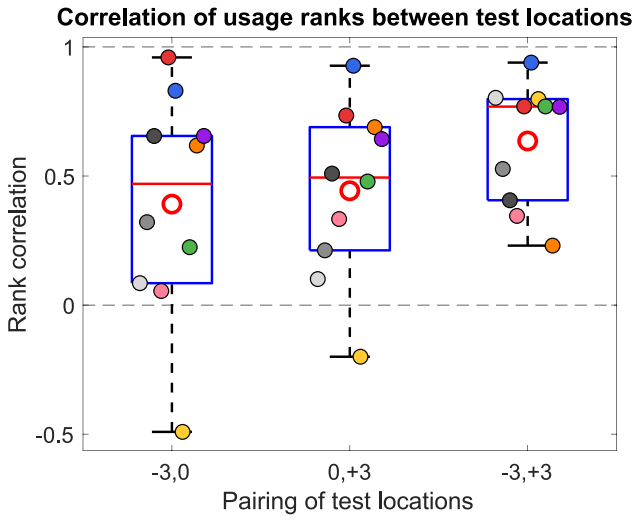

Fig. A2. For each observer, we computed the similarity in letter usage (rank correlation) across a given pair of test locations. These individual-subject values are plotted as filled symbols, with a different colour for each observer. To avoid overlap, observers S1 to S10 are displaced from left to right 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 ± S.D.	0.391 ± 0.438	0.443 ± 0.337	0.636 ± 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.

Barhoom, 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., Dziugaite, 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