Title

Optotype recognition under degradation: comparison of size, contrast, blur, noise and contour-perturbation effects

Journal

Clinical and Experimental Optometry, 99(1)

ISSN

0816-4622

Author

Westheimer, Gerald

Publication Date

2016

DOI

10.1111/cxo.12293

Peer reviewed

Optotype recognition under degradation: comparison of size, contrast, blur, noise and contour-perturbation effects

Running title: Optotype recognition under degradation

[Accepted for publication in Clinical and Experimental Optometry Vol. 99, 2016]

Gerald Westheimer ASTC, PhD , FAAO, FRS

Division of Neurobiology
University of California
Berkeley, CA 94720-3200

gwestheimer@berkeley.edu

Background. Visual acuity is measured by diminishing letter size till recognition threshold is reached, letters varying in legibility. In this experiment size reduction was compared with other means of handicapping letter legibility.
Methods. In five normal observers, discrimination thresholds for 13 sans-serif capital letters in a 5x4 format were obtained by a staircase procedure for size reduction as well as for 20 arcmin (logMAR=0.6) letters subjected to four kinds of image degradation: reducing contrast, convolving with blurring spread, embedding in noise and perturbing contour smoothness.
Results. Threshold correlation and distribution of response errors show prominent differences and indicate the degree to which the visual processing of the various modes of image degradation is distinct.
Conclusions. The validation of four other ways of impairing the recognition of optotypes in addition to size diminution reveals their potential in the differential diagnosis of defects in pattern detection, in evaluating therapeutic regimens, and in developing concepts of form perception.
Keywords: Visual acuity, contrast sensitivity, image blur, form perception

The traditional test of a patient's feature recognition competence is visual acuity in which letters are sequentially reduced in size till a threshold is reached. Performance for individual optotype letters depends to a varying degree on resolution (C or O?), detection of contour difference (V or U?) or of limb numerosity (P or R?). Hence, even within selected sets, such as those of Sloan[1] or Bailey-Lovie[2], not all letters are equally legible nor may different visual dysfunctions result in the same deterioration.[3,4]

Besides minification, there are other ways of feature degradation. Chief among them is contrast reduction, exemplified by the Pelli-Robson,[5] SKILL[6] and Reagan[7] charts, in which there is signal attenuation in the brightness domain. Letter recognition can also be compromised by blurring[8], where edge sharpness is flattened, or through pattern corruption by superimposed noise or perturbed smoothness of contours.

These methods do not interfere with vision at the same processing level. Resolution is most affected by optical spread and retinal compartmentalization; contrast by photochemical and neural retinal factors. Form discrimination surely involves early cortical visual stages. For these reasons, as well as ones of clinical utility, it is of interest to look for characteristic differences in the effect of these several methods of degradation on the legibility of standard optotype letters.

The question is here approached empirically by characterizing each of the five kinds of image degradation by a single parameter which can then be varied to determine recognition threshold. In ordinary visual acuity testing, recognition threshold is, of course, specified by letter size in minutes of arc subtended at the eye. Contrast is best identified, for a fixed letter size and background luminance L, by the Weber fraction $\Delta L/L$. Blurring is accomplished by convolving each pattern element with a Gaussian spread k, where the parameter k is the distance in minutes of arc at which the spread function has fallen to 36% of its peak. Noise can be added to patterns by randomly inverting the contrast polarity of each pattern pixel with increasing probability. Finally, contour smoothness can be perturbed by displacing rows and columns of the pixel matrix containing the patterns by random amounts with increasing amplitude.

METHODS

Stimuli

At each presentation, the observers had to identify dark upper case letters shown one at a time against a circumscribed white background on a computer screen. The letters were randomly selected from an ensemble of 13 (C, D, F, H, K, N, O, P, R, S, U, V and Z) 5x4 sans-serif letters very similar in appearance to bold capital letters in the Helvetica or Arial fonts with which the observers were familiar. The letter remained exposed until the observer responded by activating a computer key which registered the response; if incorrect, the observer was briefly shown the correct one as an error feedback.

The letters were generated on the computer screen in a *20x16* pixel rectangular matrix, in the manner of Snellen letters with each limb 4 pixels wide. Standard LCD computer screens were used at a viewing distance such that each pixel subtended *1* arcmin at the eye. Except when size was the variable, this gave letters subtending *20* arcmin (logMAR=*0.6*, or *6/24*, or *20/80*). Observation was binocular. Screen luminance was typical of high-grade LCD displays, i.e., in the range of *120-150* cd/m^2

Five methods of degrading the targets were used to measure recognition threshold by a staircase procedure. In each case, degradation proceeded by sequential change in a single parameter characterizing decreasing image quality. The quantities are given expression so that ascending threshold values signify better performance, that is, greater robustness to degradation. The typical screen appearance, reproduced here as best as possible, of the letter N in moderate levels of degradation of the five kinds is shown in Fig. 1.

Size: The overall size of the letters, measured in arcmins. was diminished, limb size remaining in proportion. To conform with the other methods where the score increases with performance improvement, a reciprocal measure is used as the parameter. Increment and decrement steps in the staircase were logMAR(0.05).

Contrast: Denoting the luminance of the letters by L and that of the background L_B , contrast was defined by the Weber fraction $(L_B - L)/L_B$ which is close to unity in the full black-on-white condition and approaches zero as contrast vanishes. In the experiment, logarithmic steps were employed and the performance registered in a reciprocal measure ($10 * \log (L_B/\Delta L)$) of contrast, in conformity with the procedure in all the tests that higher numbers mean better performance. However, where numerical comparisons were made between the five kinds of image degradation, contrasts were expressed in a linear rather than logarithmic scale to preserve uniformity.

Blur: The letters were convolved with a circularly symmetrical Gaussian spread whose parameter k, i.e., the distance from its peak to 36% height, was expressed in arcmins. A higher k value for letters that can still be recognized indicates better performance. To decouple the task from overall contrast discrimination, all patterns were given a fixed contrast: the luminance of darkest point of the convolved patterns was 50% of the bright background.

Noise: Each pixel of the whole display area (full contrast sharp letters within their 20x16 matrix and background), had its contrast polarity inverted, i.e., black became white and white black, randomly with a probability p $(0 > p > 0.5)$. This obscured the letters increasingly with the probability, which was used as the measure; the higher the probability the more robust a letter to noise corruption and the better the performance (Fig. 2B).

Contour Perturbation. The smoothness of the pattern contours was disrupted by subjecting rows and columns of the letter matrix to random displacements. Increase in the magnitude of the distribution of the displacement, in pixels, increasingly disturbs the pattern form, the higher the more resistant form perception of the particular letter to disruption of pattern coherence (Fig. 2C)

Procedure

Recognition thresholds were obtained for all 13 letters comprising the series in long runs of trials, each devoted to only one kind of degradation, in which single letters were presented one at a time. They were chosen at random on each presentation and had a degradation level that depended on the stage of the staircase for that particular letter. After three consecutive correct responses for the letter, the task was made more difficult (the parameter value increased) until two consecutive responses for that letter were incorrect, whereupon that letter at the subsequent presentation was shown with a one-step decrease in the parameter. Such up and down staircases were inderdigitated for all letters until sufficient reversals (~30) had been registered for each to compute a stable threshold.. This required a total of ~2000 responses, usually acquired in two sessions, each lasting about one hour, a day apart. Progression in the five classes of staircases was such that about 4 steps separated the levels definitely above to definitely below threshold (i.e., always to never correctly recognized).

Recognition threshold was the average of the parameter values of an even number of reversals, which also allowed the computation of standard errors. In this manner, a relative order of legibility for the 13 letters can be found separately and independently for each kind of image degradation. It is to be noted that the results depend critically and are valid only for the ensembles used. For example, legibility ranking might not be the same in ensembles containing different letters such as "I" or "W." The observer's response was required to be a letter within the specific ensemble; this allowed the generation of a confusion matrix that revealed, for example, how frequently a presented "V" was reported to be a "U." The staircase thresholds are unaffected by forcing the response to be a letter instead of merely yes/no.

Data Analysis

After completion of the runs for a given observer and method of degradation, 13 threshold values, together with their standard error, were available. A correlation coefficient was then calculated for all 4 + 3 + 2 +1 =10 pairs of degradations. The higher the correlation coefficient the more similar the sequence of letters in the series of increasing robustness to degradation of the 13 letters in the ensemble, and the more alike are the two kinds of degradation in the way they interfere with form perception.

In addition to utilizing correct and incorrect responses to identify recognition thresholds, erroneous responses were accumulated in bins according to stimulus and response letters in a 13x13 array that allowed further study of the error distribution.

Observers

Five observers participated: a senior researcher familiar with the aims of the project, and four biology undergraduate students, two male and two female, who were unaware of the direction of research when performing the experiments but had broad familiarity with and some experience in visual psychophysics. All had normal corrected visual acuity and their optometric status was unexceptional for the purposes of the study, which conformed to the Declaration of Helsinki and was approved by the Campus Committee for the Protection of Human Subjects.

Whereas it has been shown previously that there is no significant perceptual learning in foveal acuity and blur discrimination tasks, no such assurance is in place for the other three stress methods. Data shown here were, therefore, obtained only after some preliminary runs with these stimulus conditions.

RESULTS

Raw data available are values of the five degradation parameters for threshold recognition of 13 optotype letters in five observers, expressed so that the higher the value the more legible the letter. To facilitate comparison, for each observer and for each kind of degradation, the recognition threshold values of the 13 individual letters were normalized: the actual thresholds were divided by the average of the 13 thresholds.

The first question relates to the relative compactness of the five kinds of degradation – how does the spread among the set of 13 letter thresholds compare? Table I shows the average deviation from the 13-letter mean for the five kinds of degradation, across all observers. Because contrast was measured along a logarithmic scale, the contrast value was recalculated to align it with the others which were scaled linearly. The scatter in legibility among the 13 letters was somewhat less in the standard visual acuity measurement (size) than when the handicap was contrast reduction or blur, noise or perturbation. It remains to be determined whether this is to be ascribed to better familiarity of typical normal observers with the challenge of discriminating shapes near their resolution threshold. However, discrimination of blurred shapes, for one, does not improve with training.[9]

One motive for conducting this study was to find how comparable the effect of the various kinds of degradation was on individual letters. If one ranks the 13 letters according to their threshold decrement under one kind of degradation, how well does this match that under the others? This is illustrated for a typical observer and for the average of all observers in Fig. 3. The letters in each case are ordered in the sequence of monotonically increasing legibility that applied to the average in visual acuity. This procedure reveals that the various degradations indeed differ in their manner of interfering with form perception.

The differences can be expressed quantitatively by the correlation coefficients between the various pairing of the degradations. The more similar the effect on form perception of two kinds of degradation, the more highly correlated their thresholds in the ensemble of 13 letters. Table II gives quantitative expression to the closeness of the various kind of degradation to each other in their operation at the various stages of processing of form. The standard deviation of a correlation coefficient *r* based on *n* data pairs[10] is approximately $(1 - r^2)/\sqrt{n}$ and it follows that strong conclusions can be drawn only from data sets in which *r* approaches unity or *n* is large. By normalizing the data for each of the observers for each condition, all data for a given letter in each degradation category can be pooled. Then $n = 65$ and the standard error of the correlation coefficients is of the order of *0.1.*

The entries in Table II can be regarded as the inverse of the distances separating the processing mechanisms responsible for discrimination within the ensembles, across all letters and observers, in the five kinds of degradation, or expressed differently, the degree of the overlap or commonality of the neural circuits subserving them. Discrimination between patterns of low contrast appears to be remote from that of blur and traditional acuity, whereas perturbation and noise, and to some extent blur and acuity, seem to share processing channels.

In the analysis so far, all erroneous responses had been lumped together and no distinction made between, say, a "**K"** or "**N"** response to an "**H**." During data acquisition, however, this information had been preserved: responses to each of the *13* letters had been accumulated not only in correct and error categories, but also partitioned into subcategories for each letter of the ensemble of *13*, creating a confusion matrix. The five confusion matrices are displayed in Fig. 4, summed for all observers and stripped of the high values in the diagonals which represent correct responses. The confusion matrices were subjected to two kinds of analysis. One was to compute pairwise correlation coefficients, giving a measure of the similarity of the distribution of errors. The other is more specific: the difference between any two matrices was computed, cell by corresponding cell, and the average of all the differences calculated. This measure has the advantage of factoring out response idiosyncrasies of individual observers, for example, letter preferences, that would be common to all five kinds of image quality diminution.

The results of these two analytical procedures are shown in Table III. The quantities in the panels are derived from pooling the values for all observers in a *13x13* array of cells, presented letters along one axis, responses along the other. The standard errors of the statistics are then small enough to allow fairly specific conclusions. When, in two degradation regimes, the letter recognition task is accomplished by the same neural processing apparatus, the distribution and categorization of errors would be similar. On the other hand, lack of commonality of underlying mechanisms would manifest itself by low correlations and low matching of errors. The data in Tables II and III reveal some trends, e.g., corruption through noise and contour perturbations are highly correlated. But the methodology on the whole does not yet seem able to make the sharp distinctions that would enable the delineation of separate mechanisms operating in the realm of form perception.

DISCUSSION

In the century and a half since visual acuity has been formalized and standardized by Snellen and Landolt, it has been unsurpassed as a clinical probe of vision. Reducing high contrast letters in size till they can no longer be recognized serves indeed as an excellent and easily-administered measure of visual functioning; current letter charts would not have survived if there had been a better alternative. Some of the vagaries have been removed by better chart design, but some remain, for example the problem of fine gradations to determine gain or loss at diminished levels, [11] and factors involving memory[12] and scoring of results.[13]

Even staying with letter recognition as the stimulus/response medium (which in modern literate society cannot easily be equaled), merely reducing size is not a unique method of ascertaining the functional limit of spatial vision in a clinical setting. To be sure, it highlights the resolution aspect with its emphasis on optical quality of the retinal image and the integrity of the receptor mosaic, and is especially effective when finding the end point for correcting refractive errors. But from the earliest days, visual acuity has been understood to transcend optical resolution. The term *form sense*[14] was introduced in the diagnosis of the deficit when patients, who had no difficulty detecting gaps in contours or the relationship between pattern components, failed to identify letters or shapes. Hence the attempt here to find recognition limits when shapes are degraded by means other than size reduction.

Of the attributes considered, light quantity, long known to affect visual acuity, has been introduced in the clinic in the form of contrast reduction. It has been previously reported that the limits found in such tests do not fully parallel those in size reduction. This is also the case with blur as a means of image degradation.[8] In the use of optotypes, there is some resemblance of the present procedures with those of finding sampling limits[15,16] and the recent procedure of inducing artificial metamorphopsia.[17] However, the bulk of the extensive literature on shape recognition with diminished image quality (see review by Gold et al.[18]) is directed not so much toward clinical application as comparing human and ideal observers.[19] However, when the task is distinguishing between *6/24* (*20/80*) optotypes, that is, examining the relative disposition of feature components a few arcmins in size, analysis in terms of spatial-frequency channels would need phase rather than the usual amplitude consideration.

The methodology here employed seeks quantitative expression of the divergence between the modes of handicapping letter recognition in normal observers, utilizing the limiting value of the parameter that still allowed correct responses for each kind of degradation. The relative order of the discrimination threshold for the *13* letters in the ensemble could be visualised (Fig.2) and examined in a table of pairwise correlation coefficients (Table 2). Psychophysical threshold procedures inevitably include incorrect as well as correct responses. By registering responses not only as correct or incorrect, but also classifying errors according to the misperceived letter, a confusion matrix was accumulated in each experiment. Differences between the error distributions and correlation between them allowed a pairwise comparison of the five modes of image degradation (Fig. 3 and Table III). The hope of a clear separation was only partially fulfilled. It is apparent that size and contrast reduction involve largely different discrimination mechanisms, whereas noise and contour perturbation suggest some overlap. Whether more sophisticated analysis can make sharper distinctions and help establish rules governing shape discrimination remains for future studies.

Of more immediate relevance is the possible transfer of these findings to the clinic. The divergence of patient responses when letter recognition is impeded through size versus contrast reduction has already been mentioned, but the clinical value for differential diagnosis has yet to be explored. Added here to the repertoire are three more methods of image degradation. They share with contrast reduction a de-emphasis of resolution, in that their operating parameter is independent of target size, which can remain invariant and therefore factored out. Moreover, by utilizing the universal and unproblematic letter recognition paradigm, they need no additional patient instruction. The testing , scaling and scoring procedures having been individually validated and mutually compared in normal observers, they are here offered as candidates for clinical utility.

ACKNOWLEDGMENTS

Ryan N. Lee's effective contribution to the acquisition and presentation of the data is gratefully acknowledged. The author benefited from helpful discussion with Ian Bailey and Andrew Carkeet.

REFERENCES

[1]Sloan LL. A new test for the meassurement of visual acuiy at far and near distance." Am. J. Ophthalmology 1959; 48: 807-813.

[2]Bailey IL, J. E. Lovie JE New design principles for visual acuity letter charts. Am J Optom. Physiol Opt 1976; 53: 740-745

[3]Bennett AG Ophthalmic Test Types. Brit J Physiol Optics 1965;22: 238-271.

[4]McMonnies CW, Ho A.Letter legibilty and chart equivalance. Ophthalmic Physiol Opt 2000; 20: 142-152.

[5]Pelli DG, Robson JG, Wilkins AJ. The design of a new letter chart for measuring contrast sensitivity. Clinical Vision Sciences 1988; 2:187-199.

[6]Haegerstrom-Portnoy G, Brabyn J,Schneck ME,Jampolsky A. The SKILL Card: an acuity test of reduced luminance and contrast. Invest Ophthalmol Vis Sci. 1997; 38*: 207–218*

[7]Regan D. Low contrast letter charts and sine wave grating tests in ophthalmological and neurological disorders. Clinical Vision Sciences 1988; 2:235-250.

[8]Westheimer G. Measuring visual form discrimination with blur thresholds. Journal of Vision 2013; 13(5): 1-11.

[9]Westheimer G. Can perceptual learning compensate for image blur**?** *Vision Res* 2014, **94**:58-61.

[10]Bowley AL The standard deviation of the correlation coefficient. J Am Statist Assoc 1928; 23:31-34

[11]Frost NA, Brooks S, Myerscough J. Estimation of the effects of rounding visual acuity data. Br J Ophthalmol 2014; 98: 1385-1388.

[12]McMonnies CW Chart memory and visual acuity measurement**.** Clin Exp Optom. 2001 ,84:26-33

[13]Hazel CA**,** Elliott DB The dependency of logMAR visual acuity measurements on chart design and scoring rule Optom Vis Sci**.** 2002, 79:788-92.

[14] Guillery H. (1894). Einiges über den Formensinn. Archiv für Augenheilk. 1894; 28: 263-276

[15]Alexander KR, Xie W, Derlacki DJ, Szlyk JP. Effect of spatial sampling on grating resolution and letter identification. J Opt Soc Am A 1995; 12:1825-33**.**

[16]Carkeet A, Gerasimou DF, Parsonson LR, Biffin KL, Fredericksen BJ. Thresholds for sampled sloan letters are smaller than sample spacing. Optom Vis Sci. 2008; 85:1142-50.

[17]Wiecek E, Dakin SC, Bex P. Metamorphopsia and letter recognition, J of Vision 2014;14(14)1,1-10

[18]Gold J, Bennett PJ, Sekular AB. Identification of band-pass filtered letters and faces by human and ideal observers. Vision Res 1999; 39: 3537-3560

[19]Braje WL, Tjan BS, Legge GE. Human efficiency for recognizing and detecting low-pass filtered objects. Vision Res 1995, 35:2955-2966.

.

.

Table I

Average deviation from the ensemble mean of the 13 letter-recognition thresholds under the five degradation regimens

Kind of Degradation	Size	Contrast	Perturb.	Noise	Blur
Average Deviation	7.8%	10.1%	14.1%	10.3%	13.1%

TABLE II

Coefficients of correlation between letter recognition thresholds with various pairings of degradation regimens, all letters all observers. The standard error of the individual values is about *0.1*.

	Contrast	Perturb.	Noise	Blur
Size	0.14	0.38	0.02	0.32
Contrast	---	0.24	0.26	0.11
Perturb.	---	----	0.59	0.41
Noise	---	---	----	0.02

TABLE III

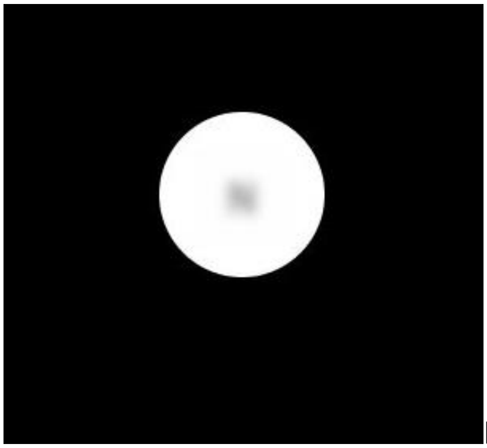
Analysis of the distribution of the errors in the confusion matrices within the responses in the five degradation experiments depicted in Fig. 4. Pairwise comparisons carried out in two ways: Average difference between the numbers in corresponding error cells, and correlations coefficients between the confusion matrices. The less commonality between any two distributions the smaller the correlation and the higher the average difference. Standard errors of the shown numbers is ~7% of their value or less.

Average difference in errors

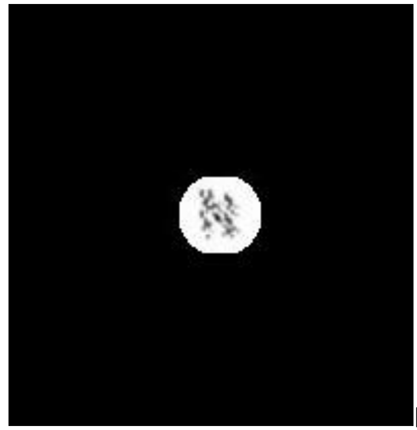
	contrast	perturb	noise	blur
size	3.28	4.06	4.74	4.44
contrast	---	3.52	3.62	4.84
perturb	----	---	2.25	4.41
noise	----	---	---	5.15

Correlation coefficient of errors

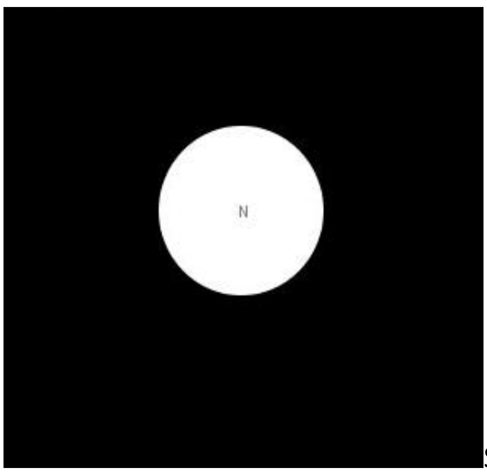
	contrast	perturb	noise	blur
size	0.62	0.63	0.42	0.68
contrast	----	0.71	0.65	0.55
perturb	---	----	0.84	0.66
noise	----	---	----	0.43



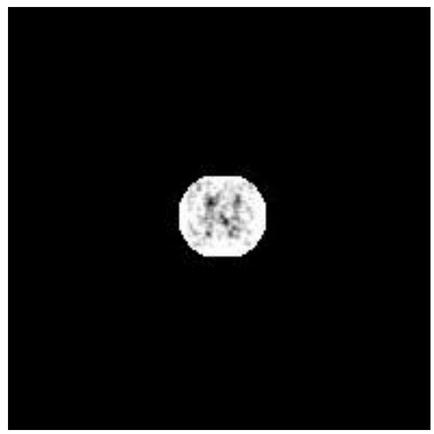
Blur



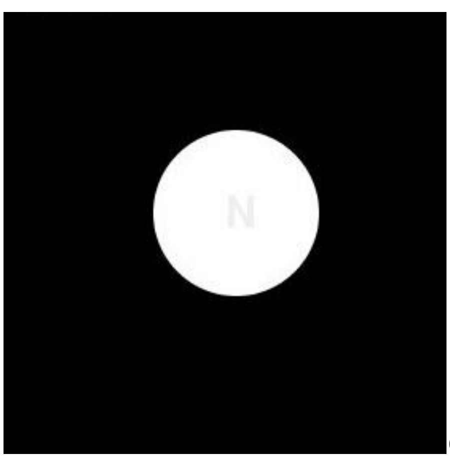
Perturbat.



Size



Noise



Contrast

Fig. 2. Sample screen views of the letter N seen under various forms of degradation:

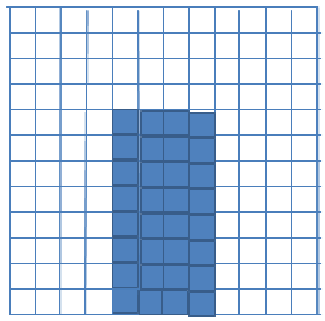
Blur increase: each pixel convolved with a Gaussian spread

Size reduction: overall letter size reduction

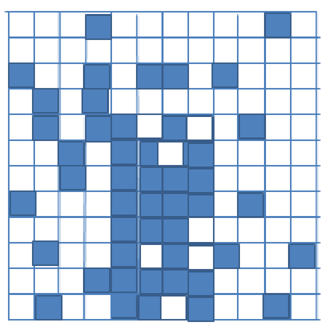
Contrast reduction: background luminance unchanged, letter luminance increased

Contour perturbation: rows and columns of letter matrix randomly jittered

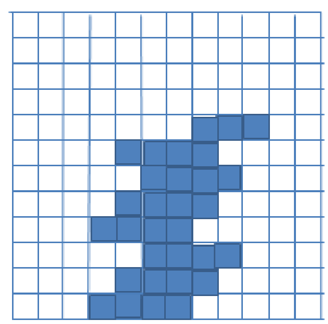
Noise Corruption: pixels randomly inverted in contrast polarity



A



B



C

Fig 2. Detail of a 12x12 pixel patch, containing the white background and black pixels making up the termination of the 4-pixel wide limb of the letter N, in three forms.

A. Normal view in an un-degraded letter

B. Appearance under noise corruption, when each pixel is individually and separately subjected to contrast inversion (black becoming white and white becoming black) with probability p $(0>p>0.5)$. That is, if, as in the example illustrated, p had been set at *0.25* for the given presentation, a random variable in the range *0* to *1.00* is generated for each pixel and if the random variable is <0.25, the particular pixel's contrast is inverted. Representative value of p for letter recognition threshold is *0.35.*

C. Appearance with contour perturbation, when rows and columns of the entire display matrix are displaced right or left, respectively up or down, by randomly, 0, 1, 2 or 3 pixels according to a Gaussian distribution with variance p. The value of p is the experimental parameter: the larger p is at recognition threshold, the more robust is the letter to perturbation of the smoothness of its contours.

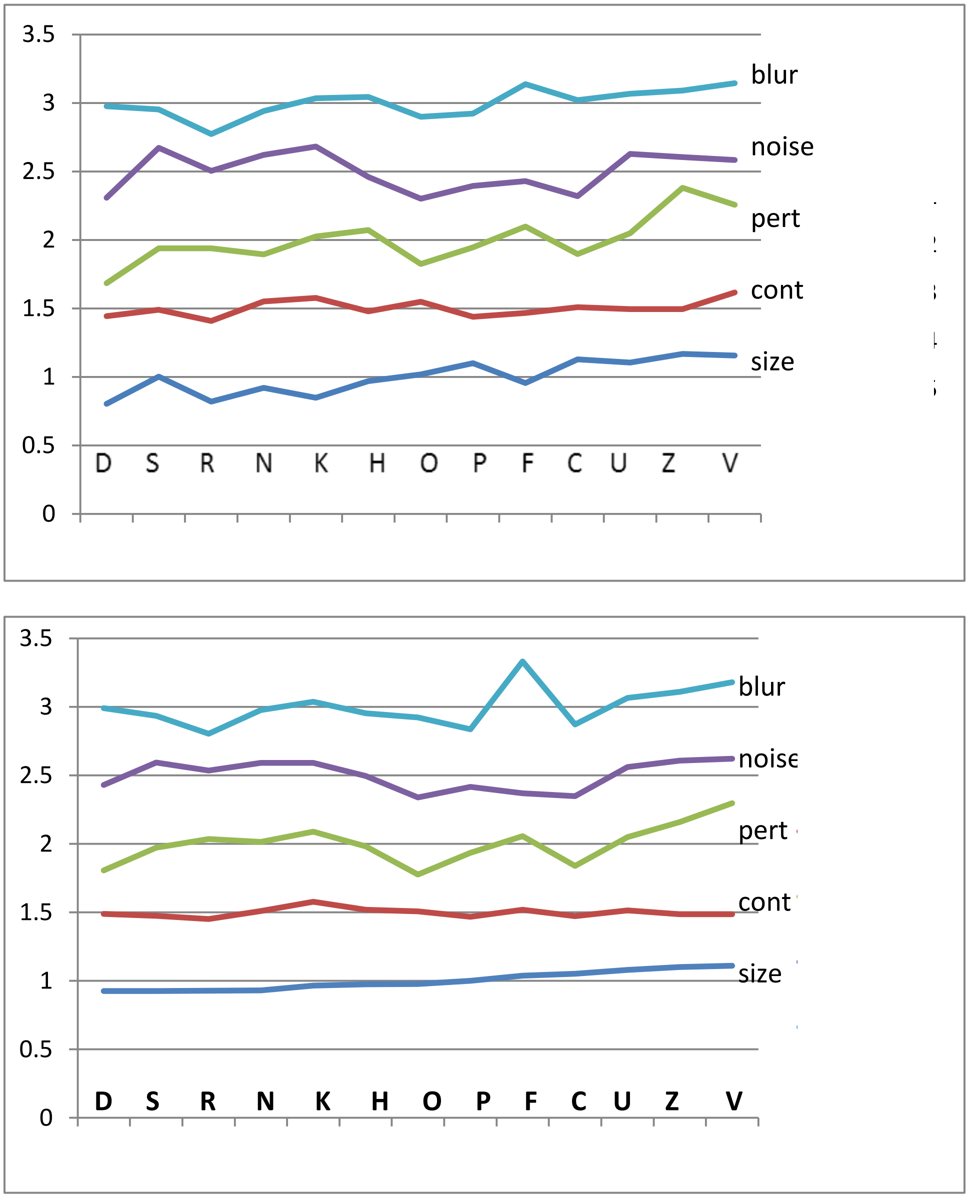
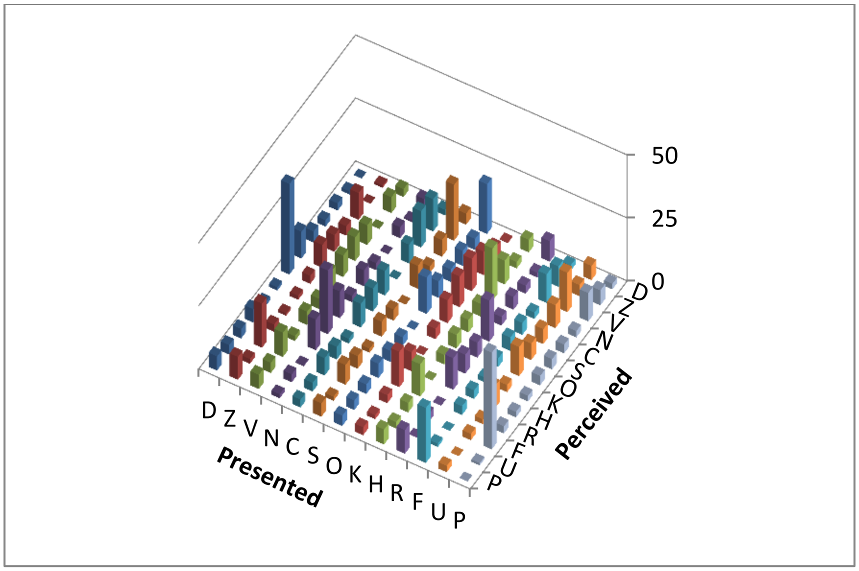
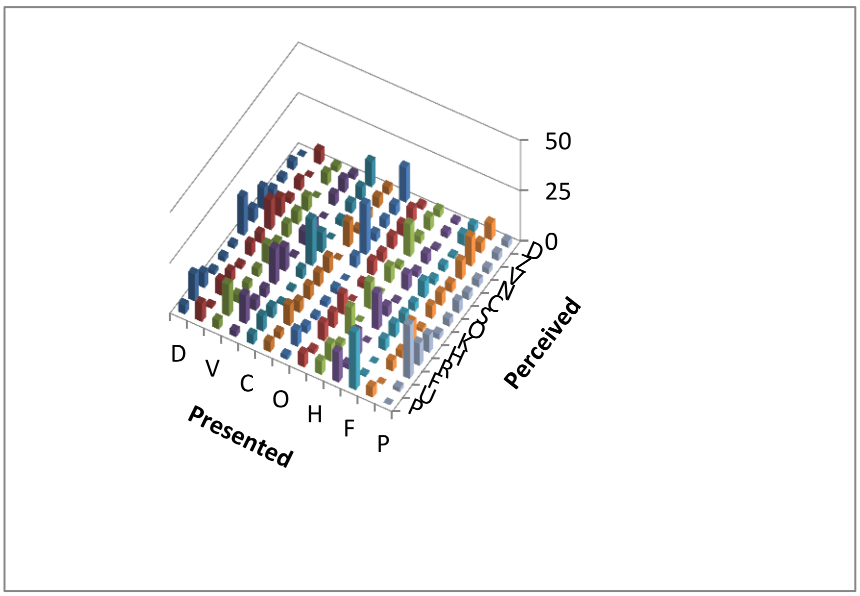


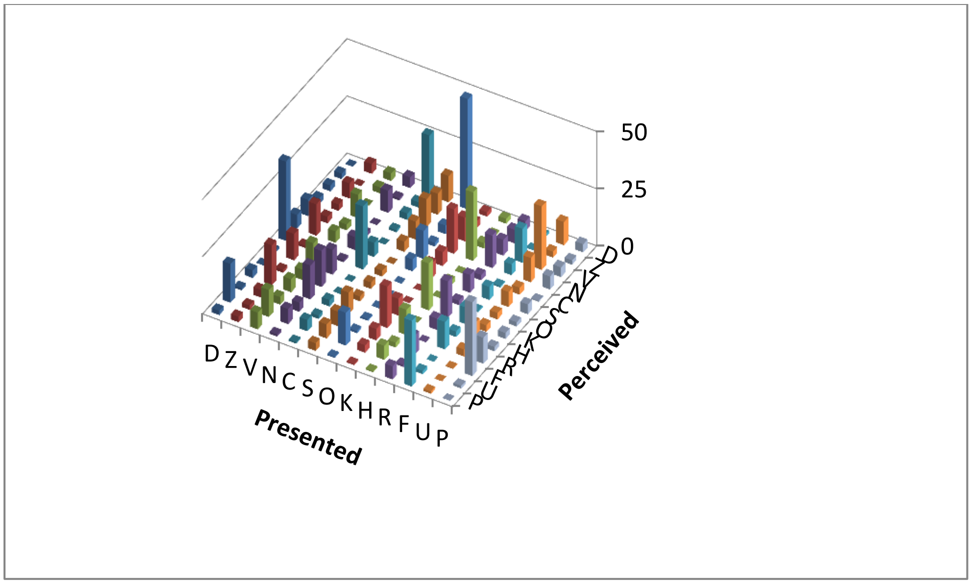
Fig. 3. Normalized recognition threshold for the 13 letters degraded respectively by size, contrast, contour perturbation, noise and blur respectively, from bottom up . For ease of inspection, curves are shifted up 0.5 unit from each other. Top panel: individual date from a typical observer, bottom, average of all five observers. Letters are arranged in ascending order of ease of recognition, as applied to the visual acuity sequence of the average. Standard error of data points is about 0.1.



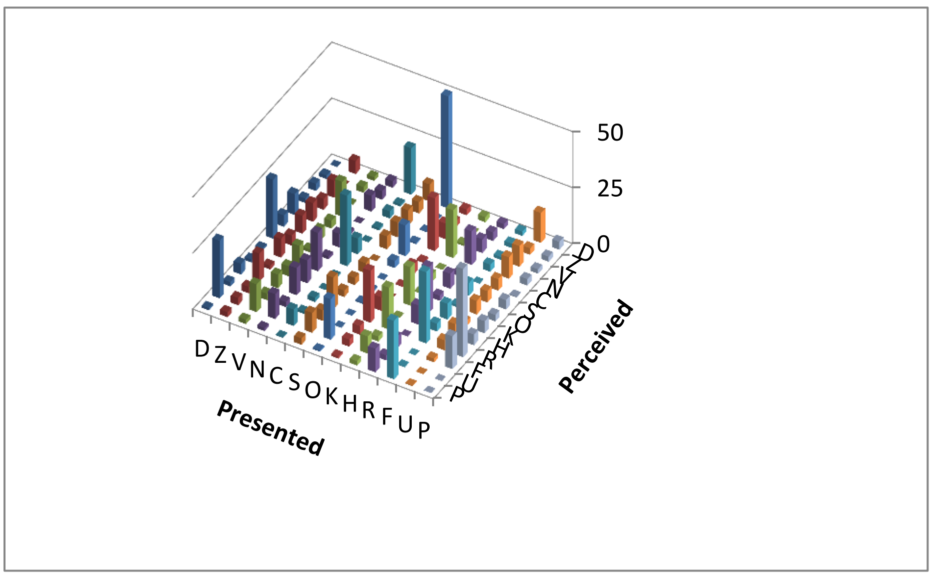
Size (Visual Acuity)



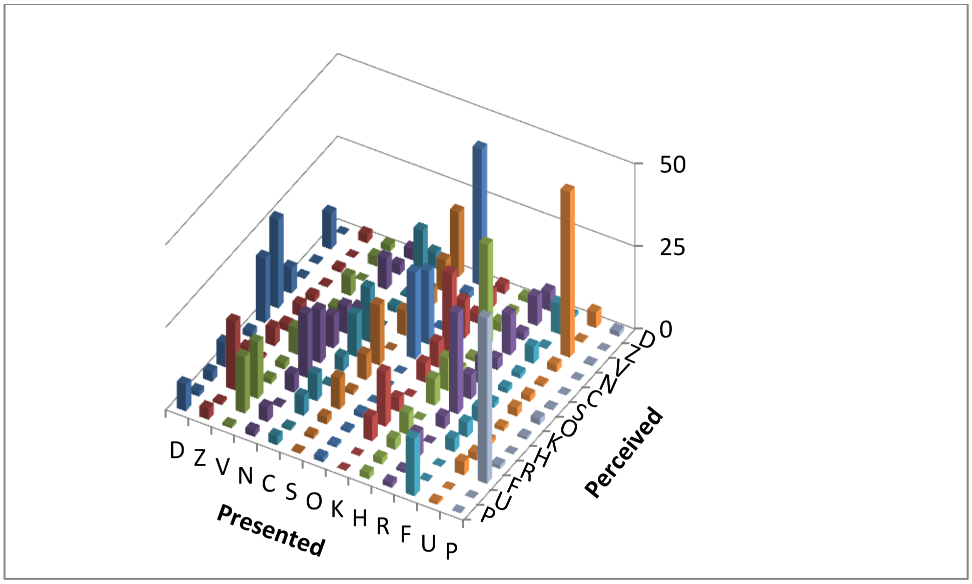
Contrast



Perturbation



Noise



Blur

Fig.4.

Confusion Matrices of the 13 letters. Height of bars represent number of specific erroneous letter responses to the presentation of the 13 letters in the ensemble. Normalized data, summed all observers, in the experiments involving degradation of the letters in the domain of, from top to bottom, size, contrast, contour perturbation, noise and blur. Entries along the diagonal, representing correct responses, have been suppressed in the interest of display clarity. The values are in the range of 400-600.

Details of the differences in error distribution between the five conditions could serve in the modelling of shape discrimination mechanisms.