Classification images predict absolute efficiency

Richard F. Murray Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

Patrick J. Bennett Department of Psychology, McMaster University, Hamilton, Ontario, Canada

Allison B. Sekuler Department of Psychology, McMaster University, Hamilton, Ontario, Canada

How well do classification images characterize human observers' strategies in perceptual tasks? We show mathematically that from the classification image of a noisy linear observer, it is possible to recover the observer's absolute efficiency. If we could similarly predict human observers' performance from their classification images, this would suggest that the linear model that underlies use of the classification image method is adequate over the small range of stimuli typically encountered in a classification image experiment, and that a classification image captures most important aspects of human observers' performance over this range. In a contrast discrimination task and in a shape discrimination task, we found that observers' absolute efficiencies were generally well predicted by their classification images, although consistently slightly (~13%) higher than predicted. We consider whether a number of plausible nonlinearities can account for the slight under prediction, and of these we find that only a form of phase uncertainty can account for the discrepancy.

Keywords: classification image, reverse correlation, linear observer model, detection, discrimination

Introduction

Classification images are an increasingly common way of characterizing human observers' strategies in visual discrimination tasks (e.g., Eckstein & Ahumada, 2002). A classification image shows the influence of small trial-to-trial variations in the stimulus on an observer's responses, and thereby shows what features the observer uses to detect or identify the stimulus. The usual method of measuring a classification image is to have an observer discriminate between two stimuli, *A* and *B*, in white Gaussian noise, and to calculate the classification image as follows:

$$C = (\overline{N_{AA}} + \overline{N_{BA}}) - (\overline{N_{AB}} + \overline{N_{BB}}). \quad (1)$$

Here $\overline{N_{SR}}$ denotes the average of the Gaussian noise fields over a stimulus-response class of trials (e.g., $\overline{N_{AB}}$ is the average of the noise fields over all trials where the stimulus is *A*, and the observer responds "*B*"). The classification image is a weighted sum of noise fields over all trials where the observer responds "A," minus a weighted sum of noise fields over all trials where the observer responds "*B*," so it shows which first-order features in the noise bias the observer toward one or the other response (Ahumada, 1996; Ahumada & Lovell, 1971; Gold, Murray, Bennett, & Sekuler, 2000; Sekuler, Gaspar, Gold, & Bennett, 2004).

The observer model used most often in connection with classification images is the linear observer model or one of its variants (Burgess, Wagner, Jennings, & Barlow, 1981). According to the linear observer model, an observer discriminates between two signals, *A* and *B*, possibly in an external noise field *N*, by adding an internal noise *Z*, cross-correlating the corrupted stimulus with a template *T*, and responding "A" if the resulting decision variable exceeds a criterion *a*, and responding "*B*" otherwise. Introducing the symbol $I_{\{A,B\}}$ for the signal that is randomly chosen as *A* or *B*, using $\otimes$ to denote cross-correlation (i.e., $X \otimes Y = \sum_{i,j} X_{ij} Y_{ij}$), and coding the observer's responses as $R = \pm 1$, we can describe the linear observer model as follows:

$$s = (I_{\{A,B\}} + N + Z) \otimes T \quad (2)$$

$$R = \mathrm{sgn}(s - a) = \begin{cases} +1 & \text{if } s \geq a \\ -1 & \text{otherwise} \end{cases} \quad (3)$$

For a linear observer, the expected value of the classification image is proportional to the template *T* (Ahumada, 2002; Murray, Bennett, & Sekuler, 2002; Richards & Zhu, 1994).

Although the linear observer model is a useful first-order approximation that captures important properties of human performance in many tasks, it is certainly not a complete model of performance. It does not incorporate spatial uncertainty, static nonlinearities, perceptual learning, or a range of other phenomena that affect performance. This leads to the question, "How well do classification images actually characterize human observers' strategies?"

doi:10.1167/5.2.5 Received July 26, 2004; published February 24, 2005 ISSN 1534-7362

A classification image recovers a linear observer's template, so we might expect that by examining a classification image we could form an idea as to how well a linear observer performs a discrimination task. This is true, as we show in the Appendix. In a two-alternative identification task in Gaussian noise, an observer's *absolute efficiency* F is equal to the squared ratio of the observer's performance d' and the ideal observer's performance d'_I on the same task (Tanner & Birdsall, 1958):

$$F = (d'/d'_I)^2 \ . \tag{4}$$

In the Appendix, we show that the following statistic, based on the cross-correlation of the classification image C with the ideal observer's template T_I, recovers an unbiased linear observer's absolute efficiency in a two-alternative identification task:

$$\hat{F} = \left(\frac{(C \otimes T_I)^2}{\sigma_C^2} - 1 \right) \frac{G(d'/2)G(-d'/2)}{ng(d'/2)^2} \ . \tag{5}$$

Here C is the classification image calculated as in Equation 1, T_I is the ideal observer's template normalized to unit sum-of-squares energy, σ^2_C is the pixelwise variance of the classification image, d' is the observer's performance level, n is the number of trials used to calculate the classification image, g is the standard normal probability density function, and G is the standard normal cumulative density function. The statistic $\hat{F}$ is slightly biased if there are fewer than 500 trials in the experiment, but normally we need far more trials than this anyway to measure the classification image accurately. (See Appendix for details.)

In the Appendix we give a full proof that Equation 5 recovers a linear observer's absolute efficiency, but here we briefly describe why the equation works. The key term is $(C \otimes T_I)^2$, the squared cross-correlation of the classification image with the ideal template. Two factors are at play in this term: the similarity of the observer's template to the ideal template – sometimes called sampling efficiency – and the observer's internal noise power. Because the classification image estimates the observer's template, this cross-correlation will be higher when the observer's template is similar to the ideal template (i.e., when the observer has higher sampling efficiency). Furthermore, because a classification image has a greater signal-to-noise ratio (SNR) for an observer with low internal noise (see discussion of Equation 16 in the Appendix), the cross-correlation will be higher for an observer with low internal noise. Thus sampling efficiency and internal noise power both affect the cross-correlation. They also affect the observer's absolute efficiency. The fortunate coincidence that makes this method of analyzing classification images possible is that they affect the cross-correlation and the observer's absolute efficiency in a quantitatively identical way, and consequently we can use the cross-correlation to recover absolute efficiency. The remaining terms in Equation 5 merely correct for the influence of the observer's performance level, the number of trials, and the stimulus noise power on the cross-correlation of the classification image with the ideal template.

To investigate whether classification images give a good description of how human observers perform visual discrimination tasks, at least over the small range of stimuli shown in a classification image experiment, we tested whether human observers' absolute efficiencies in various tasks were consistent with the efficiencies indicated by their classification images. That is, we used Equation 5 to predict human observers' absolute efficiencies from their classification images, and compared these predictions with directly measured absolute efficiencies, calculated as in Equation 4. We call this the *ideal correlation* method of testing the adequacy of a classification image.

Note that d' appears in Equation 5, so it is not the case that from just the observer's classification image and knowing nothing about performance we can use this equation to recover the observer's absolute efficiency. Indeed, if we were interested only in finding absolute efficiency, we could use d' directly in Equation 4. This is not a crucial shortcoming, however, because our goal is not to deduce an observer's efficiency from a classification image of unknown origin, but rather to see whether the strategy revealed in a classification image is consistent with the observer's performance. It so happens that the quality (i.e., SNR) of a classification image depends on the observer's performance level in a particular way, and d' enters into Equation 5 only to correct for this effect of performance level, as mentioned above and made clearer in the Appendix. Equation 5 is not in some covert way equivalent to Equation 4, and the predictions of Equation 5 can fail for observers whose strategies are not well described by classification images, as we discuss in detail below.

Methods

Experiment A: Dot contrast discrimination

We used data from Experiments 2 and 3 of Murray, Bennett, and Sekuler (2002). The signal was a contrast increment in one of two dots shown in Gaussian white noise, and observers judged whether the contrast increment was in the dot to the left or right of fixation (Figure 1). The radius of each dot was 0.11 degrees of visual angle (deg), and the center of each dot was 0.50 deg to the left or right of a small fixation point. The base contrast of each dot was 10% Weber contrast (defined as $c = (L - L_{bg})/L_{bg}$, where L is the luminance of the point of interest and L_{bg} is background luminance), and the contrast increment was set to each observer's 70% contrast threshold (which ranged between 3% and 7% contrast) as measured in a pilot session. The pixelwise root-mean-square contrast of the noise was 20%, and the pixel size was 0.027 deg, for a noise power of 29 mdeg^2. With these stimulus parameters, the base-contrast dot was well above detection threshold, so observ-

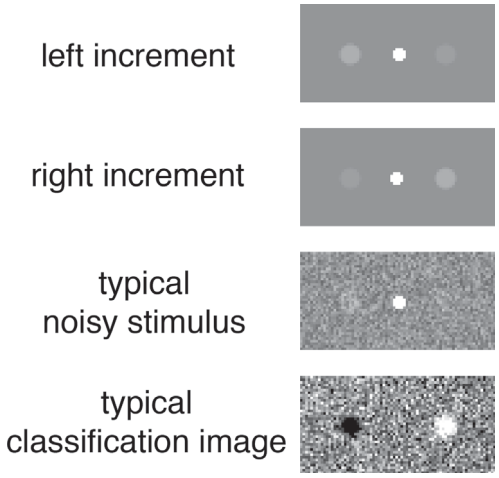

Figure 1. Stimuli in Experiment A and a typical classification image. The noiseless stimuli are shown here only to illustrate the signals being discriminated, and were not shown in the experiment. The bright central dot in the stimuli is the fixation point, and the contrast increment occurred in the fainter dots to the left or right.

ers were confident about the spatial locations of the two dots. The stimulus duration was 200 ms. There were three observers in Experiment 2 and six observers in Experiment 3. Each observer ran in approximately 10,000 trials. Further details can be found in Murray et al. (2002).

We calculated classification images using Equation 1, and predicted the observers' absolute efficiency using Equation 5.

Experiment 2 of Murray et al. (2002) was a two-pass experiment, in which every second 100-trial block was an exact repetition of the previous 100-trial block (Burgess & Colborne, 1988). Equation 5 assumes that the external noise fields shown on different trials are independent, so we analyzed the first- and second-pass trials separately. This gave us two classification images per observer, each based on approximately 5,000 trials.

Experiment 3 of Murray et al. (2002) was a rating-scale experiment, in which observers rated their confidence that the contrast increment was in the left or right dot, on a scale of 1 to 6. We collapsed this rating scale, and considered responses 1 to 3 to indicate a "left" response, and responses 4 to 6 to indicate a "right" response. In Experiment 2 (the two-pass experiment), observers responded "left" on between 47% and 58% of trials, whereas in this experiment observers responded "left" (i.e., gave ratings 1-3) on between 41% and 60% of trials, so using collapsed responses resulted in slightly larger response biases. However, response biases this small have a negligible effect on the accuracy of predictions made using Equation 5 (Murray, 2002).

Experiment B: Kanizsa square shape discrimination

We used data from Section 2.3 of Murray (2002). The stimuli were real, illusory, occluded, and fragmented Kanizsa squares shown in Gaussian white noise (Figure 2). The four stimulus types were shown in separate blocks. Observers judged whether the stimulus was "fat" or "thin." The radius of the Kanizsa inducers was 0.5 deg, and the

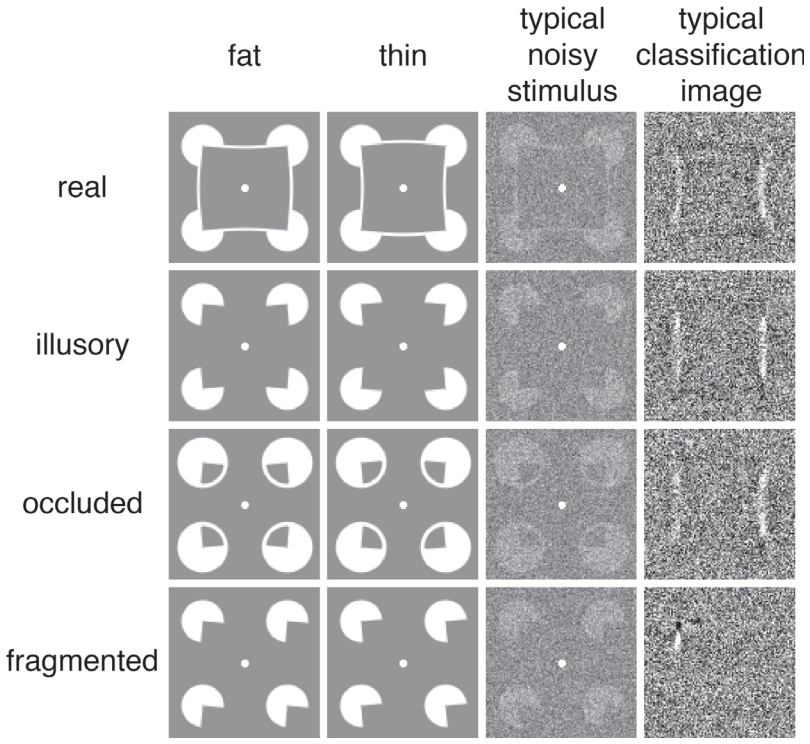

Figure 2. Stimuli in Experiment B and typical classification images.

inducers were spaced 2.0 deg apart, center-to-center. The angle of the inducers from horizontal-vertical was held constant from trial to trial, and ranged from 5° to 8° for different observers. Stimulus contrast was set at each observer's 70% threshold (which ranged between 8% and 30% contrast) as measured in a pilot session. The pixelwise root-mean-square contrast of the noise was 18%, and the pixel size was 0.035 deg, for a noise power of 40 mdeg^2. The stimulus duration was 240 ms. There were five observers, each of whom ran in between 2,800 and 9,000 trials with each of the four types of stimuli. Further details can be found in Murray (2002).

We calculated classification images using Equation 1, and predicted the observers' absolute efficiency using Equation 5. We discarded three data points based on relatively small numbers of trials (less than 4,500 trials each), for which the *SE* of predicted absolute efficiency was greater than 40% of the prediction itself.

Results

Figures 3 and 4 show predicted and actual absolute efficiency in each condition of the two experiments. Actual efficiency is plotted against predicted efficiency, so perfect predictions would lie on the diagonal. These figures show that observers have approximately the absolute efficiency that one would expect from comparing their classification images to the ideal template, but that performance is consistently slightly higher than predicted. This holds true over absolute efficiencies ranging from less than 2% to more than 35%.

Actual efficiency exceeds predicted efficiency 12 times out of 12 in Experiment A, and 14 times out of 16 in Experiment B. In both cases a sign test easily rejects the hypothesis that the predictions are exact ($p < .01$).

Nevertheless, in Experiment A, the maximum-likelihood weighted average of the ratios of actual to predicted efficiency is only 1.12, and in Experiment B it is only 1.14. Because absolute efficiency is the *squared* ratio of real and ideal d', these values correspond to only a 6% to 7% under prediction of d'. Thus, although the predictions are not exact, they are not far off either. (Figures 3 and 4 show that some data points have ratios much larger than 1.14, but these points typically also have large *SE*s, so they have little effect on a maximum-likelihood average in which values are weighted by the inverse of their squared *SE* [Taylor, 1982].)

In Experiment A, maximum-likelihood linear regression of actual against predicted efficiency gives a slope of 1.01 ± 0.20 and a *y*-intercept of 0.004 ± 0.004. (The error values are *SE*s, calculated by bootstrapping.) The *SE*s of these values are too large for us to conclude whether actual efficiency differs from predicted efficiency by a fixed offset or by a scale factor. In Experiment B, the slope and *y*-intercept parameters are 0.99 ± 0.10 and 0.005 ± 0.001, respectively, indicating that the difference is mostly due to a fixed offset.

In summary, in both experiments the predictions of Equation 5 are generally accurate, but they under predict absolute efficiency by about 13%. Linear regression does not indicate that that the difference is consistently due to either a fixed offset or a scale factor.

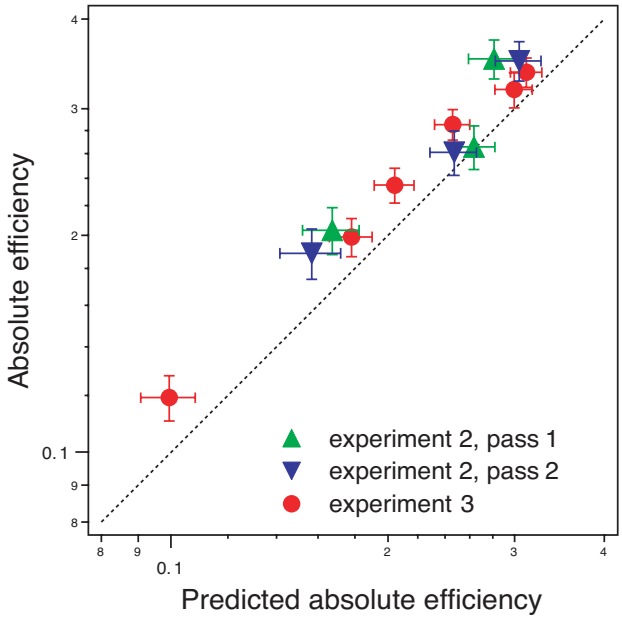

Figure 3. Predicted and actual absolute efficiency in Experiment A (dot contrast discrimination). Performance was similar on the two passes of Experiment 2 of Murray et al. (2002), and the pairs of upward and downward triangles that appear close together in the plot correspond to the two passes made by each observer. Error bars represent *SE*s.

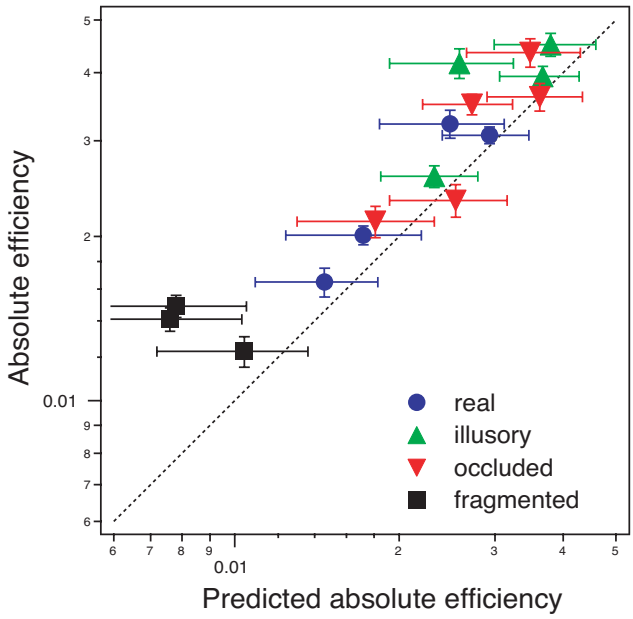

Figure 4. Predicted and actual absolute efficiency in Experiment B (Kanizsa square shape discrimination). Error bars represent *SE*s.

Discussion

The ideal correlation method predicts absolute efficiency quite well in our contrast discrimination and shape discrimination tasks, but this will not be the case in all tasks. Ahumada and Beard (1999) measured classification images in a task where the observer detected a grating, but was uncertain about the phase of the grating. They found that the part of the classification image measured on signal-absent trials was blank, aside from the inevitable sampling noise. They noted that empty classification images are expected on signal-absent trials for a phase-uncertain observer, because the observer will report a signal present when a grating-like pattern of *any* phase appears in the noise, and over a large number of trials the contribution to the classification image of any grating-like noise pattern will be cancelled by the contribution of the corresponding phase-reversed grating-like noise pattern. This phase-uncertain strategy is a feasible (though suboptimal) strategy for detecting gratings, but by examining the portion of such an observer's classification image collected on signal-absent trials, one would mistakenly infer that the observer's performance was near chance. In this case, Equation 5 would fail to predict absolute efficiency, and this failure could provide a warning that the observer's strategy was poorly described by a classification image.

Thus, phase uncertainty might account for the small discrepancy we found between predicted and actual efficiency. What other departures from the linear observer model could also account for this discrepancy? As a preliminary exploration of this question, we simulated a number of model observers who departed from the linear model in various ways, and we tested whether these simulated observers also performed better than one would expect from their classification images. In the long run, a theoretically grounded understanding of possible decision strategies is preferable to model observer simulations, but simulations can provide clues as to which avenues of research are worth following, and we have found simulations to be helpful in understanding how nonlinearities can affect observers' classification images. We simulated model observers who were uncertain about various aspects of the stimulus (e.g., spatial position or phase), who viewed the stimulus after transformation by a pointwise nonlinearity (e.g., a logarithmic transform of luminance), who exhibited response biases (i.e., were more likely to respond "*A*" or "*B*"), and who simply guessed at the correct response on some proportion of trials. We will discuss in detail only our results with model observers who were uncertain about spatial position or phase, as these results are representative of the other cases. The remaining simulations are discussed in Murray (2002).

All simulations used the same stimuli. Model observers judged whether a single Kanizsa inducer, like the top-left inducer in the fat and thin illusory Kanizsa squares in Figure 2, was rotated clockwise or counterclockwise of horizontal-vertical. The angle of rotation was ±5°, the inducer radius was 14 pixels, and the whole stimulus measured 50 × 50 pixels. The inducers were shown in Gaussian white noise with a pixelwise root-mean-square contrast of 20%. The contrast of the inducer was chosen separately for each model observer, to yield 70% correct performance.

Spatial uncertainty

To examine the effect of spatial uncertainty, we simulated model observers who had spatial uncertainty of 0, 1, 2, 4, or 8 pixels. An observer who had an uncertainty of p pixels cross-correlated the ideal template with the stimulus, centered on each pixel within a distance of p pixels of the correct location. The decision variable was the maximum of these cross-correlations: If the maximum exceeded a criterion, the observer gave one response, and otherwise gave the other response (Green & Birdsall, 1978). We chose the criterion to produce unbiased responses. We gave the model observers an early internal noise source that resulted in an internal-to-external noise ratio of 1.0, which is a typical value for human observers (Gold, Bennett, & Sekuler, 1999; Green, 1964). We simulated each model observer in 100 repetitions of an experiment with 10,000 trials. In each repetition of the experiment, we calculated predicted and actual efficiency, and we averaged these values over the 100 repetitions.

Figure 5 shows predicted and actual absolute efficiency (the black circles), as well as typical classification images. Predictions were accurate at low levels of uncertainty, and were too high at high levels of uncertainty, indicating that this form of spatial uncertainty cannot explain the under prediction in our experiments. (The pedestals in Experiment A minimized spatial uncertainty [Pelli, 1985], so it is not surprising that spatial uncertainty alone cannot account for the discrepancy in that experiment.) The ratio of actual to predicted performance at spatial uncertainty levels of 0, 1, 2, 4, and 8 pixels was 1.00, 1.00, 1.00, 0.84, and 0.56, respectively.

In simulations of model observers with early pointwise nonlinearities and with some other types of stimulus uncertainty, we obtained results similar to those reported for low levels of spatial uncertainty: Departures from the linear model worsened performance, but also reduced the cross-correlation of the classification image with the ideal template, with the result that absolute efficiency was accurately predicted from classification images (Murray, 2002).

Phase uncertainty

As discussed earlier, we expect some forms of phase uncertainty to lead to discrepancies between predicted and actual absolute efficiency, so we simulated a set of model observers that incorporated phase uncertainty. The stimuli were the same as in the simulations of spatially uncertain observers. Observers' decision variables were calculated from the cross-correlation of the stimulus with a pair of

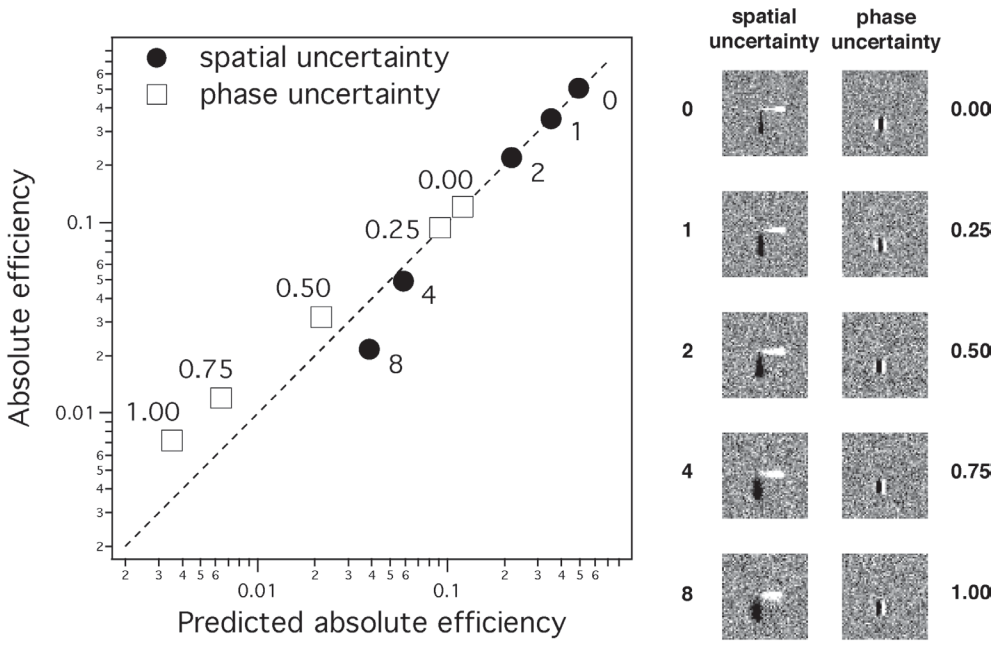

Figure 5. Predicted and actual absolute efficiency for simulated model observers with spatial uncertainty and phase uncertainty. Spatially uncertain observers had spatial uncertainty of 0, 1, 2, 4, and 8 pixels. Phase-uncertain observers had phase uncertainty parameters 0.00, 0.25, 0.50, 0.75, and 1.00. Error bars representing *SE*s are smaller than the data points. Typical classification images are shown at right.

vertical sine- and cosine-phase Gabor templates, located on the vertical edge of the Kanizsa inducer. The model observers calculated decision variables from the outputs of the Gabor templates, as follows:

$$s = (1-\lambda)(S \otimes G_{\cos}) + \lambda((S \otimes G_{\cos})^2 + (S \otimes G_{\sin})^2) \quad . \quad (6)$$

Here S is the stimulus with external and internal noise added (i.e., $S = I_{\{A,B\}} + N + Z$), $G_{\cos}$ is the cross-correlation with the cosine-phase Gabor template, and $G_{\sin}$ is the cross-correlation with the sine-phase Gabor template. This is the ideal decision variable for detecting a Gabor when the observer knows that on average the Gabor is in cosine-phase, but the observer is uncertain about the exact phase on any given trial (Van Trees, 1968, pp. 335-341). The parameter λ determines the degree of phase uncertainty. When $\lambda = 0$, the observer has no phase uncertainty, and the decision variable reduces to cross-correlation with a cosine-phase Gabor. When $\lambda = 1$, there is complete phase uncertainty, and the decision variable reduces to energy detection (i.e., the summed, squared outputs of Gabor filters that are an approximate quadrature pair). When $0 < \lambda < 1$, there is partial phase uncertainty. Our model observers had uncertainty parameters of 0.00, 0.25, 0.50, 0.75, and 1.00. The model observers also had an early internal noise source that resulted in an internal-to-external noise ratio of 1.0.

We simulated each model observer in 100 repetitions of an experiment with 10,000 trials. In each repetition of the experiment, we calculated predicted and actual efficiency, and we averaged these values over the 100 repetitions. As expected, absolute efficiency was much higher than predicted when phase uncertainty was large (Figure 5, white squares). The ratio of actual to predicted efficiency for phase uncertainty values of 0.00, 0.25, 0.50 0.75, and 1.00 was 1.00, 1.03, 1.48, 1.87, and 2.03, respectively. For our human observers, absolute efficiency was about 13% higher than predicted, so even a small amount of intrinsic phase uncertainty could account for this discrepancy.

One caveat is that it is unclear under what conditions phase uncertainty results in under predictions of efficiency. It is unlikely that human observers base their decisions on the outputs of a single pair of Gabor filters, so in further simulations, we examined model observers who based their decisions on the output of a large bank of Gabor filters distributed across the stimulus, with phase uncertainty in each filter (Murray, 2002). The results were similar to those we reported for small amounts of spatial uncertainty: Absolute efficiency and cross-correlation of the classification image with the ideal template declined in tandem, and classification images accurately predicted absolute efficiency. (The simulations in Murray, 2002, did not adjust signal contrast across model observers so as to hold performance constant. In further simulations with performance held constant, we found that for model observers with large amounts of phase uncertainty, performance was slightly higher than predicted, but even with complete phase uncertainty the discrepancy was less than the 13% difference found in the present experiments.) Thus not just any kind of phase uncertainty can explain the under prediction of absolute efficiency that we found with human observers.

In any case, our results suggest that some form of phase uncertainty may have had a significant effect on human observers' performance, because of all the nonlinearities we have examined, only phase uncertainty can account for the discrepancy between predicted and actual absolute efficiency. Furthermore, there is independent evidence that under some conditions, observers are poor at discriminating between gratings of different phases, consistent with the hypothesis of intrinsic phase uncertainty (Bennett, 1993).

Limitations of the current analysis

The ideal correlation method does not give a final and definitive test of the linear observer model as a model of human performance. First of all, it only tests whether human performance is consistent with the classification image over the narrow range of stimuli encountered in the classification image experiment (i.e., two signals at a fixed contrast, perturbed by Gaussian noise). Outside this range, the linear model will often fail (e.g., the linear model predicts that psychometric functions are linear when plotted as d' versus signal contrast [see Equation 10 in the Appendix], but it is well known that psychometric functions are often nonlinear [Legge, Kersten, & Burgess, 1987; Pelli, 1985]). Second, other tests may reveal failings of the linear model that are not detected by the ideal correlation method, even within the same narrow range of stimuli (e.g., early pointwise nonlinearities are not detected by the ideal correlation method, as discussed earlier, but they can be detected by histogram contrast analysis) (Chubb, Econopouly, & Landy, 1994; Murray et al., 2001). Our goal in introducing the ideal correlation method is not to see whether there is a need for more complex models than the linear model as a general model of human vision, for there is ample evidence that more complex models are needed. Rather, our goal is to provide one test of whether classification images provide a reasonable phenomenological description of observers' strategies, even over a limited range of stimuli.

In this regard, our simulations of observers with spatial uncertainty are instructive. The model observers were substantially different from the linear model, in ways that had large effects on their performance. Nevertheless, in some cases we were able to predict observers' performance from their classification images. For instance, the model observers with zero and two pixels of spatial uncertainty had absolute efficiencies of 0.50 and 0.20, respectively – thus two pixels of spatial uncertainty had a large effect on performance – and yet even for the observer with two pixels of uncertainty, the predictions were accurate. Obviously, then, just because these predictions succeed, we cannot conclude that an observer is linear. Rather, if the predictions succeed, this only suggests that over the narrow range of stimuli encountered in the experiment, the observer's strategy is described reasonably well by the linear model. For our model observers, spatial uncertainty is reflected in the classification image as a smearing of the underlying template. In general, smearing of a linear template is an inadequate model of spatial uncertainty, but our simulations suggest that over a small stimulus range, such smearing gives an adequate description of how different parts of the stimulus affect an observer's responses. Again, this does not rule out the possibility that other tests could detect failings of the linear model, even over the same limited range of stimuli.

An analogy is the linear Taylor series approximation to an arbitrary function: $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$. If this approximation is accurate over an interval surrounding x_0, we do not conclude that f is a linear function. We conclude that although the full definition of f may involve nonlinear terms, over the interval of interest, the behavior of f is adequately characterized by a linear approximation, and we will not be greatly misled about the values of f over that interval by examining the linear approximation. Just so, visual processes can often be approximated as being linear over a limited range (e.g., Ahumada, 1987), and this is the approximation that underlies use of classification images to characterize observers' strategies.

Relationship to other methods

Others have also considered the question of whether classification images give an adequate description of observers' strategies. Neri and Heeger (2002) and Neri (2004) proposed methods to characterize nonlinear aspects of observers' strategies; their methods and the ideal correlation method are complementary, in that their methods describe the nonlinear components of observers' strategies, and the ideal correlation method tests whether observers' performance is consistent with a model that disregards such nonlinearities. Barth, Beard, and Ahumada (1999) and Ahumada (2002) developed a test to determine whether a linear model incorporating an observer's classification image can account for an observer's performance level. We believe that the ideal correlation method represents an improvement on their test, which involves fitting a parametric approximation to the classification image: If their test fails, it is unclear whether the linear model itself is rejected, or whether the parametric approximation is inadequate. Levi and Klein (2002) and Li, Levi, and Klein (2004) showed that human performance covaries with the squared cross-correlation of the classification image and the ideal template; the ideal correlation method extends their method by taking account of the effects of parameters such as performance level and internal noise power, thereby permitting a quantitative prediction of human efficiency from a classification image.

Finally, one convenient property of the ideal correlation method of testing the adequacy of classification images is that we do not need a separate estimate of the observer's internal noise power. Not all methods will have this property. Consider, for example, the method of cross-validation, in which one would measure a classification image from, say, 80% of the trials in an experiment, and test its adequacy by trying to predict the observer's responses on the remaining 20% (Baddeley & Tripathy, 1998; Camstra & Boomsma, 1992; Mosier, 1951). What proportion of successfully predicted trials would validate the classification

image? Even without a quantitative model, we can see that we would have to know the observer's internal noise power to determine the criterion for accepting the classification image as adequate. The higher the observer's internal noise power, the lower the proportion of successfully predicted trials we would require: first, because the classification image would be noisier (i.e., it would have a lower SNR; see discussion of Equation 16 in the Appendix), and so would be less accurate at predicting subsequent responses, and second, because the observer's responses would be intrinsically more difficult to predict, as they would be more strongly influenced by internal noise. Testing the adequacy of classification images with the ideal correlation method circumvents this problem, because sampling efficiency and internal noise power affect both the ideal correlation and absolute efficiency in quantitatively the same manner.

Conclusions

All aspects of a linear observer's strategy are reflected in its classification image, and so from a classification image, we can predict a linear observer's absolute efficiency. Testing whether the same is true for human observers provides one test of whether classification images adequately characterize observers' strategies. We found that in contrast discrimination and shape discrimination tasks, human efficiency was generally well predicted by classification images, although efficiency was slightly (~13%) higher than predicted. We found that several plausible departures from the linear observer model cannot account for this discrepancy, and the only nonlinearity we found that can account for it is a form of phase uncertainty.

Appendix

Here we show that a linear observer's absolute efficiency can be predicted from the cross-correlation of its classification image with the ideal template.

First, we will derive an expression for a linear observer's absolute efficiency, defined as the ratio of the ideal observer's contrast energy threshold E_I and the observer's contrast energy threshold E on the same task (Tanner & Birdsall, 1958):

$$F = E_I / E \quad . \tag{7}$$

In a two-alternative identification task in Gaussian noise, the ideal observer's performance d'_I is proportional to the square root of the signal contrast energy (Peterson, Birdsall, & Fox, 1954), and consequently absolute efficiency can be calculated as the squared ratio of an observer's performance d' and the ideal observer's performance d'_I:

$$F = (d'/d'_I)^2 \quad . \tag{8}$$

A linear observer's two-alternative identification responses are based on a decision variable s that is the cross-correlation of a signal I corrupted by external noise N and internal noise Z, with a unit-energy template T, as described by Equations 2 and 3. (By unit-energy, we mean $\|T\|^2 \equiv \sum_i T_i^2 = 1$.) In a discrimination task with signals A and B, the linear observer's performance is given by the difference between the expected value of s on signal-A and signal-B trials, divided by the SD of s:

$$d' = \frac{(A-B)\otimes T}{(\sigma_N^2 + \sigma_Z^2)^{1/2}} \tag{9}$$

$$= \frac{\|A-B\|(T_I \otimes T)}{\sigma_N (1+\rho)^{1/2}} \tag{10}$$

Here $\rho = \sigma^2_Z/\sigma^2_N$ is the observer's internal-to-external noise ratio, and $T_I = \frac{(A-B)}{\|A-B\|}$ is the ideal observer's unit-energy template. When the linear observer is the ideal observer, the template is $T = T_I$ and the internal-to-external noise ratio is $\rho = 0$, so Equation 10 shows that the ideal observer's performance is

$$d'_I = \frac{\|A-B\|}{\sigma_N} \quad . \tag{11}$$

Substituting Equation 10 and 11 into Equation 8, we find that the linear observer's absolute efficiency is

$$F = \frac{(T \otimes T_I)^2}{1+\rho} \quad . \tag{12}$$

Second, we note that because the expected value of a linear observer's classification image is proportional to the observer's template T, we can express the classification image, considered as a random variable, as the sum of a scaled template kT and a field of sampling noise N_C with a pixelwise mean of zero:

$$C = kT + N_C \quad . \tag{13}$$

The scale factor k is necessary because the expected value of the classification image is merely proportional to the unit-energy template T, and not necessarily equal to it.

Third, we define the signal-to-noise ratio (SNR) of a classification image as the energy of its expected value kT, divided by the pixelwise variance of the sampling noise N_C:

$$SNR[C] = \|kT\|^2 / \sigma_C^2 \tag{14}$$

$$= k^2 / \sigma_C^2 \tag{15}$$

Murray et al. (2002; Equation 4) showed that the SNR of an unbiased linear observer's classification image is

$$SNR[C] = \frac{n}{(1+\rho)} \cdot \frac{g(d'/2)^2}{G(d'/2)G(-d'/2)} \quad . \tag{16}$$

Here n is the number of trials, ρ is the observer's internal-to-external noise ratio, d' is the observer's performance, g is the standard normal probability density function, and G is the standard normal cumulative distribution function.

Finally, consider the statistic $\hat{F}$ that we introduced in Equation 5:

$$\hat{F} = \left(\frac{(C \otimes T_I)^2}{\sigma_C^2} - 1 \right) \frac{G(d'/2)G(-d'/2)}{ng(d'/2)^2} \quad . \tag{17}$$

We would like to find the expected value of this statistic. The values of C, σ^2_C, and d' are all calculated from a single reverse correlation experiment, and in general they are correlated, so finding the expected value of this equation is not straightforward. However, after the large number of trials typically used in a reverse correlation experiment, σ^2_C and d' will be known quite precisely. We will make the simplifying assumption that σ^2_C and d' are known constants to derive the expected value of Equation 17, and then we will test the accuracy of the resulting expression in simulations.

We will substitute Equation 13 for the linear observer's classification image into Equation 17:

$$\hat{F} = \left(\frac{((kT + N_C) \otimes T_I)^2}{\sigma_C^2} - 1 \right) \frac{G(d'/2)G(-d'/2)}{ng(d'/2)^2} \tag{18}$$

$$= \left(\begin{array}{l} \frac{k^2}{\sigma_C^2}(T \otimes T_I)^2 \\ + \frac{2k}{\sigma_C^2}(T \otimes T_I)(N_C \otimes T_I) \\ + \frac{1}{\sigma_C^2}(N_C \otimes T_I)^2 - 1 \end{array} \right) \frac{G(d'/2)G(-d'/2)}{ng(d'/2)^2} \tag{19}$$

The expected value of $\hat{F}$ is

$$E[\hat{F}] = \left(\begin{array}{l} \frac{k^2}{\sigma_C^2}(T \otimes T_I)^2 \\ + \frac{2k}{\sigma_C^2}(T \otimes T_I)E[N_C \otimes T_I] + \\ \frac{1}{\sigma_C^2}E[(N_C \otimes T_I)^2] - 1 \end{array} \right) \frac{G(d'/2)G(-d'/2)}{ng(d'/2)^2} \quad . \tag{20}$$

Now, $E[N_C \otimes T_I] = 0$ (because the pixelwise mean of N_C is zero) and $E[(N_C \otimes T_I)^2] = \sigma_C^2$, so this simplifies to

$$E[\hat{F}] = \frac{k^2}{\sigma_C^2} \cdot \frac{G(d'/2)G(-d'/2)}{ng(d'/2)^2} (T \otimes T_I)^2 \quad . \tag{21}$$

Equation 15 shows that $\frac{k^2}{\sigma_C^2}$ in this equation is just the SNR of the classification image, and substituting Equation 16 for a linear observer's SNR, we find

$$E[\hat{F}] = \frac{(T \otimes T_I)^2}{1+\rho} \quad , \tag{22}$$

which Equation 12 shows to be the linear observer's absolute efficiency, F. Thus $\hat{F}$ is an unbiased estimator of a linear observer's absolute efficiency.

All terms in Equation 17 are easily calculated. The classification image C, the ideal template T_I, the number of trials n, and the observer's performance d' are obtained by the usual methods. The classification image is just a weighted sum of noise fields, so the pixelwise variance σ^2_C of the sampling noise can be very closely approximated by assuming that the individual noise fields are independent samples from a multivariate normal distribution (e.g., if the pixelwise variance is 0.04, and the number of hits, misses false alarms, and correct rejections are 75, 25, 25, and 75, respectively, then σ^2_C = 0.04/75 + 0.04/25 + 0.04/25 + 0.04/75 = 0.0043). In fact, the noise fields within each stimulus-response class of trials are not quite multivariate normal (noise fields fall in a given class because they have a particular relation to the observer's template), but this approximation is adequate for estimating the variability of a classification image.

We have made a number of approximations in this derivation. We have assumed that σ^2_C and d' are known exactly, although actually they are random variables. We have also assumed that the noise fields in each stimulus-response class of trials are multivariate normal, which again is only approximately true. To validate Equation 17 as an estimator of absolute efficiency, we ran simulated linear observers in classification image experiments, and tested whether Equation 17 really does predict their absolute efficiency. The relevant parameters of the linear model are sampling efficiency $(T \otimes T_I)^2$ and internal-to-external noise ratio ρ, and we tested linear model observers with a wide range of these parameters. We also varied the number of trials in the simulated experiment. Figure 6 plots the results for linear model observers with $\rho = 1$, and with a range of sampling efficiencies, as a function of the number of trials in the classification image experiment. The horizontal lines show the model observers' actual absolute efficiencies, and the data points show the efficiencies predicted by Equation 17, averaged over many simulations of the experiment. The plot shows that so long as there are 500 trials or more, the predictions are effectively unbiased. Results for linear model observers with different internal-to-external noise ratios ranging from $\rho = 0$ to $\rho = 2$ were virtually identical.

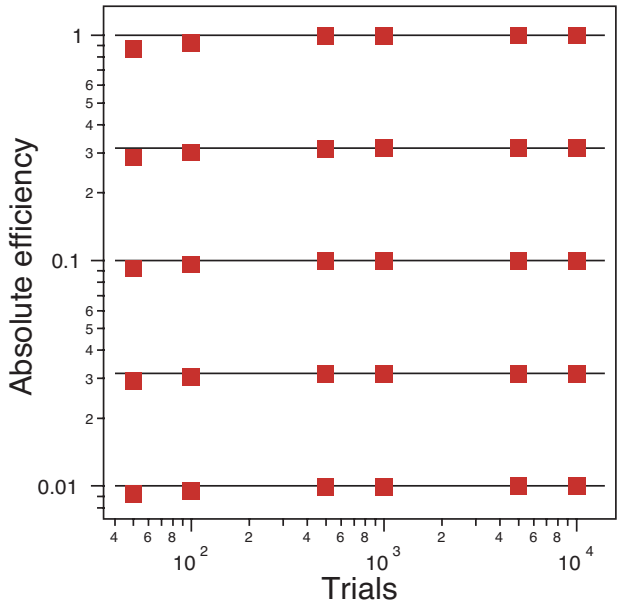

Figure 6. Predicted and actual efficiency for simulated linear observers. Observers had sampling efficiencies of 0.25 to 2.0, in steps of 0.25, and all observers had an internal-to-external noise ratio of $\rho = 1$, resulting in absolute efficiencies of $F = 0.125$ to $F = 2.0$, in steps of 0.125 (see Equation 12). Each horizontal line indicates one simulated observer's absolute efficiency (which does not depend on the number of trials). The red data points indicate the absolute efficiency predicted for the simulated observers from their classification images, using Equation 17. The predictions were effectively unbiased, so long as there were 500 trials or more.

Acknowledgments

This research was supported by grants from the National Eye Institute, the Natural Sciences and Engineering Research Council of Canada, and the Canada Research Chair Council. We thank Jason Gold, Stanley Klein, and Peter Neri for helpful discussions, and two anonymous reviewers for their comments.

Commercial relationships: none.
Corresponding author: Richard F. Murray.
Email: rfmurray@psych.upenn.edu.
Address: 3401 Walnut Street, room 302C, Philadelphia, PA 19104-6228.

References

Ahumada, A. J. (1987). Putting the visual system noise back in the picture. *Journal of the Optical Society of America A, 4*(12), 2372-2378. [PubMed]

Ahumada, A., Jr. (1996). Perceptual classification images from vernier acuity masked by noise [Abstract]. *Perception*(ECVP Suppl.), *25*, 18.

Ahumada, A., Jr. (2002). Classification image weights and internal noise level estimation. *Journal of Vision, 2*(1), 121-131. http://journalofvision.org/2/1/8/, doi:10.1167/2.1.8. [PubMed][Article]

Ahumada, A. J., & Beard, B. L. (1999). Classification images for detection [ARVO Abstract]. *Investigative Ophthalmology and Visual Science, 40*(4), S572.

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America, 49*(6), 1751-1756.

Baddeley, R., & Tripathy, S. P. (1998). Insights into motion perception by observer modeling. *Journal of the Optical Society of America A, 15*(2), 289-296. [PubMed]

Barth, E., Beard, B. L., & Ahumada, A., Jr. (1999). Nonlinear features in vernier acuity. In B. E. Rogowitz & T. N. Pappas (Eds.), *SPIE Proceedings: Vol. 3644: Human Vision and Electronic Imaging IV* (pp. 88-96). Bellingham, WA: SPIE.

Bennett, P. J. (1993). The harmonic bandwidth of phase-reversal discrimination. *Perception and Psychophysics, 53*(3), 292-304. [PubMed]

Burgess, A. E., & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A, 5*(4), 617-627. [PubMed]

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science, 214*(4516), 93-94. [PubMed]

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods and Research, 21*(1), 89-115.

Chubb, C., Econopouly, J., & Landy, M. S. (1994). Histogram contrast analysis and the visual segregation of IID textures. *Journal of the Optical Society of America A, 11*(9), 2350-2374. [PubMed]

Eckstein, M. P., & Ahumada, A. J. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision, 2*(1), i-i, http://journalofvision.org/2/1/i/, doi:10.1167/2.1.i. [PubMed][Article]

Gold, J. M., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature, 402*(6758), 176-178. [PubMed]

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology, 10*(11), 663-666. [PubMed]

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review, 71*(5), 392-407. [PubMed]

Green, D. M., & Birdsall, T. G. (1978). Detection and recognition. *Psychological Review, 85*(3), 192-206.

Legge, G. E., Kersten, D., & Burgess, A. E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A, 4*(2), 391-404. [PubMed]

Levi, D. M., & Klein, S. A. (2002). Classification images for detection and position discrimination in the fovea and parafovea. *Journal of Vision, 2*(1), 46-65. http://journalofvision.org/2/1/4/, doi:10.1167/2.1.4. [PubMed][Article]

Li, R. W., Levi, D. M., & Klein, S. A. (2004). Perceptual learning improves efficiency by re-tuning the decision 'template' for position discrimination. *Nature Neuroscience, 7*(2), 178-183. [PubMed]

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11*, 5-11.

Murray, R. F. (2002). Perceptual organization and the efficiency of shape discrimination. Doctoral thesis, Department of Psychology, University of Toronto. Electronic copy available from the author.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2001). No pointwise nonlinearity in shape discrimination [Abstract]. *Journal of Vision, 1*(3), 52a. http://journalofvision.org/1/3/52/, doi:10.1167/1.3.52.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision, 2*(1), 79-104. http://journalofvision.org/2/1/6/, doi:10.1167/2.1.6. [PubMed][Article]

Neri, P., & Heeger, D. J. (2002). Spatiotemporal mechanisms for detecting and identifying features in human vision. *Nature Neuroscience, 5*(8), 812-816. [PubMed]

Neri, P. (2004). Estimation of nonlinear kernels. *Journal of Vision, 4*(2), 82-91. http://journalofvision.org/4/2/2/, doi:10.1167/4.2.2 [PubMed][Article]

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A, 2*(9), 1508-1532. [PubMed]

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory, 4*, 171-212.

Richards, V. M., & Zhu, S. (1994). Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients. *Journal of the Acoustical Society of America, 95*(1), 423-434. [PubMed]

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology, 14*(5), 391-396. [PubMed]

Tanner, W. P. J., & Birdsall, T. G. (1958). Definitions of d' and η as psychophysical measures. *Journal of the Acoustical Society of America, 30*(10), 922-928.

Taylor, J. R. (1982). *An introduction to error analysis: The study of uncertainties in physical measurements*. Mill Valley, CA: University Science Books.

Van Trees, H. L. (1968). *Detection, estimation, and modulation theory*. New York: Wiley.