Research Report

SUPERSTITIOUS PERCEPTIONS REVEAL PROPERTIES OF INTERNAL REPRESENTATIONS

Frédéric Gosselin[1] and Philippe G. Schyns[2]

[1]*Université de Montréal, Montréal, Québec, Canada, and* [2]*University of Glasgow, Glasgow, Scotland, United Kingdom*

Abstract—*Everyone has seen a human face in a cloud, a pebble, or blots on a wall. Evidence of superstitious perceptions has been documented since classical antiquity, but has received little scientific attention. In the study reported here, we used superstitious perceptions in a new principled method to reveal the properties of unobservable object representations in memory. We stimulated the visual system with unstructured white noise. Observers firmly believed that they perceived the letter* S *in Experiment 1 and a smile on a face in Experiment 2. Using reverse correlation and computational analyses, we rendered the memory representations underlying these superstitious perceptions.*

For several decades, face, object, and scene recognition researchers have sought to understand the properties of representations in memory. However, the relationship between representations and behavior is tenuous, leaving researchers with few options except to test the validity of hypothesized representational schemes.

A few decades ago, Wiener (1958) showed that noise could be used to analyze the behavior of a black box, even suggesting that the brain could be studied this way. Here we propose a principled method, combining Wiener's idea with visual perception, for reconstructing the internal representation of an observer. We started from an unstructured external stimulus (white noise), and we led observers to believe that the stimulus comprised a signal. As white noise does not represent coherent structures in the image plane, the superstitious perception of a signal had to arise from the observers' share. To characterize these internal representations, we reverse correlated (e.g., Ahumada & Lovell, 1971; Beard & Ahumada, 1998; Gold, Murray, Bennett, & Sekuler, 2000; Neri, Parker, & Blakemore, 1999; Oshawa, De Angelis, & Freeman, 1990) the observers' detection and rejection responses with the corresponding white-noise stimuli.

EXPERIMENT 1: "S" AS IN "SUPERSTITIOUS"

Method

In Experiment 1, we instructed 3 paid naive observers (R.C., N.L., and M.J.; ages 21–24) to detect in white noise the presence of a target black letter *S* on a white background filling the image. The observers were told that the letter *S* (for "superstitious") was present on 50% of the 20,000 trials, which were equally divided into 40 blocks and completed over a fortnight. No more detail was given regarding the shape of the letter. The image presented on each trial consisted of static bit noise spanning 50 × 50 pixels (2° × 2° of visual angle), with a black-pixel density of 50%. No signal was ever presented. The experiment ran on a G4 Macintosh computer using a program written with the Psychophysics Toolbox for Matlab (Brainard, 1997; Pelli, 1997).

Results

The observers detected an *S* in noise on 22.7% (R.C.), 45.9% (N.L.), and 11% (M.J.) of the trials, respectively. They claimed that they responded positively whenever they saw an *S* and estimated the quantity of added noise to vary between 30% and 50%. Observer R.C. described her response strategy as, "I simply waited to see if the *S* jumped out at me."

To depict the information eliciting these superstitious perceptions, we applied reverse correlation. For each observer, we computed a "yes image" (vs. "no image") by adding together all the stimuli leading to detections (vs. rejections). We then subtracted the no image from the yes image to produce a classification image (see Fig. 1, inset a, for R.C., N.L., and M.J.). For each observer, the classification image represents the template of information that drove the detection of the target *S* letter; formally, it is the best least square linear fit to the detection data.[1]

Input white noise has equal energy at all spatial frequencies. It is therefore unbiased, and the expected energy of the classification image is constant across the whole spatial frequency spectrum. Such an "empty" classification image will occur if the observer responds randomly to the white-noise stimuli, either because the observer ignores the stimuli or because the observer hallucinates *S*s without any systematicity. A superstitious[2] (as opposed to a blind or hallucinating) observer will respond positively to white-noise fields when these correlate (even very weakly) with the observer's internal representation of an *S*. Consequently, any bias appearing in the spectral analysis

Address correspondence to Frédéric Gosselin, Département de Psychologie, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, Québec, Canada, H3C 3J7, e-mail: frederic.gosselin@umontreal.ca, or to Philippe Schyns, Department of Psychology, University of Glasgow, 58 Hillhead St., Glasgow, Scotland, United Kingdom, G12 8QB, e-mail: philippe@psy.gla.ac.uk.

1. We suppose that the observer matches two vectors on each trial of the experiment: a stimulus vector of dimensionality k and a template vector $\boldsymbol{\beta}$, of the same dimensionality, representing the memorized pattern to match against the input (e.g., the letter *S*). We also suppose that the observer's response is a linear function of this match. We can arrange the n stimulus vectors of the experiment in the $n * k$ matrix $\mathbf{X}$. A linear equation then describes the behavior of the observer in the experiment: $\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is an n-dimensional response vector, and $\boldsymbol{\varepsilon}$ is an n-dimensional vector of error random variables with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $V(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. The least square estimate of $\boldsymbol{\beta}$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Because the stimulus vectors are uncorrelated, we have $(\mathbf{X}'\mathbf{X})^{-1} = (k\mathbf{I})^{-1} = k^{-1}\mathbf{I}$. Therefore, $\boldsymbol{\beta} = k^{-1}\mathbf{X}'\mathbf{y}$. Leaving the constant k aside and assuming that the responses can have only the values 1 or −1, this last equation reduces to summing all the stimulus vectors that led to a response of 1 and subtracting from it the sum of all the stimulus vectors that led to a response of −1.

2. We call these perceptions superstitious because the correlation between the information template and input noise was extremely weak ($r = .026$ on average in Experiment 1, and still smaller in Experiment 2), even if we assume that the observer used the same unique template and detection criterion throughout the experiment.

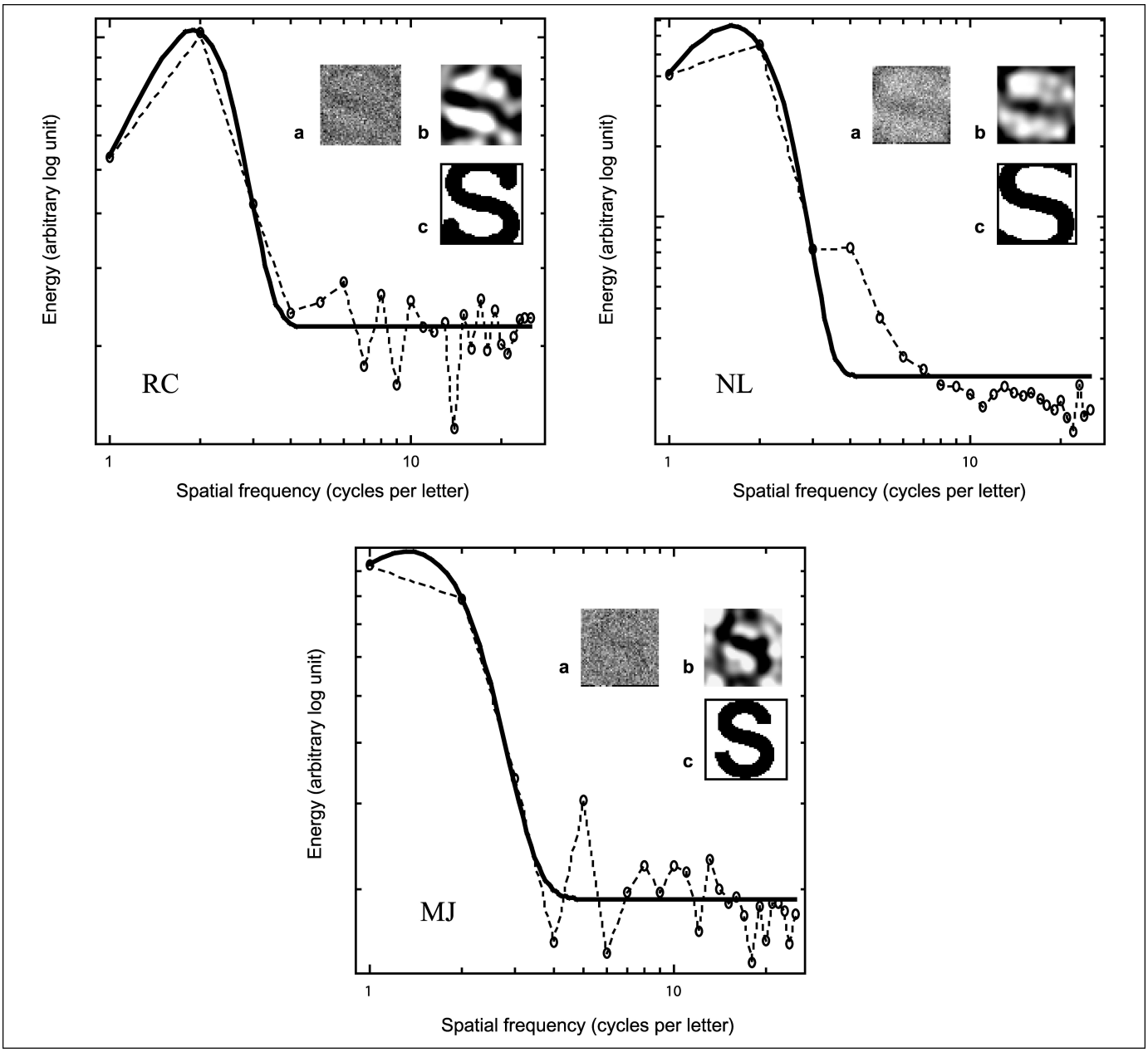

Fig. 1. Results of Experiment 1. The graphs show, for each observer, the distribution of the average squared amplitude energy for different spatial frequencies (collapsed across all orientations) of the raw classification images (expected energy = constant). The solid lines are the best Gaussian fits. The insets show (a) the raw classification images, (b) the classification images filtered with a smooth low-pass (Butterworth) filter with a cutoff at 3 cycles per letter, and (c) the best matches between the filtered classification images and 11,284 letters, each resized and cut to fill a square window in the two possible ways. For (b), we squeezed pixel intensities within 2 standard deviations from the mean.

of the raw classification images (see the curves in Fig. 1 for R.C., N.L., and M.J.) indicates the presence of structures that underlie the superstitious perceptions of the letter *S*. It also provides the means to render the observer's share.

We found such biases for information at slightly different bandwidths, depending on observer (R.C., 1.5–2.3 cycles/letter, peak = 1.9 cycles; N.L., 1.21–2.01 cycles/letter; peak = 1.61 cycles; M.J., 0.05–2.1 cycles/letter, peak = 1.3 cycles). Technically, we best-fitted a Gaussian density function (see Fig. 1, solid lines) to the energy distribution of each observer's raw classification image (Fig. 1, open circles). To determine the observer-specific bias, we computed the mean of each best Gaussian fit (R.C., R^2 = .97; N.L., R^2 =.99; M.J., R^2 = .97) and included all spatial frequencies 1 standard deviation away—that is, a bandwidth comprised between 0 and 3 cycles per letter. We rendered this information by filtering the classification images with a smooth low-pass filter (Butterworth) with a cutoff at 3 cycles. The outcomes

were black *S*s on white backgrounds filling the images (see Fig. 1, inset b, for R.C., N.L., and M.J.). Their spectral compositions were consistent with psychophysical findings indicating that letter identification is most efficient at about 3 cycles per letter (Pelli, Burns, Farell, & Moore, in press; Solomon & Pelli, 1994).

To test that the *S*s in the classification images were not simply the result of our own superstitious perceptions, we correlated these classification images with the 26 letters of the alphabet from 31 fonts,[3] in 7 styles (normal, italic, bold, underline, outline, condense, extend) and in upper- and lowercase, for a total of 11,284 Pearson correlations. All letters were resized and cut to fill a square window in the two possible ways (i.e., with the width of the letter occupying the whole width of the window or the height of the letter occupying the whole height of the window). The highest correlations were obtained with the following letters (see Fig. 1, inset c, for each observer): for R.G., an uppercase Courier New bold *S* scaled horizontally ($r = .557$); for N.L., a lowercase Verdana regular-style *S* scaled horizontally ($r = .553$); and for M.L., an uppercase Arial bold *S* scaled vertically ($r = .704$). On average, confounding font, style, case, and observer, the largest correlation between the classification images and the 26 letters of the alphabet was found for *S* (see Fig. 2), a pattern true for each observer.

In sum, we induced superstitious perceptions of *S*s by instructing 3 observers to detect this letter in noise. They did not know that the stimuli never comprised the letter, but only white noise. Thus, if the observers had performed only according to the stimulus (i.e., in a bottom-up manner), their classification images should have had the same properties as averaged white noise—that is, constant energy across all spatial frequencies. However, there were marked peaks of energy below 3 cycles per letter. These must have arisen from top-down influences on the interpretation of white noise—very low correlations between input noise and the memory representations of the letter. Further analyses revealed the shape of the letters that the observers thought they saw.

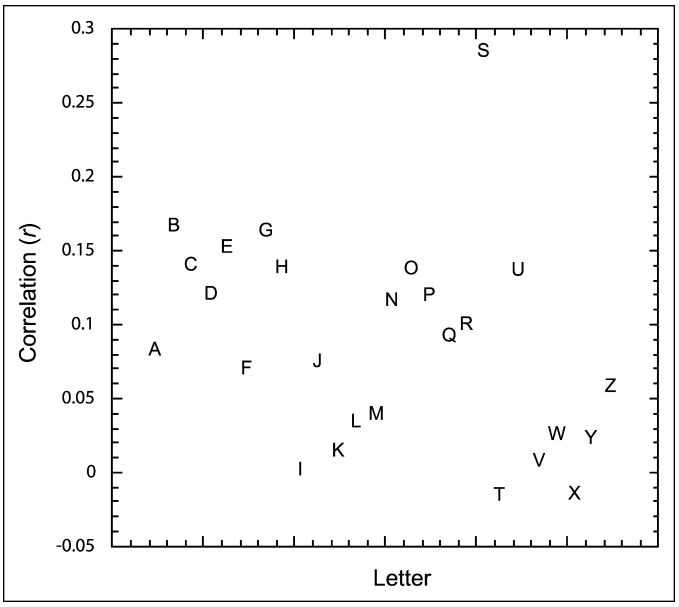

Fig. 2. Average Pearson correlation coefficients between the three filtered classification images depicted in insets b in Figure 1 and the 26 letters of the alphabet in 31 fonts, seven styles, and upper- and lowercase versions (total of 11,284 stimuli; each letter was resized and cut to fill a square window in the two possible ways).

EXPERIMENT 2: SIMILE SMILE

Method

In Experiment 2, we sought to generalize this rendering of represented visual information to a more complex representation, using 2 other observers. We instructed 2 female observers (A.R., age 26; H.P., age 23) to discriminate between a smiling and nonsmiling face embedded in noise. Each observer was told that the smiling face was present on 50% of the 20,000 trials, which were equally divided into 40 blocks and ran over a fortnight. To ensure that the observers focused on detecting the features of a smile, we gave no details regarding the alternative expressions. In each trial, one sparse image spanning 256×256 pixels ($5.72° \times 5.72°$ of visual angle) was presented. To create each image, we randomly sampled 27.5% of the black pixels of the contours of a face without a mouth (indicated in red in Fig. 3, insets a and b) and filled the remainder of the image with bit noise with the same density of black pixels. No signal was therefore presented in the mouth area. The experiment ran on a Macintosh G4 computer using a program written with the Psychophysics Toolbox for Matlab (Brainard, 1997; Pelli, 1997).

Results

The observers detected a smile on 7.07% (H.P.) and 48.4% (A.R.) of the trials. Observer H.P. explained that she had been very conservative and responded "yes" only when she was absolutely certain that the face was indeed smiling. She added that she looked for teeth and used the eyes and the nose to locate the mouth. Observer A.R. reported that she was focusing mostly on the junctions of the lips.

To render their internal representations of a smile, we first computed the raw classification images as explained for Experiment 1 (see Fig. 3, inset a, for H.P. and A.R.). Following spectral analyses of the classification images and Gaussian fits (H.P., $R^2 = .83$; A.R., $R^2 = .95$) of their energy distributions, an information bias appeared for H.P. between 0 and 13.69 cycles per face, with a peak at 0.65 cycles; for A.R., the critical bandwidth was between 0.92 and 5.47 cycles per face, with a peak at 3.192 cycles (see the curves in Fig. 3).

These biases are consistent with the most efficient bandwidth described in the literature on the identification of facial expressions—that is, maximum efficiency centered at 8 cycles per face (Bayer, Schwartz, & Pelli, 1998). We revealed this information by filtering the classification images with a low-pass cutoff (at 14.27 cycles/face for H.P. and 5.84 cycles for A.R.; see Fig. 3, insets b). The outcomes rendered the internal representation of a smile revealing the teeth for H.P., and a smile with well-defined junctions of the lips for A.R.

As in Experiment 1, we validated the content of the classification images by correlating their mouth area (see Fig. 3, inset c, for H.P. and A.R.) with the mouths of 16 individuals (8 males and 8 females, from

3. The fonts were Andala Mono, Apple Chancery, Arial, Book Antiqua, Bookman Old Style, Capitals, Century Gothic, Century Schoolbook, Charcoal, Chicago, Comic Sans MS, Courier, Courier New, Gadget, Geneva, Georgia, Helvetica, Impact, LED, Letter Gothic MT, Mishawaka, Monaco, New York, Sand, Techno, Teletext, Textile, Times, Times New Roman, Trebuchet MS, and Verdana.

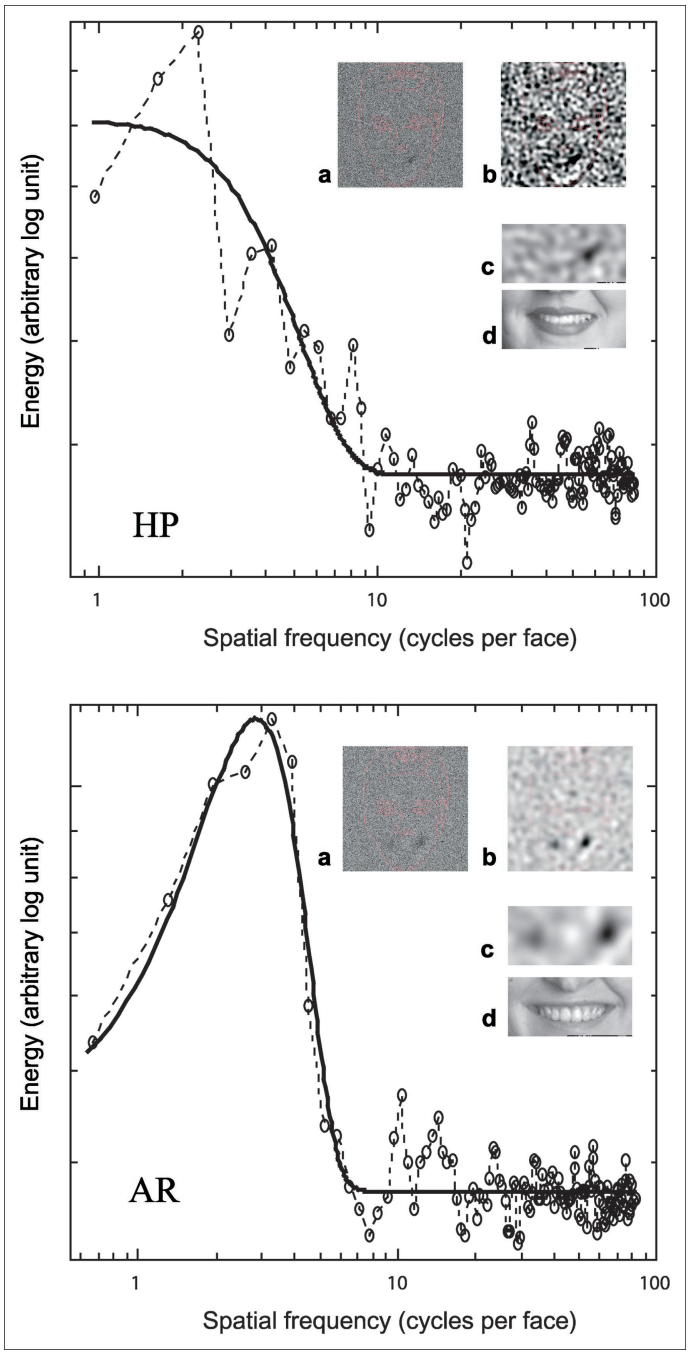

Fig. 3. Results of Experiment 2. The graphs show, for each observer, the distribution of the average squared amplitude energy for different spatial frequencies (collapsed across all orientations) of the raw classification images. (Each open circle is the average of two successive data points for H.P.) The solid lines are the Gaussian best fits. The insets show (a) the raw classification images, (b) the classification images filtered with a low-pass (Butterworth) filter (cutoff = 14.27 and 5.84 cycles/face, for H.P. and A.R., respectively), (c) areas of the filtered classification images that were correlated with the corresponding area of 48 face stimuli (three expressions each for 8 men and 8 women), and (d) the face stimuli with the largest correlations with the filtered classification images. In (a) and (b), the red indicates the contour of the mouthless face. For (c), we squeezed pixel intensities within 2 standard deviations from the mean.

Schyns & Oliva, 1999), each displaying three different expressions (neutral, happy, and angry), for a total of 48 correlations per observer. The largest correlations (r = .24 and .32, for H.P. and A.R., respectively) were with happy women (see Fig. 3, inset d, for H.P. and A.R.). The average correlations for the happy, neutral, and angry mouths were .08, −.005, and .002, respectively, for H.P. and .137, −.0004, and .009, respectively, for A.R.

DISCUSSION

We have presented a method for reconstructing unobservable representations from superstitious perceptions. In white noise, we elicited superstitious perceptions of *S*s in Experiment 1 and of smiles in Experiment 2. Reverse correlation rendered the internal representations underlying these perceptions. Notably, the internal representations had spectral properties compatible with those that recognition studies have reported observers use. This result is in line with recent findings demonstrating that some neurons in the human medial temporal lobe respond both to bottom-up visual inputs and to the top-down mental visualizations of these visual inputs (Kreiman, Koch, & Fried, 2001). To our knowledge, this is the first time that the representations underlying object recognition have been depicted in the absence of a systematic bottom-up signal.

The depiction of complex, psychologically validated object representations from white noise is genuinely new. Studies close in spirit to this research have successfully mapped the low-level receptive fields of single neurons (see De Angelis, Ohzawa, & Freeman, 1995, for a review). However, these representations depict a level of visual organization considerably lower than object recognition. At such higher levels, noise has been applied to study the nature of illusory contours (Gold et al., 2000) and the discrimination of letters (Watson, 1998). However, in all of these cases, the statistics of the input space have comprised a signal in addition to the noise. This stimulation with a signal plus noise is biased, unlike our stimuli, which comprised only unbiased white noise.

Consider the study of Gold et al. (2000), in which reverse correlation depicted the illusory contours observers perceived to classify noisy Kaniza squares as concave versus convex. In 20% of the trials of a within-subjects design, the real contours (concave vs. convex) were presented in white noise. This signal biased the input distribution, which, in turn, could have biased perception in the illusory conditions (when only noise was presented). Whether or not the classification images derived in these conditions represent more than input signal artificially internalized for the sole purpose of the experiment is a generic issue that pervades the field. To avoid these difficulties altogether, we went back to Wiener's (1958) original idea and used only white noise to depict the observer's share.

To the extent that white noise can be weakly correlated with every visual stimulus, the technique could be applied to a wide range of visual and auditory events. However, there are also serious limitations to the technique, arising from its linearity. For example, the observer needs to be properly instructed that the same target is always presented, that it does not change position across trials, that it always has the same black-on-white contrast, and so forth. Even though it is theoretically possible to extend the technique to nonlinear problems (Wiener, 1958), it is practically difficult to do so because of (a) the required number of trials and (b) the required time for data analyses. In any case, psychology needs new techniques to characterize the properties of memorized information. Here, we developed a technique that reveals these properties from the superstitious perceptions of objects in white noise.

Acknowledgments—We thank Liza Paul, Lizann Bonnar, Benoit A. Bacon, and Simon Garrod for proofreading drafts of this article. This research was supported by Economic and Social Research Council Grant R000237901.

REFERENCES

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

Bayer, H.M., Schwartz, O., & Pelli, D. (1998). Recognizing facial expressions efficiently. *Investigative Ophthalmology & Visual Science*, *39*, S172.

Beard, B.L., & Ahumada, A.J. (1998). A technique to extract the relevant features for visual tasks. In B.E. Rogowitz & T.N. Pappas (Eds.), *Human vision and electronic imaging III* (SPIE Proceedings Vol. 3299) (pp. 79–85). Bellingham, WA: International Society for Optical Engineering.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

De Angelis, G.C., Ohzawa, I., & Freeman, R.D. (1995). Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences*, *18*, 451–458.

Gold, J., Murray, R.F., Bennett, P.J., & Sekuler, A.B. (2000). Deriving behavioral receptive fields for visually completed contours. *Current Biology*, *10*, 663–666.

Kreiman, G., Koch, C., & Fried, I. (2001). Imagery neurons in the human brain. *Nature*, *408*, 357–361.

Neri, P., Parker, A.J., & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature*, *401*, 695–698.

Oshawa, I., De Angelis, G.C., & Freeman, R.D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, *249*, 1037–1041.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Pelli, D.G., Burns, C.W., Farell, B., & Moore, D.C. (in press). Identifying letters. *Vision Research*.

Schyns, P.G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, *69*, 243–265.

Solomon, J.A., & Pelli, D.G. (1994). The visual filter mediating letter identification. *Nature*, *369*, 395–397.

Watson, A.B. (1998). Multi-category classification: Template models and classification images. *Investigative Ophthalmology and Visual Science*, *39*, S912.

Wiener, N. (1958). *Nonlinear problems in random theory.* New York: Wiley.

(RECEIVED 4/25/01; REVISION ACCEPTED 1/29/02)