



Contents lists available at ScienceDirect

Vision Research

journal homepage: [www.elsevier.com/locate/visres](http://www.elsevier.com/locate/visres)

## Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model

Yaffa Yeshurun<sup>a,\*</sup>, Marisa Carrasco<sup>b</sup>, Laurence T. Maloney<sup>b</sup>

<sup>a</sup> Department of Psychology, University of Haifa, 31905 Haifa, Israel

<sup>b</sup> Department of Psychology, Center for Neural Science, New York University, 6 Washington Place, New York, NY 10003, USA

### ARTICLE INFO

#### Article history:

Received 23 April 2007

Received in revised form 24 April 2008

#### Keywords:

Psychophysics

Threshold estimation 2-IFC

Bias

### ABSTRACT

We assess four common claims concerning the two-interval forced choice (2-IFC) procedure and the standard Difference Model of 2-IFC performance. The first two are (1) that it is unbiased and (2) that the structure of the 2-IFC task does not in itself alter sensitivity. The remaining two concern a claimed  $\sqrt{2}$  enhancement in sensitivity in 2-IFC relative to that measured in a Yes–No task. We review relevant past research and re-analyze seventeen experiments from previous studies across three laboratories. We then report an experiment comparing 2-IFC performance with performance in a second task designed to elucidate observers' decision processes. This second task is simply two successive Yes–No signal detection tasks with the same timing as in the 2-IFC experiment. We find little evidence supporting the claims that 2-IFC is unbiased and that it does not alter sensitivity and we also reject the two claims associated with the Difference Model as a model of performance in our own experiment.

© 2008 Elsevier Ltd. All rights reserved.

### 1. Introduction

Human performance in psychophysical experiments depends on both sensory and decisional processes. To reliably measure the sensitivity of the sensory processes one should ensure that decisional processes and psychophysical methods do not distort sensitivity measurements. A widely used method for assessing sensitivity is the two-interval forced choice (2-IFC) paradigm. In a 2-IFC task, a single experimental trial consists of two temporal intervals. The signal is presented in one and only one of the intervals and the observer is required to report the interval (first vs. second) in which the signal was presented.

The 2-IFC procedure is employed in a wide range of applications in vision, audition, cognition and other fields. Many users believe that (1) it is “unbiased” ( $p_1 = p_2$  where  $p_i$  is the probability of a correct response when the signal is in the  $i$ -th interval), (2) that the procedure itself does not affect the sensitivity of the observer, and that (3) there is a  $\sqrt{2}$  enhancement of the signal consistent with a particular model of 2-IFC performance, the Difference Model, which we present below. Over the course of this article we will formulate these claims precisely (splitting the third claim into two claims) and review previous work relevant to each claim. We will report re-analyses of data from published experiments to evaluate the first claim and we will report an experiment designed to test the latter claims. To summarize our conclusions, we find little sup-

port for any of these claims. In the last part of the paper, we discuss the implications for use of 2-IFC and other procedures.

#### Claim 1: $p_1 = p_2$

The 2-IFC paradigm is widely used because it is considered to reduce or eliminate bias. Green and Swets (1973), for example, state that “... the principal value of the forced choice task is that it practically eliminates the need to deal with the observer's decision criterion. Since the errors in a forced choice, unlike the errors in a Yes–No task, do not differ intrinsically in cost, observers find it more natural to maintain the symmetrical criterion.” (p. 108). Egan (1975, pp. 44ff) gives a succinct summary of all of the assumptions needed to derive the standard Difference Model of 2-IFC performance (which we present below) and the claim that performance will be unbiased. However, he does not assess whether the assumptions are likely to be satisfied in any particular application. More recent texts are cautious. For instance, as we will review below, both Wickens (2002, pp. 93ff) and Macmillan and Creelman (2005, pp. 175ff) discuss possible failures of the Difference Model or its assumptions, but nevertheless recommend the use of the 2-IFC procedure, stating that “... the procedure discourages bias. (p. 179)”.

In this article, we first report the results of re-analyses of published data from seventeen 2-IFC experiments for which we were able to recover data for performance in both intervals separately. The experiments were carried out in three separate laboratories and were used to measure visual sensitivity of very different kinds

\* Corresponding author. Fax: +972 4 8240966.

E-mail address: [yeshurun@research.haifa.ac.il](mailto:yeshurun@research.haifa.ac.il) (Y. Yeshurun).

of tasks. In conclusion, we find large interval biases in all of these studies. Across studies we found biases that favored either the first or the second interval. We report all of the data we could obtain. Moreover, we found that many of our colleagues could not provide us with data to analyze for possible biases because they do not separately record responses in both intervals of a 2-IFC trials. Consequently, there is no way to determine whether or not large interval biases were present in their experiments.

The results of this initial analysis led us to consider other claims concerning 2-IFC commonly made. In addition to the issue of bias, a second issue surrounding the 2-IFC procedure concerns sensitivity: what exactly do 2-IFC experiments measure? For many applications, the experimenter wants only indices of sensitivity that can be compared across two or more conditions within a single experiment. However, if we wish to translate 2-IFC performance into a measure of sensitivity that can be compared to measures of sensitivity derived from other psychophysical procedures, we need a credible model of what the observer is doing in 2-IFC tasks. Accordingly, we briefly review the literature testing the standard Difference Model of 2-IFC and find that it is inconclusive: while there is little reason to accept the model as an accurate model of what observers do in 2-IFC experiments, previous work also gives us no conclusive ground to reject the Difference Model.

Accordingly, we performed an experiment where we compared performance in a 2-IFC task with performance in a second task composed of two independent Yes–No tasks with the same stimuli and timing as the 2-IFC task: stimuli can appear in either, both or neither interval. The results of this experiment allow us to reject the Difference Model and also give us some insight into the observer's decision processes in two-interval experiments.

Our conclusions are two. First, 2-IFC in many applications is not bias free or approximately so. Second, based on previous work and the experiment reported here, the Difference Model of 2-IFC performance is not appropriate for many applications. As a field, we currently do not have a credible process model of what observers do in psychophysical experiments that involve comparison across time and we have no basis to generalize experience with one sort of stimulus to experiments with different stimuli. We recommend that experimenters use 2-IFC methods with caution, if at all, recording data by interval, testing for bias, and reporting it when found. Moreover, we stress that if a difference is found between the two intervals, it is necessary to analyze the effect of each variable or possible variable interactions in each interval. Finally, before any claim could be made based on the Difference Model, one needs to test whether the assumptions of the model hold under the specific experimental conditions of the study.

## 2. Re-analysis of previous 2-IFC studies

To test whether the 2-IFC paradigm is bias free, or nearly so, we re-analyzed several sets of data from seventeen experiments that have employed a 2-IFC paradigm. These different studies are all published and a detailed description of their goals, methods, and findings can be found in the referenced papers. A short summary of their methods, including a description of stimuli, is given in Appendix A. Here, we are primarily interested in the problem raised by Klein (2001). Does performance as measured by proportion-correct differ significantly in the two intervals? For all data sets, we performed nested hypothesis tests (Mood, Graybill, & Boes, 1974, pp. 441) to measure the difference in performance in Interval 1 vs. Interval 2. We tested the hypothesis  $H_0 : p_1 = p_2$ , where  $p_1$  is the probability of a correct response when the signal is in the first interval, and  $p_2$  is the probability of a correct response when the signal is in the second interval, vs. the alternative  $H_1 : p_1 \neq p_2$ . The details of the test are described in Appendix B.1.

We refer to a difference in probability of correct response in the two intervals as an *interval bias* for brevity.

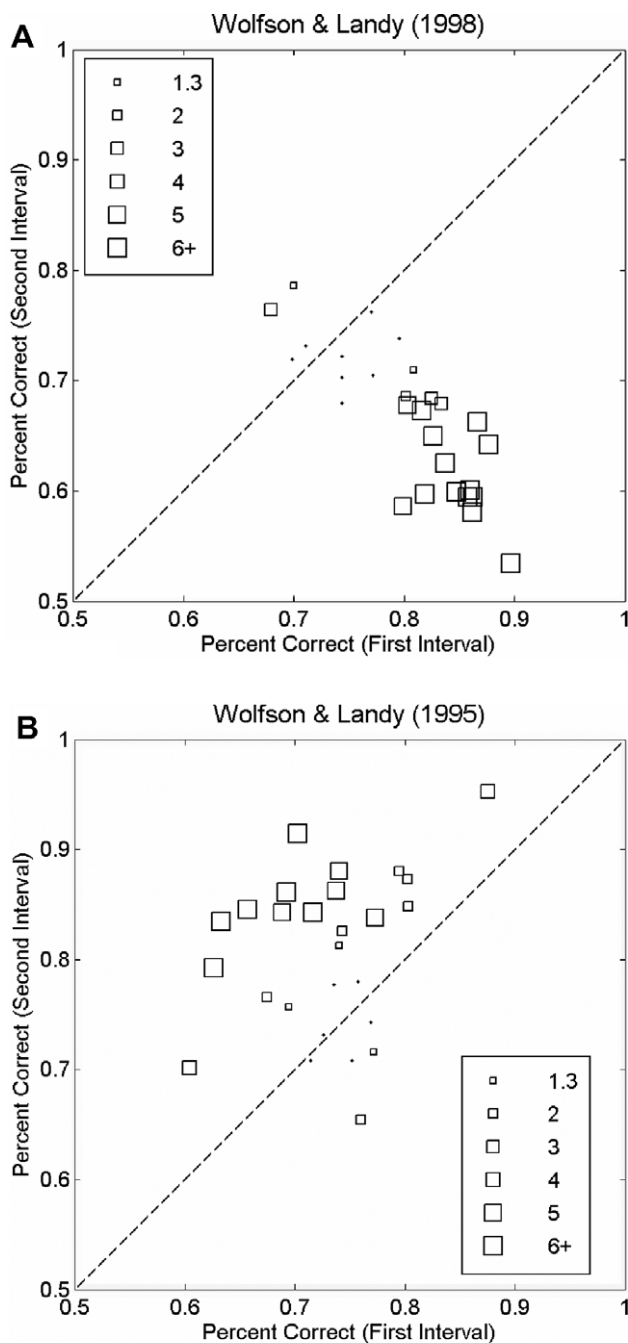
The results of the hypothesis tests include  $p$ -values that express the consistency between the observed data and the hypothesis under test. Under the null hypothesis, we expect that  $p$  will itself be a random variable distributed uniformly across the interval [0,1]. Under the alternative hypothesis, the values of  $p$  will tend to be small, and values of  $p$  below a conventional level (e.g.  $p < .05$ ) are taken as evidence to reject the null hypothesis.

Here, we will report the exact  $p$ -values of tests. These values have a wide range and it is hard to tell  $p < .00001$  from  $p < .000001$  at a glance. It is convenient to use  $\Gamma = \lfloor -\log_{10} p \rfloor$  as an index of the evidence against the null hypothesis in the data. The brackets in the definition imply that we round the value of  $\Gamma$  down to the nearest integer. Thus, if  $p = 10^{-6}$ ,  $\Gamma = 6$ , and if  $p = 10^{-3}$ ,  $\Gamma = 3$ . The conventional level of  $p < .05$  translates to  $\Gamma > 1.3$  and in many of the plots we will report values of  $\Gamma$  between 1.3 and 2 as well. We report values of  $\Gamma > 6$  as 6. Note that any value of  $\Gamma \geq 1.3$  would usually lead to rejection of the null hypothesis  $H_0 : p_1 = p_2$  in a single null hypothesis test. Note that the  $\Gamma$  measure is not a measure of magnitude of the effect (or any information associated with effect size or Type II Error). It is simply Type I Error, transformed logarithmically and reversed in sign.

In Fig. 1A, we plot  $p_2$  vs.  $p_1$  for seven observers in each of four experimental conditions reported in a texture discrimination study by Wolfson and Landy (1998; see description in Appendix A.1). The value of  $\Gamma$  is encoded as the radius of the symbol plotted at each point when it is in the range 1.3–6 as shown in the legend. When  $\Gamma < 1.3$  (and therefore  $p > .05$ ) the point is plotted as a single dot. Note first that there are 20 out of 28 points for which  $\Gamma \geq 1.3$  ( $p \leq .05$ ), there are 17 points with  $\Gamma \geq 3$  corresponding to  $p < .001$  in the nested hypothesis test, and 12 points with  $\Gamma \geq 6$ , corresponding to  $p < .000001$  in the nested hypothesis test. The distance from the diagonal line is an index of the magnitude of the difference  $p_2 - p_1$ . It is clear that there are large, significant deviations from the hypothesis that  $p_1 = p_2$  for the majority of observer-conditions in this experiment.

In Fig. 1B, we plot  $p_2$  vs.  $p_1$  for three observers in each of nine experimental conditions reported in a texture segmentation study by Wolfson and Landy (1995; see description in Appendix A.2). The plotting format is the same as that of Fig. 1A. Again, it is clear that there are large, extremely significant deviations from the hypothesis that  $p_1 = p_2$  for the majority of the observer-conditions. We note that the differences favored the first interval in Fig. 1A but favor the second in Fig. 1B. Since there is good agreement among observers within each experimental condition it seems unlikely that the observed interval biases are based on idiosyncratic preferences by observers for one interval over the other. We note that trained observers participated in both studies.

A possible explanation for the observed difference in direction of the interval bias in Wolfson and Landy (1995), (1998) is that the temporal spacing between the two presentations of possible targets is too short and one interval is somehow “masking” the other (Alcalá-Quintana & García-Pérez, 2005). In Fig. 2, we present the results for 15 observers for two ISI (inter-stimulus interval) conditions in a single experiment involving spatial frequency discrimination (McIntosh et al., 1999; see description in Appendix A.3). The ISI was either 500 ms (red symbols) or 4000 ms (blue symbols). The plot formats are identical to those of Fig. 1A and B. Again, there are marked and highly significant interval biases for several observers and both ISIs. Note that in both conditions some observers do significantly better in Interval 1 than in Interval 2 and some do significantly better in Interval 2 than in Interval 1, but the former is more frequent. In Fig. 3, we plot the value  $-\log_{10} p_{4000}$  vs.  $-\log_{10} p_{500}$  (the  $\Gamma$  values without truncation or rounding) to see



**Fig. 1.** The results of the re-analysis performed on: (A) Data from a texture discrimination study by Wolfson and Landy (1998). (B) Data from a texture segmentation study by Wolfson and Landy (1995). The distance of each data point from the diagonal line is an index of the difference between Percent Correct in the second interval and Percent Correct in the first interval ( $p_2 - p_1$ ). The radius of the symbol plotted at each point represents the  $p$ -value of the hypothesis tests [ $H_0 : p_1 = p_2$ ;  $H_1 : p_1 \neq p_2$ ] when it is in the range 0.05–0.000001 (i.e., when  $-\log_{10}p$  is in the range 1.3–6 as shown in the legend). When  $p > .05$  the point is plotted as a single dot. It is clear that there are large, extremely significant deviations from the hypothesis that  $p_1 = p_2$  for the majority of observer-conditions in these studies.

whether observers who had a bias of performance with one ISI also had a bias with the other. The Pearson's product moment correlation between the two measures is  $\hat{\rho} = 0.44$  (if we omit the two evident outliers near the top of the plot, it rises to  $\hat{\rho} = 0.51$ ). Thus, while the same observers tended to have biases in both conditions, the correlation is modest and accounts for less than 20% of the variance. To summarize, varying the ISI from 500 to 4000 ms seemed

to have no effect on the presence or absence of interval biases in a substantial number of observers.

Last of all, we summarize results for five visual search experiments with a total of 56 observers taken from Carrasco and Yeshurun (1998; see description in Appendix A.4). The results are shown in Fig. 4 for all of the different experiments combined. The different experiments are coded by color as explained in the figure caption. There are fewer violations proportionately, possibly because the accuracy in these experiments was relatively high to allow collection of reliable measurements of RT, but still  $\Gamma \geq 1.3$  ( $p < .05$ ) for more than 40% of observer-experiments (24 out of 56). Further analysis of the two attentional conditions included in these experiments (cued vs. neutral trials; see Appendix A.4 for details) revealed biases in performance even on trials in which a spatial cue attracted spatial attention to the target location prior to the presentation of the search display. That is, the asymmetrical performance was present regardless of whether attention was focused in advance on the relevant location.

The data sets presented were those for which we could obtain separate data for the first and second intervals. There were large interval biases in all of the studies considered. Our results support Klein's conjecture: there can be large differences in proportion-correct in the two intervals of a 2-IFC experiment. Interestingly, the presence or absence of these biases does not depend in any obvious way on: (a) the type of experiment or the complexity level of the display (e.g., a single sinusoidal grating in McIntosh et al., 1999 vs. a complex multi-element display of various color-shape combinations in Carrasco & Yeshurun, 1998); (b) whether or not spatial attention is focused on the target location (cued vs. neutral trials in Carrasco & Yeshurun, 1998); (c) the duration of the ISI between the two presentations (500-ms vs. 4000-ms in McIntosh et al., 1999); or (d) whether or not the observers are experienced (e.g., highly experienced observers in Wolfson & Landy, 1995 vs. naïve, inexperienced observers in Carrasco & Yeshurun, 1998).

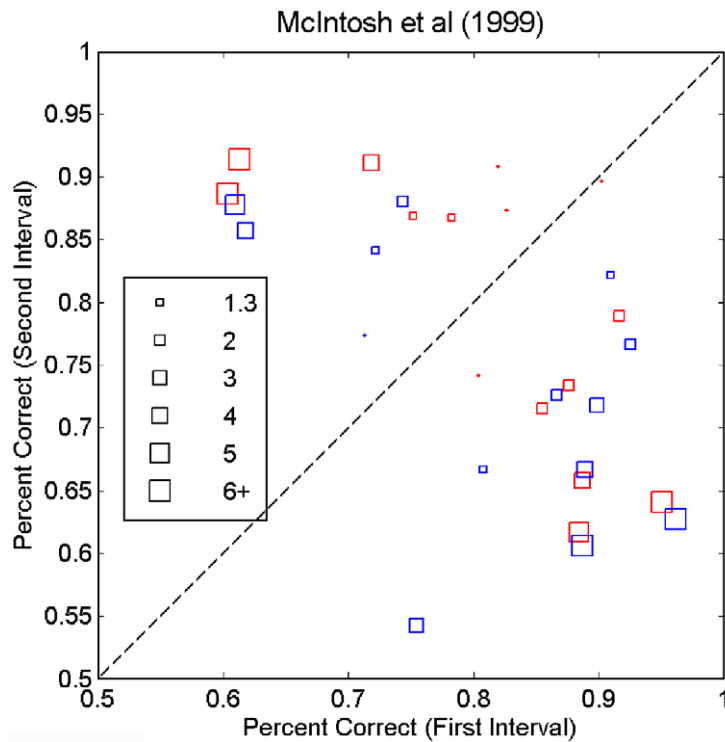
It is sometimes claimed that experienced psychophysical observers show little or no bias, that is, bias is due to lack of experience: "Occasionally a subject will show a preference for one or another interval; such a preference is usually eliminated with further practice ... (Green & Swets, 1973, p. 108)". Others make similar observations (Köhler, 1923; Needham, 1934) in discussing interval biases. We find no evidence to support these claims.

Finally, based on Fig. 3, some observers tend to have biases across conditions, but the correlation is modest. Thus, this finding of an asymmetrical performance shows no obvious patterned dependence on factors such as attentional state, complexity, ISI or practice.

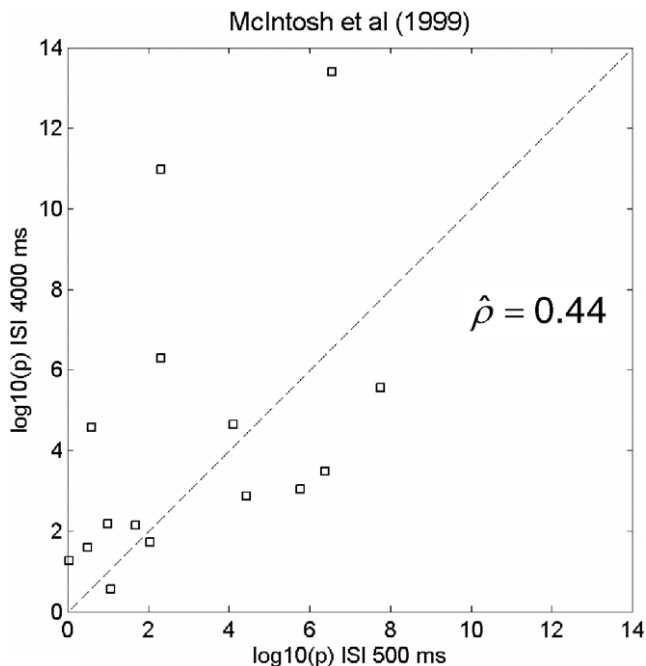
The reader's initial reaction might be to attribute these biases to a bias in "guessing": when observers have "no idea" whether the signal was in the first or second interval, they guess: some observers stereotypically select the first interval in guessing, some the second. A first difficulty with this explanation is that observers in Wolfson and Landy (1995) and in Wolfson and Landy (1998) tend to have biases in the same direction within each experiment. It is difficult to see how observers coordinated their "stereotypical preferences". Moreover, there is a major theoretical difficulty with this explanation. We remind the reader that, in the Difference Model of 2-IFC performance, there is no role assigned to guessing, and for that reason it is often described as "criterion-free". Consequently, whatever the origin of the bias, it corresponds to a failure of this standard Difference Model which we review next.

### 3. The difference model

All of the models we consider are examples of optimal Bayesian classifiers which, subject to constraints imposed on the observer,



**Fig. 2.** The results of the re-analysis performed on data from a spatial frequency discrimination study by McIntosh et al (1999). The distance of each data point from the diagonal line is an index of the difference: Percent Correct in second interval – Percent Correct in first interval ( $p_2 - p_1$ ). The radius of the symbol plotted at each point represents the  $p$ -value of the hypothesis tests [ $H_0 : p_1 = p_2$ ;  $H_1 : p_1 \neq p_2$ ] when it is in the range 0.05–0.000001 (i.e.,  $-\log_{10}p$  is in the range 1.3–6 as shown in the legend). When  $p > .05$  the point is plotted as a single dot. The red symbols depict data from the condition in which the ISI between the intervals was 500 ms and blue symbols depict data from the condition in which the ISI was 4000 ms. Here too, there are marked and highly significant differences in performance between the two intervals for several observers and both ISI conditions.



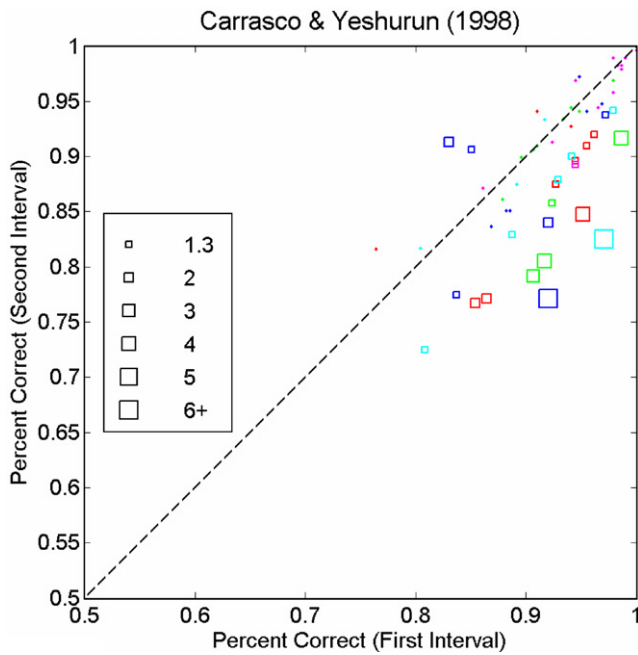
**Fig. 3.** A plot of  $-\log_{10}p$  by ISI condition across observers (McIntosh et al, 1999),  $p$  is the  $p$ -value of the hypothesis tests. The Pearson's product moment correlation between the two measures is  $\hat{\rho} = 0.44$ . This modest correlation can account for less than 20% of the variance, and it indicates that, for a substantial number of observers, varying the ISI from 500 to 4000 ms seemed to have no effect on the presence or absence of interval biases.

maximize the expected proportion of correct responses (Duda, Hart, & Stork, 2000, chap. 2). In the standard Difference Model (Egan, 1975, p. 44ff; Macmillan & Creelman, 2005, pp. 165ff; Wickens, 2002, pp. 93ff), the observer records a measure of sensory activity,  $S_1$ , in the first interval and compares it to a measure of sensory activity,  $S_2$ , in the second. This model is sometimes referred to as the 'trace' model (e.g., Berliner & Durlach, 1973; Durlach & Braitin, 1969).

These measures  $S_1, S_2$  are typically assumed to be statistically independent and identically distributed. They may be multivariate, corresponding to the activities of many visual mechanisms. If the distributions of the two measures are known, the optimal Bayesian classifier simply computes the posterior probability that the observed pattern of activity is due to the presence of a signal in the first interval, taking into account the prior probabilities that signals occur in each interval. If this posterior probability is greater than .5, the observer responds "first interval," otherwise, "second". The rule just described is optimal in the sense that the observer has the highest possible expected rate of correct response (Duda et al., 2000, chap. 2).

In the special case where the random variables  $S_1, S_2$  are both univariate Gaussians<sup>1</sup> with mean  $d' > 0$  and variance 1 when the signal is present in the corresponding interval and mean 0 and variance 1 when the signal is absent, then we can derive a simple form for the decision rule. In Fig. 5A, for example, we show a hypothetical distribution for activity in a task where the sensory signal in each interval is unidimensional. The distribution is bimodal, one mode corresponding to the signal in the first interval, the second to the signal in the second interval. Each mode is bivariate Gaussian and we

<sup>1</sup> See Duda et al. (2000) for the general case.



**Fig. 4.** The results of the re-analysis performed on data from five visual search experiments performed by Carrasco and Yeshurun (1998). The different experiments are coded by color as follows: Magenta – Orientation; Green – Color X Orientation; Blue – Long Color X Orientation; Red – Color X Shape; Cyan – Ls (see Appendix A.4 for details regarding the different experiments). The distance of each data point from the diagonal line is an index of the difference: Percent Correct in second interval – Percent Correct in first interval ( $p_2 - p_1$ ). The radius of the symbol plotted at each point represents the  $p$ -value of the hypothesis tests [ $H_0 : p_1 = p_2$ ;  $H_1 : p_1 \neq p_2$ ] when it is in the range 0.05–0.000001 (i.e.,  $-\log_{10}p$  is in the range 1.3–6 as shown in the legend). When  $p > .05$  the point is plotted as a single dot. Significant ( $p < .05$ ) performance biases were found for 24 out of the 56 observers. These results indicate that this bias is present regardless of task complexity.

assume that its standard deviations are  $\sigma_x = \sigma_y = 1$  and that activity in the intervals is uncorrelated. On each trial, the observer records the two sensory signals and compares them, responding “Interval 1” precisely when  $S_1 > S_2$ . If we represent the pair  $(S_1, S_2)$  as a point in the plot of Fig. 5A then the rule becomes “respond ‘Interval 1’ precisely when  $(S_1, S_2)$  is below the diagonal dashed line shown.”

This model include the assumption that sensitivity  $d'_1$  in the first interval is equal to that in the second  $d'_2$  and this is our second claim:

*Claim 2:*  $d'_1 = d'_2$ .

Another way to describe this computation is that the observer computes as decision variable the difference  $D = S_2 - S_1$  and responds “Second Interval” if  $D > 0$  and otherwise “First Interval”.<sup>2</sup> Interpreted this way, the observer’s task is simply an ordinary equal-variance Gaussian signal detection judgment on the random variable  $D$ .

Let  $d'_{YN}$  be the difference in activity due to the presence of the signal in an interval (Fig. 5A). This notation needs explanation. If the second interval was not presented to the observer and the observer were instructed to perform a Yes–No task on the first interval (judging whether the signal were present or absent in the first interval) then the observer’s measured sensitivity corresponds to  $d'_{YN}$ . Of course in the 2-IFC task the observer has additional infor-

mation from the second interval and, intuitively we expect that the sensitivity measure for the task on the difference signal  $D = S_1 - S_2$  should be greater than  $d'_{YN}$ . The intuition is correct and, for the standard Difference Model,  $d'_{FC} = \sqrt{2}d'_{YN}$  as we derive next.

The variance of  $D = S_2 - S_1$  is 2 since the variance of the difference of two independent random variables is the sum of the variances of the two random variables each of which is 1. When the signal is in the first interval the expected value of  $D$  is  $E[D|1] = -d'_{YN}$  and when it is in the second interval, the expected value of  $D$  is  $E[D|2] = d'_{YN}$ .

The ratio of the signal range to standard deviation is then  $d'_{FC} = \sqrt{2}d'_{YN} = (E[D|2] - E[D|1]) / \sqrt{2}$ . Thus the effective sensitivity measurement  $d'_{FC}$  from the 2-IFC task is  $\sqrt{2}$  greater than the sensitivity  $d'_{YN}$  that would be expected in an ordinary single-interval Yes–No signal detection task using the same stimuli (Fig. 5B).

We previously cited one reason that Macmillan and Creelman (2005) give for recommending the use of 2-IFC (discourages bias). Another reason they give is that “... the predicted  $\sqrt{2}$  difference between Yes–No and 2-AFC permits measurement of sensitivity to smaller stimulus differences than may be practical with Yes–No ... (p. 179)”. The 2-IFC data we re-analyzed in the previous section provide considerable evidence of interval bias in 2-IFC tasks. Thus we are left with the second advantage claimed by Macmillan and Creelman, which is based on the assumption that 2-IFC  $d'_{FC}$  is greater than Yes–No  $d'_{YN}$ .

We can separate two claims concerning the values  $d'_{FC}$ ,  $d'_{YN}$ . The first is that  $d'_{FC} > d'_{YN}$  as presupposed by Macmillan and Creelman. The second is that  $d'_{FC} = \sqrt{2}d'_{YN}$  (which Macmillan and Creelman do not assume). We will see below that the  $\sqrt{2}$  factor is only correct if the observer’s sensitivity is the same in the two intervals of the 2-IFC task. When the observer is more sensitive in one interval than the other we will replace this factor by a factor  $\tau > 1$  to get the modified claim,  $d'_{FC} = \tau d'_1$  where  $d'_1$  is the observer’s sensitivity in the first interval. In summary, we have derived the following claims:

*Claim 3:*  $d'_{FC} = \tau d'_1$

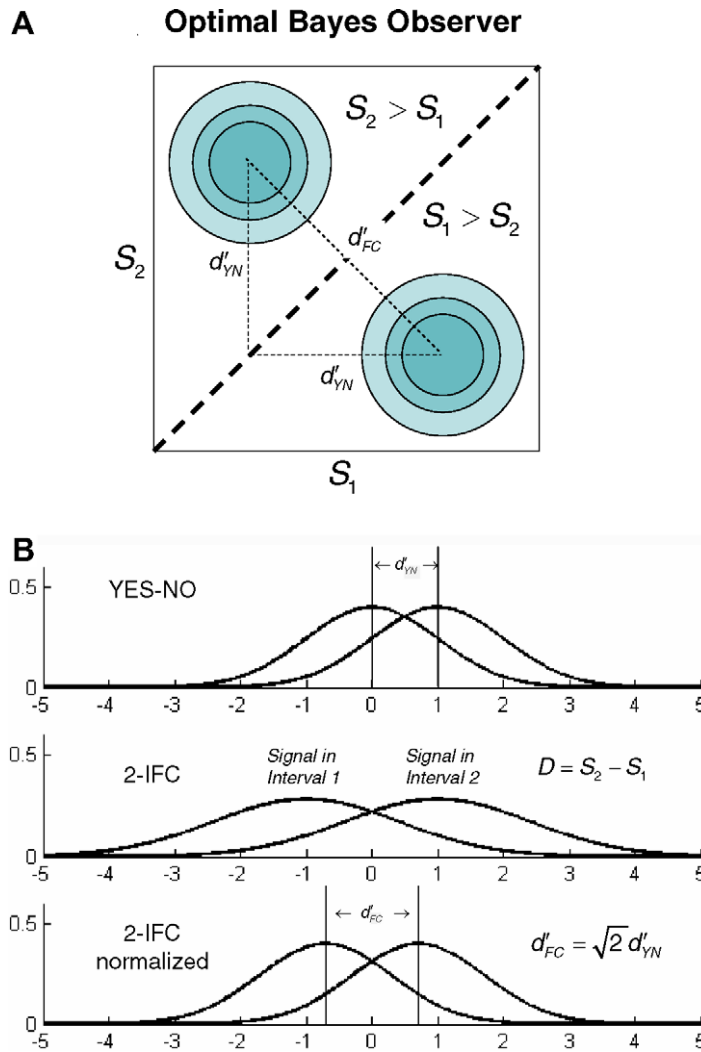
*Claim 4:*  $d'_{FC} > d'_{YN}$

where  $\tau > 1$  can be computed from the observer’s sensitivities  $d'_i, i = 1, 2$  in the two intervals and  $\tau = \sqrt{2}$  when  $d'_1 = d'_2 = d'_{YN}$ . Claim 4 is derived from the Difference Model but it can be considered apart from the Model as the claim that the observer can use the second interval of the 2-IFC procedure to do better in detection. Claim 3 implies Claim 4 but the converse does not hold.

We emphasize that our notation is not standard (although it is close to that of Wickens 2002, p. 100). Many researchers prefer to use  $d'$  to denote only  $d'_{YN}$  and assume that sensitivities measured using other methods such as 2-IFC will be converted to  $d'_{YN}$ , i.e., the sensitivity measured with 2-IFC could be scaled by  $1/\sqrt{2}$  to get a measure of sensitivity comparable to  $d'_{YN}$ . Of course, we cannot make this assumption since the link between 2-IFC and Yes–No measures of sensitivity (Claim 3) is in question.

If we trust the Difference Model as a model for what observers do in 2-IFC experiments, we can readily convert sensitivity measured with 2-IFC tasks to equivalent sensitivity measured with Yes–No tasks and vice versa. Note that, with this model of the 2-IFC task, the ideal observer is never reduced to guessing. A side benefit of the Difference Model is that it can readily account for interval biases. In the standard model of 2-IFC the sensory difference is compared to criterion 0. If instead the observer assumes a criterion of  $c < 0$  then she will be biased to respond in the first interval and will have a higher percent correct on average in that interval. If the observer assumes a criterion of  $c > 0$  then she will have a higher percent correct on average in the second interval.

<sup>2</sup> The case  $S_1 = S_2$  occurs with probability 0 and it does not matter how the observer chooses to respond.



**Fig. 5.** (A) A hypothetical activity distribution for the optimal Bayesian classifier in the standard Difference Model. According to this model, on each trial, the observer records a measure of sensory activity  $S_1$  in the first interval and compares it to a measure of sensory activity  $S_2$  in the second interval, responding “Interval 1” when the former is larger than the latter and vice versa. If we represent this pair of activity measurements as a point in this plot then the optimal decision rule, according to the model, becomes “respond ‘Interval 1’ when this point is below the 45 degree dashed line shown and “respond interval 2” when the point is above the diagonal. The two vicariate Gaussian distributions have standard deviation  $\sigma = 1$ . The vertical and horizontal dashed lines mark  $d'_{YN}$  for the Yes–No signal detection task involving only interval 1 or only interval 2. The dashed line connecting the centers of the two distributions corresponds to the sensitivity of the observer in the 2-IFC task and  $d'_{FC} = \sqrt{2}d'_{YN}$ . (B) The top panel shows the sensory activity  $S_i$  in a single-interval when a signal is presented and when no signal is presented. The distributions are normalized to have standard deviation  $\sigma = 1$  and separation  $d'_{YN}$ . The middle panel shows the distributions of  $D = S_2 - S_1$  when the signal is in the first or second interval of a 2-IFC task. Note that the standard deviations of the distributions are now  $\sqrt{2}$ , the standard deviation of the difference and the separation is  $2d'_{YN}$ . The bottom panel is the middle panel normalized so that the standard deviations of the two distributions are  $\sigma = 1$ . The separation is now  $d'_{FC} = \sqrt{2}d'_{YN}$ .

If we refer to the Difference Model with  $c = 0$  as the standard model of 2-IFC, we could refer to the Difference Model with possibly non-zero choices of criterion as the Difference Model with Bias (Klein, 2001). Note however, that beyond the choice of criterion the observer has no freedom in how he or she executes the Difference Model. By choice of criterion, we can mimic any degree of bias across intervals but, of course, one cannot explain interval biases in performance by means of a model for which the only evidence is the presence of these same interval biases.

If the observed interval biases are just the result of a shift of criterion in the standard Difference Model, then there is a standard correction procedure (Green & Swets, 1973, p. 410; Klein, 2001, p. 1425) that can be used to correct for such biases and recover an accurate estimate of sensitivity. We illustrate this correction procedure in Section 5. However, there is another, darker possibility, noted first by Green and Swets (1973, p. 408): the observed bias may reflect a difference in sensitivity in the two intervals. They and also Macmillan and Creelman (2005, p. 176–177) adopt

this explanation for observed interval biases: failures of Claim 1 are the result of failures of Claim 2. If this were the case, then the experimenter is in the position of trying to measure visual sensitivity using a procedure, 2-IFC, that alters that sensitivity in at least one of the two intervals of presentation. This failure of the model would correspond to a violation of Claim 2.

Klein (2001) has questioned whether 2-IFC procedures correctly measure observer sensitivity. He describes his own experience in one particular 2-IFC task and notes that an identical signal seemed more intense in one interval than in the other. He suggests that this experience reflects a possible bias in this task, and proposes a bias correction procedure. His observation is likely an example of the ‘time error’ reported by Köhler (1923; see also Needham, 1934). Observers in a task similar to 2-IFC tend to favor one interval over the other and this tendency depends on the time between the presentations of stimuli.

Alcalá-Quintana and García-Pérez (2005) carried out an experiment designed to test Klein’s hypothesis but found only that, when

the time between intervals or between trials is too short, the two intervals can interact by masking each other, reducing overall sensitivity to the signal in one interval more than in the other. Were this interaction the only problem for the interpretation of 2-IFC results, it would be easily remedied by increasing the spacing between intervals. However, as demonstrated by the re-analysis of McIntosh et al.'s (1999) data, increasing the ISI between intervals from 500 to 4000 ms did not prevent the emergence of a bias (Fig. 2).

An evident difficulty with 2-IFC designs is that the presentation of the first stimulus may affect the second and vice versa. Unless an experimenter conducts analyses that rule out this possibility, it is possible that the  $d'$  measure that the experimenter seeks to measure is really two  $d'$  measures, one for each interval. Note also that  $d'$  could differ in the two intervals because of a change in noise variance (Wickens, 2002, p. 100ff) and not only because of a change in signal strength (Green & Swets, 1973, p. 408).

There are other possible reasons why Claim 2 might fail. Another evident difficulty for the observer is that, in a 2-IFC experiment, s/he must hold information from the first interval in memory before making a judgment based on information from both intervals (Wickelgren, 1968). This asymmetry could lead to an asymmetry in performance in the two intervals. Nachmias (2006) compared performance in various discrimination tasks and found that performance was superior when the standard stimulus is presented in the first interval rather than the second. He suggested that this asymmetry might imply the use of a memory-based virtual standard even with the 2-IFC paradigm, but it is not clear whether these findings carry any implications for 2-IFC detection tasks.

In the experiment below, we will directly measure the sensitivity of the observer in the two intervals of a 2-IFC task to test Claim 2. Before we report that experiment, we discuss Claims 3 and 4 and review previous relevant work.

### 3.1. Claims 3 and 4

As noted above, for the Difference Model of Fig. 5, the observer's performance can be represented by a measured  $d'_{FC}$  based on data and standard signal detection analyses. As derived above, this measured  $d'_{FC}$  is a random variable that is an estimate not of the  $d'_{YN}$  value associated with a single-interval Yes–No task but of  $\sqrt{2}d'_{YN}$ . When there are distinct  $d'$  values in the two intervals, denoted  $d'_i, i = 1, 2$ , we need to modify Claim 3:  $d'_{FC} = \sqrt{(d'_1)^2 + (d'_2)^2} = d'_1 \sqrt{1 + \rho^2}$  where  $\rho = d'_2/d'_1$  (Wickens, 2002, p. 100ff).

When Claim 2 holds,  $\rho = 1$  and the  $\sqrt{2}$  rule described above results. Previous studies have compared performance in 2-IFC tasks with performance in Yes–No tasks with identical stimuli. The pattern of results in many of these studies did not support the  $\sqrt{2}$  rule. Some studies (e.g., Creelman & Macmillan, 1979; Jesteadt & Bilger, 1974; Leshowitz, 1969; Markowitz & Swets, 1967; Pynn, Braida, & Durlach, 1972; Schulman & Mitchell, 1966; Swets & Green, 1961; Viemeister, 1970; Watson, Kellogg, Kawanishi, & Lucas, 1973) found  $d'_{FC}/d'_{YN}$  ratios that were larger than  $\sqrt{2}$ . For instance, Jesteadt and Bilger (1974) compared frequency discrimination and intensity discrimination with both 2-IFC and Yes–No tasks and found that for both frequency and intensity discrimination performance in the 2-IFC task is better than performance in the Yes–No task by a factor that is considerably larger than  $\sqrt{2}$  (2.1 and 2.13, respectively). Similarly, Viemeister (1970) found for intensity discrimination a  $d'_{FC}/d'_{YN}$  ratio of 1.91, and Creelman and Macmillan (1979) found for both frequency and phase discrimination that the ratio  $d'_{FC}/d'_{YN}$  for 2-IFC and Yes–No paradigms was about 2 rather than  $\sqrt{2}$ . Large  $d'_{FC}/d'_{YN}$  ratio – 1.75 – was also found for the detection of a brief 1000 Hz sinusoid in the presence of noise

plus pedestal (Leshowitz, 1969). Other studies (e.g., Leshowitz, 1969; Markowitz & Swets, 1967; Swets & Green, 1961) have found  $d'_{FC}/d'_{YN}$  ratios that were smaller than  $\sqrt{2}$ . Markowitz and Swets (1967) compared performance in auditory detection with the two paradigms and found a mean ratio of 1.15, with one of the observers being more sensitive in the one-interval experiment. Similar results were found by Leshowitz (1969) when the task involved simple detection of sinusoid added to a noise that was either continuously present or presented only during the observation interval (ratios of 1.19 and 1.35, respectively). As in Markowitz and Swets (1967) study, one of Leshowitz' three observers actually performed better in the one interval than the two intervals condition. Finally, Schulman and Mitchell (1966) also examined auditory signal detection with both paradigms, and found an averaged  $d'_{FC}/d'_{YN}$  ratio of 1.46, which is close to  $\sqrt{2}$ , but the ratios of individual observers revealed considerable deviations from  $\sqrt{2}$ , ranging from 1.17 to 1.77.

Many of the studies just cited lack estimates of the standard error of the ratios of  $d'$  values obtained in different tasks, making it difficult to assess the status of the Difference Model given these results. We note, for example that values of  $\tau = \sqrt{1 + \rho^2} < 1$  as reported for one observer each in the studies of Leshowitz (1969) and Markowitz and Swets (1967) are inconsistent with the Difference Model. However, we cannot determine from the published work that the reported estimates of  $\tau$  were significantly less than 1.

If we accept that Claim 2 is false then any measured value of  $\tau > 1$  can be explained as the result of a hypothetical difference in sensitivity in the two intervals:  $d'_1 \neq d'_2$ . Such results do not challenge Claim 3, only Claim 2.

Based on the literature just summarized, there is little reason to accept the Difference Model or any of the conclusions based on it – and little reason to reject it. If the experimenter does not run a control experiment comparing  $d'_{FC}$  and  $d'_{YN}$ , there is little reason to assume that sensitivity measured by 2-IFC can be used to estimate the sensitivity that would be measured with the same stimuli in a Yes–No task. We conclude that, given only the results  $d'_{FC}$  of a 2-IFC experiment, the experimenter knows little about  $d'_{YN}$ .

Moreover, the literature just summarized included experiments with many different kinds of stimuli. It is plausible that the observed differences in estimates of  $\tau$  are in part due to the kinds of stimuli employed. Based on these results, there is little reason to think that the experimenter can generalize experience with one sort of stimulus to another. We return to this point in Section 5.

In the next section, we report an experiment designed to directly investigate how and whether observers guess and to test the Difference Model. We do not compare a Yes–No task with a 2-IFC task based on the same stimuli. Instead we compare a 2-IFC task with a task in which the observer is simply required to carry out two Yes–No tasks with the same timing as the 2-IFC task. In particular, the target stimulus could appear in either, both or neither interval. With this design we allow for the possibility that sensitivity differs in the two intervals and measure  $d'_i, i = 1, 2$  separately for each interval under conditions that duplicate the conditions of the 2-IFC. With this design, we can estimate  $d'_1, d'_2$  directly and test Claim 2. With knowledge of  $d'_1, d'_2$  we estimate  $\tau$  and test Claim 3.

## 4. Experiment

Twenty-two observers participated in an experiment that included two sessions. The 2-way session employed a typical 2-IFC paradigm. Each trial included two temporal intervals and the target appeared in one of the intervals. The 4-way session employed a novel adaptation of that paradigm. A single trial also included two intervals, but in the 4-way condition the target could appear

in one of the intervals, in both intervals, or in neither of them. The four possible outcomes were equally likely. We adopt the convention that the four possible kinds of trials in the 4-way condition are labeled YN, NY, NN, YY (i.e., the target appears in: 1st interval, 2nd interval, neither interval, or both intervals), to emphasize that the 4-way task is simply two successive Yes–No signal detection tasks. On trials in the 4-way where YN or NY is presented (approximately half of the trials), the observer's sensory state should be equivalent to that of corresponding trials in the 2-way task. However, in the 4-way task, the states nn – the sensory activity in both intervals is below the criterion – and yy – the sensory activity in both intervals is above the criterion – are now legitimate possible outcomes of a trial and the observer is permitted to respond “nn” and “yy”. The 4-way task allows us to test the standard model and also to learn something about the observer's decision process on those trials (YN and NY) where sensory events are identical to those in the 2-way task.

To our knowledge this experiment is the first to combine measurement of 2-IFC performance with separate measurement of sensitivity in the two intervals of the task with the same timing and design. Had we simply compared 2-IFC performance to Yes–No performance, we could not be sure that the presence of the second interval had not altered measured 2-IFC (violating Claim 2) as several researchers have suggested (reviewed above).

In Fig. 6, we show data sets for one of our observers to illustrate the design and its analysis. We have labeled the 2-way trials as YN and NY by analogy to the 4-way, and the observer's responses as yn and ny. In the actual experiment, observers responded 1, 2, 3, 4 for yn, ny, nn, yy, respectively, in the 4-way condition and, in the 2-way condition, could only respond 1 (yn) or 2 (ny). Note that the first two rows of the 4-way table are the trials that could occur in the 2-way table as well, except now the observer can legitimately report that he or she is in state nn or yy.

We wish to examine estimated  $d'_{FC}$  in the 2-way session and  $d'_i, i = 1, 2$  in the 4-way session to test Claim 2 and Claim 3. We describe how we fit 2-way via the Difference Model in Appendix B.2. We describe how we fit 4-way data in Appendix B.3, also by an application of the optimal Bayesian classifier.

A 2-way design				B 4-way design					
	Response			Response					
	yn	ny		yn	ny	nn	yy		
YN	168	36	204	YN	76	1	12	13	102
NY	44	160	204	NY	4	80	6	12	102
				NN	19	10	66	6	101
				YY	11	17	2	72	102

Results for a single subject (S02).

**Fig. 6.** Design and sample data of: (A) The 2-way condition, employing a typical 2-IFC paradigm. In this condition two kinds of trials were possible: YN – in which the target was present in the first interval, and NY – in which the target was present in the second interval. Accordingly, two kinds of response were possible: yn and ny corresponding to the YN and NY trials. (B) The 4-way condition employing a novel adaptation of the 2-IFC paradigm, in which four kinds of trials were possible: YN and NY which are identical to the 2-way condition, and NN – in which the target was present in both intervals, and YY – in which the target was present in neither of the intervals. Accordingly, four kinds of response were possible: yn, ny, yy, and nn, corresponding to the YN, NY, YY, and NN trials. The response distribution across the different kinds of trials is reported for one of our observers (S02) to illustrate the design.

#### 4.1. Methods

##### 4.1.1. Observers

Twenty-two naïve observers at the University of Haifa participated in the experiment; all had normal or corrected to normal vision. Each completed the 2-way and 4-way conditions in separate sessions. For two of the observers we could not compute a stable estimate of  $d'$  in either the 2-way or 4-way condition because they made few or no errors in one or the other task. We excluded these observers from further analysis.

##### 4.1.2. Stimuli

The target was a single 8 cpd Gabor patch with 30° orientation and 10% contrast presented at the center of the screen. The standard deviation of the Gaussian window of the Gabor was 2 cycles of the windowed grating. In the 2-way condition the target was present in the first or second interval (YN, NY, respectively); each occurred at random on 50% of the trials. In the 4-way condition there were four types of trials, each at random occurring on 25% of the trials: YN – the target was present only in the first interval; NY – the target was present only in the second interval; YY – the target was present in both intervals; NN – the target was present in none of the intervals.

##### 4.1.3. Procedure

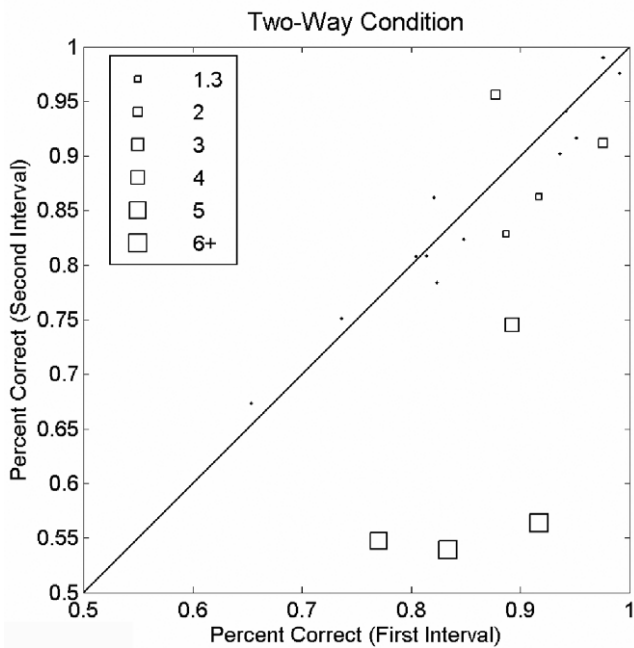
Each temporal interval began with a fixation dot presented for 600-ms at the center of the screen. On intervals that contained the target, a Gabor patch was then briefly presented at the center. The duration of the target presentation was set individually to keep performance level at approximately 85% correct. It varied between 15 and 105 ms (mean = 46 ms and SD = 26 ms). On intervals without a target, the screen was blank for the corresponding duration. An additional 300-ms blank screen served as the ISI between intervals. In the 2-way condition, the observers were asked to hit the '1' key for a YN trial and the '2' key for a NY trial. In the 4-way condition observers were asked to hit the '1' key for a YN trial; the '2' key for a NY trial; the '3' key for a NN trial; and the '4' key for a YY trial. The order of the different trial types was randomized within a condition. Each observer participated in 816 trials (408 per condition). All observers performed the 2-way condition first to ensure that whatever strategy they employed during the 4-way condition would not contaminate their performance in the 2-way condition.

#### 4.2. Analysis and results

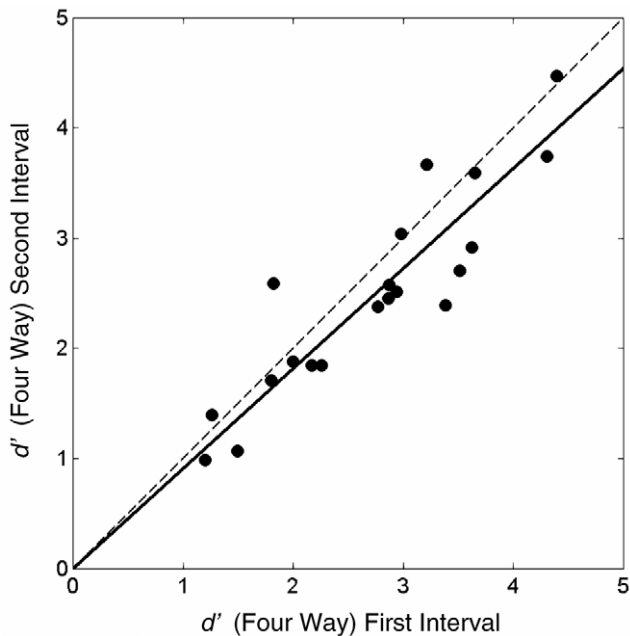
We first compared  $p_1$  (proportion-correct, first interval) and  $p_2$  (proportion-correct, second interval), just as we did for the data sets in the first section, to evaluate whether there were marked interval biases for any observers. Over a third of the observers (8/22) had marked 2-way biases, seven favoring the first interval. These data are presented in Fig. 7, in the same format as Fig. 1.

Next, we estimated the  $d'_i, i = 1, 2$  values and criterion values in the two intervals from the 4-way by maximum likelihood methods as explained in Appendix B.3. In Fig. 8, we plot  $d'_2$  vs.  $d'_1$ . A least-square regression of  $d'_2$  on  $d'_1$  (without intercept) results in a slope of 0.908 which is significantly different from 1 ( $p < .05$ ). The 95%-confidence interval for the slope is (0.844, 0.973). The observers are, overall, slightly more sensitive in the first interval than the second. This outcome is the opposite of what Klein (2001) reported based on his own experience (discussed above). We reject Claim 2 but note that the difference in  $d'_1 > d'_2$  is only about 9%.

Were the 2-way and 4-way analyses consistent with the Difference Model, we would expect the  $d'_{FC}$  value for the 2-way condition to be greater than that for the intervals in the 4-way condition (Claims 3 and 4). Because the  $d'_2$  values (second interval) are about 9% less than the  $d'_1$  values (first interval), this factor will not be  $\sqrt{2}$



**Fig. 7.** A Plot of Performance in the First Interval vs. the Second for the 2-way condition. The distance of each data point from the diagonal line is an index of the difference between Percent Correct in the second interval and Percent Correct in the first interval ( $p_2 - p_1$ ). The radius of the symbol plotted at each point represents the  $p$ -value of the hypothesis tests [ $H_0 : p_1 = p_2$ ;  $H_1 : p_1 \neq p_2$ ] when it is in the range 0.05–0.000001 (i.e., when  $-\log_{10}p$  is in the range 1.3–6 as shown in the legend). When  $p > .05$  the point is plotted as a single dot. Again, significant interval biases were found: over a third of the observers (8/22) had marked 2-way biases.



**Fig. 8.** A Plot of  $d'$  in the Second Interval vs. the First for the 4-way condition. A least-square regression of  $d'$  in Interval 2 on  $d'$  in Interval 1 (without intercept) results in a slope of 0.908 (solid line) which is significantly different from 1 ( $p < .05$ ). The 95%-confidence interval for the slope is (0.844, 0.973). The observers are, overall, slightly more sensitive in the first interval than the second.

but instead  $\tau = \sqrt{1 + \hat{\rho}^2}$  where  $\hat{\rho} = 0.908$  is the ratio of  $d'$  values just estimated (Wickens, 2002, p. 100). We derived this rule earlier in discussing the optimal Bayesian observer. Thus,  $\hat{\tau} = 1.35$ , slightly less than  $\sqrt{2}$ .

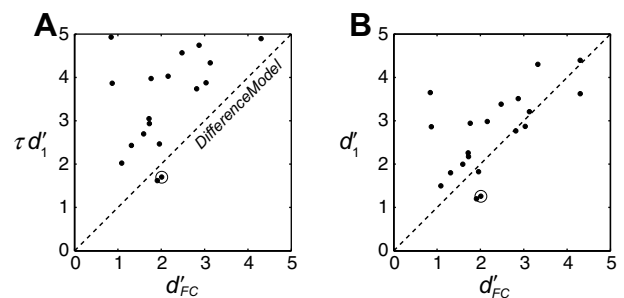
We plot the mean of the estimated  $d'_1$  for the 4-way condition vs. the estimated 2-way  $d'_{FC}$  in two formats in Fig. 9. First, we plot estimates of  $\tau d'_1$  vs. estimates of  $d'_{FC}$  in Fig. 9A. If Claim 3 is correct and  $d'_{FC} = \tau d'_1$  then we expect to find that the plotted points fall symmetrically about the dashed diagonal line. Examination of the scatter plot indicates that this is not the case. We regressed the estimates of  $\tau d'_1$  vs. estimates of  $d'_{FC}$  (without intercept term). The estimate of regression slope is 1.49 with 95%-confidence interval (1.25, 1.74). We reject Claim 3 and the Difference Model.

Next, we plot estimates of  $d'_1$  (without the correction factor  $\tau$ ) vs. estimates of  $d'_{FC}$  in order to test Claim 4 (Fig. 9B). We regressed the estimates of  $d'_1$  vs. estimates of  $d'_{FC}$  (without intercept term). The estimate of regression slope is 1.10 with 95%-confidence interval (0.92, 1.29). The slope is not significantly different from 1 at the 0.05 level. We find no support for Claim 4:  $d'_{FC} > d'_{VN}$ .

If we had not run the 4-way condition, then we could interpret the fitted slope for Fig. 9B as an estimate of  $\tau$  and we could solve for  $\rho = d'_2/d'_1 = 0.47$ . We would conclude that Claim 3 could hold if the observer is remarkably insensitive in the second interval relative to the first:  $d'_2 = 0.47d'_1$ . But we did measure  $\tau$  separately in the 4-way condition, it is not 0.47 but rather 0.908, and we reject this possibility and Claim 3. These results are inconsistent with the Difference Model for 2-IFC data even if we allow for the possibility of differential sensitivity in the two intervals (failures of Claim 2).

Finally, to ensure that the lack of significant differences between the  $d'$  values of the 2-way and 4-way conditions was not due to practice effects, we compared performance in the first and second halves of the 2-way condition. If performance could still be improved with practice when the observers finished the 2-way condition and started the 4-way condition there should be evidence of this improvement when we compare performance in the first half of the 2-way condition to that in the second half. We found no significant differences in performance between the two halves ( $F < 1$ ). In fact, percent correct in the two halves was almost identical (1st half: 84.9%; 2nd half: 84.3%).

To summarize the findings of this experiment: (a) considerable interval biases were found with the 2-way condition (corresponding to the typical 2-IFC task). These findings are consistent with previous reports of biases and with the data sets re-analyzed above, and together they question the claim that the 2-IFC paradigm discouraged biases (Claim 1); (b) no support was found for



**Fig. 9.** (A) The mean  $d'$  estimates for the 4-way condition multiplied by  $\tau$  vs. the corresponding  $d'$  estimates for the 2-way condition for each observer. The multiplier  $\tau$  is based on the actual measured sensitivity of observers (Fig. 8 and text) and would be  $\sqrt{2} = 1.414 \dots$  only if the observers' sensitivity in the two intervals were the same. Here, it is slightly smaller,  $\tau = 1.35$ . If the Difference Model accurately described these observers, we expect the points to be distributed around the diagonal dashed line  $d'_{FC} = \tau d'_{VN}$ . They are not and we reject the Difference Model. The failure of the Difference Model cannot be explained by a hypothetical difference in sensitivity in the two intervals since we measured sensitivity in the two intervals and factored it into the prediction. The data point corresponding to the observer in Fig. 6 is circled. (B) The same data is re-plotted without the multiplier  $\tau$ . The diagonal dashed line now corresponds to the prediction that there is no benefit whatsoever from taking a difference of sensory signals and  $d'_{FC} = d'_{VN}$ . The data point corresponding to the observer in Fig. 6 is circled.

the  $\sqrt{2}$  rule or its modification to allow for differential sensitivity in the two intervals (Claim 3), nor was there any support for the claim that  $d'_{FC} > d'_{YN}$  (Claim 4). We also rejected Claim 2 ( $d'_1 = d'_2$ ) but the difference is estimated to be only 9%.

### 5. Discussion

We considered four claims concerning performance in 2-IFC tasks. The first claim was that 2-IFC performance is unbiased: the probability of a correct response when the stimulus is presented in the first interval  $p_1$  is equal to the probability of a correct response when the stimulus is presented in the second interval  $p_2$ :

*Claim 1:*  $p_1 = p_2$

The term “interval bias” is used to describe a failure of this claim.

We first reported the results of re-analyses of data sets collected in seventeen experiments from three laboratories using 2-IFC methods with very different stimuli. We found large interval biases for all data sets. The presence or absence of these biases did not depend in any obvious way on: (a) the type of experiment or the complexity level of the display; (b) whether or not spatial attention was focused on the target location; (c) the duration of the ISI between the two presentations; or (d) whether or not the observers were experienced. Thus, this finding of an asymmetrical performance shows no obvious patterned dependence on factors such as attentional state, complexity, ISI or practice. Given that biased observers in Wolfson and Landy (1995), (1998) tended to favor the same interval, it is unlikely that the observed biases are due to idiosyncratic preferences for one interval over the other that lead to shifts of criterion in the Difference Model (Wickens, 2002, p. 99). The biases seem to be evoked by the stimulus conditions in the experiment.

We then presented the standard Difference Model of 2-IFC and emphasized that interval biases in themselves do not present a challenge to the Model. The standard Difference Model is equivalent to a Yes–No signal detection task on the difference between the sensory signals in the two intervals and any interval bias could be explained as a bias in choice of criterion. However, as noted by Green and Swets (1973, p. 408) it could also be due to a difference in sensitivity in the two intervals, bringing us to the second claim considered: that measured sensitivity in the first interval  $d'_1$  is equal to that in the second  $d'_2$ .

*Claim 2:*  $d'_1 = d'_2$

A failure of this claim implies that the 2-IFC procedure has altered the sensitivity in at least one of the two intervals, raising the question of what it is that 2-IFC is measuring. Of course, given only 2-IFC data there is no direct way to test Claim 2 as any observed bias may be due to a preference for one interval over the other or to a failure of Claim 2.

The last two claims concern the relation of  $d'_{FC}$ , sensitivity as measured using the 2-IFC procedure, to  $d'_{YN}$ , the sensitivity measured using a Yes–No procedure. In deriving the relation between them, we allowed for the possibility that Claim 2 is false and derived  $d'_{FC} = d'_1 \sqrt{1 + \rho^2}$  where  $\rho = d'_2/d'_1$  (Wickens, 2002, p. 100ff). For convenience we set  $\tau = \sqrt{1 + \rho^2}$  giving us

*Claim 3:*  $d'_{FC} = \tau d'_1$ .

*Claim 4:*  $d'_{FC} > d'_1$

When,  $d'_1 = d'_2$ , and  $d'_1 = d'_2 = d'_{YN}$  (the 4-way task is two Yes–No tasks), then Claim 3 becomes the familiar  $d'_{FC} = \sqrt{2} d'_{YN}$ .

We next reviewed the many comparisons on the literature of sensitivity measured using 2-IFC and other methods. As we noted there, these studies typically report summary results without any tests of significance. Many of these studies assumed that the difference signal is square root of 2 greater than would be found in the corresponding Yes–No task but, as we showed, this would only be the case if the sensitivity to the signal in the two intervals were equal  $d'_1 = d'_2$  and further  $d'_1 = d'_2 = d'_{YN}$ . If it were not, then other factors  $\tau$  are possible. The entire literature that we reviewed is subject to the criticism that the results are consistent with a slight extension of the Difference Model that allows for different sensitivity in the two intervals. It is difficult to see how one can evaluate the Difference Model when one does not separately measure the sensitivity of the observer in each interval of the 2-IFC task. Consequently, there is little to conclude from this body of work.

Last, we reported an experiment in which we paired a 2-IFC (“2-way”) task with a 4-way task comprising two independent Yes–No tasks with the same stimuli and timing as the 2-way task. On trials in the 4-way task where the correct responses were Yes–No or No–Yes, the observer experienced the same sensory events with the same timing as in the 2-way task. We could observe performance in the 2-way task while measuring  $d'$  separately in the two intervals with the 4-way task. We found marked biases for 8/22 observers and a small but significant difference in  $d'$  of about 9% in the two intervals, contradicting Claims 1 and 2. Given the measured  $d'_i, i = 1, 2$  we can estimate  $\tau$  and test Claim 3. We found that Claim 3 failed and that  $d'_{FC} < \tau d'_{YN}$ , moreover, we could not reject the claim that  $d'_{FC} = d'_{YN}$ . We described how the Difference Model could be modified to allow for failures of Claim 1 and Claim 2. We do not see how it can be modified to allow for the failure of Claim 3. The observer seems to do little better with 2-IFC than with just one interval in a Yes–No task.

To our knowledge, we are the first to measure sensitivity in both intervals in our 4-way task and therefore we are the first who are able to compare performance with the predictions of the Difference Model allowing for different sensitivities in the two intervals. The comparison of sensitivity across the two experiments was inconsistent with the predictions of the Difference Model as seen in Fig. 9A and the accompanying discussion. We emphasize that simply measuring  $d'_{FC}$  and  $d'_{YN}$  in two separate experiments using the 2-IFC procedure and the Yes–No procedure could not have led to the conclusions we reached. Since Claim 2 is under test we must allow for the possibility that the 2-IFC procedure itself leads to different sensitivities in the two intervals of a 2-IFC procedure and we must measure  $d'_{YN}$  in both intervals with the same timing as in the 2-IFC task.

#### 5.1. What is going on?

As mentioned in the introduction, there are two common explanations for interval biases in the literature. The first possible explanation is that the two stimuli in a 2-IFC trials interact, that the effective  $d'$  in the two intervals differ, and that observers favor the interval with higher  $d'$  (Macmillan & Creelman, 2005, pp. 176–177; Green & Swets (1973, p/ 408)). Our results in the 2-way vs. 4-way experiment support this claim. Seven out of eight biased observers in the 2-way (2-IFC) task favored the first interval; and in the 4-way task we found that the  $d'$  for the first interval was roughly 9% higher than that of the second. However, if we accepted this explanation of interval bias, there is an evident conceptual problem in the use of 2-IFC: it is unclear whether  $d'_i, i = 1, 2$  for the first or for the second interval is the sensitivity we want to measure. When the difference is as small as 9%, the experimenter may choose to ignore the difference in  $d'_i, i = 1, 2$ . However, we have no reason to think that that the differences in  $d'_i, i = 1, 2$  for the data sets re-analyzed in the first part of the article are just 9%.

A second possible explanation of interval imbalances is that they are due to memory limitations: the observer either cannot or does not record an accurate sensory intensity in the first interval for later comparison with that of the second interval (Wickelgren, 1968). The bias is due to the guessing rule. But then the simplicity of the Difference Model is replaced by an unspecified model where the observer can retain only a limited portion of sensory information from the first interval to compare to that of the second. Given this explanation, we simply do not know what the observer is doing during a 2-IFC trial and do not know how to interpret observed biases or the resulting measures of sensitivity.

The two explanations (interval interaction and memory limitations) are not mutually exclusive and both interval interactions and memory limitations may affect 2-IFC performance. There may be other factors that could lead to interval biases or deviations from the  $d'_{FC} = \tau d'_{YN}$  rule. Wickelgren (1968, p. 116) mentions correlation between noise activity in the two intervals. If, for example, the additive noise processes in the two intervals had correlation 1, then the difference between sensory activity in the intervals would be the signal or its negative uncontaminated by noise. Whatever  $d'_{YN}$  might be in a Yes–No task with these stimuli, it is infinity in the corresponding 2-IFC task with perfectly correlated noise. If the additive noise processes in the two intervals had correlation  $-1$ , it is easy to show that the  $d'$  resulting from an application of the Difference Model will be  $\sqrt{2}$  less than that measured in the corresponding Yes–No task consistent with the performances of single anomalous observers in Markowitz and Swets (1967) and Leshowitz (1969) that we discussed above. Correlation would not lead to interval bias but it would lead to 2-IFC measures that were not  $\sqrt{2}$  times greater than that for the corresponding Yes–No task. Wickelgren also points out the possible effects of attention. Our re-analysis of the Carrasco and Yeshurun (1998) study in which attention was explicitly manipulated does not support this idea. One last possibility is the possibility that observers just happen to prefer one response to the other (Wickens, 2002, p. 99), although this possibility is not consistent with the finding (noted above) that the biased observers in Wolfson and Landy (1995), (1998) tended to favor the same interval.

The results of the experiment (Fig. 9A and B) are inconsistent with the standard Difference Model but consistent with the claim that observers in the 2-way (2-IFC) task are classifying sensory events in the two intervals just as they do in the 4-way task. If this were the case then we expect that there would be no  $\sqrt{2}$ -like benefit from taking a difference of sensory signals, consistent with the outcome of the experiments. Observers simply treat the 2-IFC task as two signal detection tasks just as the 4-way observer is asked to do.

The observer in the 2-way task could readily translate the outcomes  $yn$  and  $ny$  into judgments that the stimulus occurred in the first or second interval respectively. However, he or she would have to transform  $yy$  and  $nn$  states into either a  $yn$  or  $ny$  response via a guessing rule since he or she knows that  $YY$  or  $NN$  were not presented in 2-IFC tasks and  $yy$  and  $nn$  are not valid responses.

The results of the 4-way data can help us clarify this point. As an example we use the data set presented in Fig. 6. Specifically, we use the  $YN$  and  $NY$  trials of the 4-way data set (the first two rows in the 4-way table of Fig. 6B), and apply to them different guessing rules as if this is the data set of an observer in a 2-way experiment who has to convert the  $yy$  and  $nn$  outcomes into a  $yn$  or a  $ny$  response. For instance, applying a ' $p_{yy} = 1, p_{nn} = 1$ ' guessing rule (i.e., always respond  $yn$  when forced to guess) to the data set in Fig. 6B will result in the following distribution of  $yn$  and  $ny$  responses: 101  $yn$  responses and 1  $ny$  response in the  $YN$  trials and 22  $yn$  responses and 80  $ny$  responses in the  $NY$  trials.

Klein (2001) and Green and Swets (1973, p. 408ff) suggest compensating for the interval bias by averaging the z-scores

rather than probabilities:  $z = (z_1 + z_2)/2$ ; where  $z_1$  is the z-score of percent correct in the first interval and  $z_2$  is the z-score of percent correct in the second interval. Applying this bias correction to the above distribution of responses results in an estimate of  $d'_{FC} = 2.21$ . Repeating this procedure with a ' $p_{yy} = 0, p_{nn} = 0$ ' guessing rule (i.e., always respond  $ny$  when forced to guess) results in  $d'_{FC} = 1.71$ . Similarly, a ' $p_{yy} = 0, p_{nn} = 1$ ' guessing rule (i.e., trust the outcome of the most recent interval) results in  $d'_{FC} = 1.69$ , and a ' $p_{yy} = 1, p_{nn} = 0$ ' guessing rule (i.e., trust the outcome of the first interval) results in  $d'_{FC} = 1.52$ . Finally a 'fair' guessing rule – ' $p_{yy} = 0.5, p_{nn} = 0.5$ ' (i.e., guess  $yn$  or  $ny$  with equal probability) results in  $d'_{FC} = 1.59$ . Thus, even with the bias correction, the same 'sensory states' result in different estimates of  $d'_{FC}$  depending on the guessing rule. The guessing rule used by an observer is typically unknown to the experimenter.

## 5.2. What is to be done?

Our results provide compelling evidence that 2-IFC tasks are not simple, they are not 'bias free', and they are potentially difficult to interpret. We therefore recommend a degree of caution in using 2-IFC methods. At a minimum the experimenter should separately record data for the first and second intervals and analyze this data for evidence of interval biases. If there are large biases, then it is unclear what the experimenter can do about them. The best advice is likely that of Green and Swets: "Of course, the assumption of symmetry [interval bias] can be checked directly, since one can determine whether the subject has in fact detected more signals in one interval than in another. Therefore, one might choose to test first for symmetry, or to select data where the asymmetry is slight and use only such data to check further predictions ...". (Green & Swets, 1973, p. 45).

Even if we knew that the biases resulted from sensitivity differences in the two intervals, the problem cannot be resolved by a bias correction procedure such as that suggested by Klein (2001) since it is based on the Difference Model and Claim 2. However, for many applications, the experimenter may only wish to examine how sensitivity varies as function of experimental condition and is not concerned with what the sensitivity measure is measuring so long as it is comparable across conditions. The use of 2-IFC measures and Klein's correction would be justified in such an application.

The failure of Claim 3 implies that we know of no way to predict performance in a Yes–No task from 2-IFC performance for arbitrary choice of stimuli. If the experimenter's goals require comparison of sensitivity measured with different psychophysical measures then he or she should run the appropriate control experiment to determine the relation between sensitivities measured with different procedures.

Alternatively, the experimenter might consider use of a 2-AFC procedure where stimuli are presented simultaneously but at different spatial locations. It is possible that simultaneous presentation will get around some of the problems with 2-IFC.

A second alternative is to use a Yes–No task rather than 2-IFC. This choice minimizes the difficulties just mentioned. Interval interactions (including correlations in noise) are still possible but now only between trials and any memory burden is plausibly reduced. Arbitrary preferences for one or the other response may occur but they are readily detected in the data. Yes–No is not the only alternative: there are other promising psychophysical procedures based on explicit or implicit models of what observers are doing such as virtual standard models (see, for example, Morgan, Watamaniuk, & McKee, 2000; Nachmias, 2006).

Of course, the Yes–No paradigm has its own faults. The best-known drawback of this paradigm is that the threshold obtained with a Yes–No task may be contaminated by the observer's

decisional criterion (e.g., Green & Swets, 1973; Jäkel & Wichmann, 2006; Kaernbach, 2001; Klein, 2001; Miller & Ulrich, 2001). Specifically, with the Yes–No task a reliable estimation of threshold depends heavily on the ability of the observers to maintain a stable criterion – fluctuations of the decisional criterion can lead to fluctuations in the threshold estimates (Kaernbach, 2001).

A related drawback is that there are hardly any adaptive methods available for a Yes–No task in which the false alarm rate is measured so that  $d'_{YN}$  can be calculated (Klein, 2001). Kaernbach (1990) describes an unbiased adaptive procedure for a Yes–No task in which an equal number of blanks and signal trials are intermixed. Thus, the Yes–No task may serve as an alternative for the 2-IFC task, but one has to be aware of decisional biases and accept the fact that the collection of unbiased adaptive methods for this task is limited.

Based on the results presented here, there is little reason to think that 2-IFC tasks or any psychophysical tasks will have the same pattern of success or failure for all possible choices of stimuli. The reader may believe that with different stimuli or timing, the outcome of our experiment could have been very different. Perhaps we would have found no interval bias (Claim 1), no differences in sensitivity (Claim 2) and we might have found that  $d'_{FC} = \sqrt{2}d'_{YN}$  (Claims 3 and 4). We agree. We do not claim that our results with 2-IFC and a particular choice of experimental conditions will generalize to other experimental conditions. We saw even in Fig. 1 that two slightly different experimental designs (Wolfson & Landy, 1995, 1998) produced opposite patterns of bias. We are not in a position to generalize how 2-IFC or any psychophysical procedure will behave in novel applications and we cannot count on any of the claims being true or false in a novel application of 2-IFC. If an experimenter wishes to claim that any of the four claims is valid in a particular application, the burden is on him or her to demonstrate that it is, experimentally. Past rules of thumb based on shared opinion, standard texts, untested models or imperfect experimental tests are no reliable guide to what observers actually do in 2-IFC experiments.

*A comment on notation.* When we first defined  $d'_{FC}$  we noted that our definition is not the standard definition used by many researchers. The standard definition is our measure divided by  $\sqrt{2}$  and it is evidently based on Claim 3 with  $\tau = \sqrt{2}$ . If Claim 3 were correct with  $\tau = \sqrt{2}$  then the standard definition of  $d'_{FC}$  would be equal to  $d'_{YN}$ . We chose to use a non-standard definition of  $d'_{FC}$  precisely to avoid an implicit assumption that we would later reject. However, the standard definition is based on a second implicit assumption, the assumption that there is a single measure of sensitivity that is independent of the psychophysical method used to measure it. The standard definition of  $d'_{FC}$  effectively presupposes that this single measure is  $d'_{YN}$  which is estimated from 2-IFC data by normalization by a  $\sqrt{2}$ . What we found in the experiment reported, was that the temporal structure of the 2-IFC task altered sensitivity in one or both intervals and that measured sensitivity is not independent of the psychophysical method used to measure it. Accordingly, we propose that our definition of  $d'_{FC}$  be adopted by other researchers since it does not presuppose questionable assumptions and it makes explicit that sensitivity was measured using a particular psychophysical method.

We emphasize that we are not challenging the value of the use of optimal statistical observers as a standard for analysis in psychophysical experiments. We argue only that we do not currently know how to model what observers actually do in 2-IFC tasks and that we have no reason to think that models appropriate to one choice of stimuli can be generalized to others. In particular our own experiment, and those of previous studies reviewed above cast considerable doubt regarding the validity of the standard Difference Model and its predictions. We regard the development and

testing of models of what observers actually do in 2-IFC tasks as a promising area for research.

## Acknowledgments

Supported by Grants EY016200 (M.C.) and EY08266 (L.T.M.) from the National Institutes of Health. We thank Sabina Wolfson and Cigdem Penpeci for sharing data sets. We thank Ron Kinchla for helpful discussions, and Stuart Fuller, Stan Klein, Samuel Ling, Taosheng Liu, Barbara Montagna and Michael Morgan for helpful comments on a draft of this manuscript.

## Appendix A

### A.1. Texture discrimination (Wolfson & Landy, 1998)

This data set was collected with texture stimuli composed of randomly placed, short line segments. Each stimulus was composed of two textures that either abutted to form an edge or were separated by a blank region. The lines' orientation was chosen randomly using Gaussian distributions. In one temporal interval both textures were created based on the same distribution (i.e., same mean and standard deviation). In the other temporal interval the two textures were created using different distributions. In this case, the distributions of the two textures differed either in their mean orientation or standard deviation. Six experienced observers had to indicate which temporal interval included different textures. The results for this study are summarized in Fig. 1A.

### A.2. Texture segmentation (Wolfson & Landy, 1995)

This study explored observers' ability to discriminate between two textures, one with a straight texture edge and one with a "wavy" texture edge. Each temporal interval included a circular texture image composed of randomly placed, oriented line segments. On each side of the edge all line segments shared a common orientation. This difference in texel orientation produced an illusory or orientation-defined edge that was straight in one interval and wavy in the other. The task of three experienced observers was to identify the interval containing the straight edge texture. The re-analysis results for the nine experiments<sup>3</sup> are summarized in Fig. 1B.

### A.3. Spatial frequency discrimination (McIntosh et al., 1999)

This study employed two different age groups – young (age range = 20–30) and old (age range = 60–79), and two ISI conditions. In one ISI condition the two temporal intervals were separated by 500 ms, and in the other by 4000 ms. Eight, old participants and seven, young participants saw a single vertical sinusoidal grating in each temporal interval and they had to indicate which interval included a grating with a higher spatial frequency. The results for this study are summarized in Figs. 2 and 3.

### A.4. Visual search (Carrasco & Yeshurun, 1998)

Five visual search data sets were re-analyzed, all involved detection of a pre-specified target appearing among non-relevant distracters and was performed by naïve, inexperienced observers. Each trial included two temporal intervals, both intervals included an array of elements but only one of them included a target. In the other interval all the elements were distracters. The observers had

<sup>3</sup> The study included 10 experiments, but due to technical difficulties only nine experiments were re-analyzed.

to indicated which interval included the target. A manipulation of spatial attention was added to this basic visual search task. Each presentation of the elements array was preceded by a cue. On *cued trials* a spatial cue appeared above the target location, allowing observers to focus their attention in advanced on the target location. On *neutral trials* a cue presented in the center of the display did not convey information regarding the upcoming target location. There were two additional independent variables: the number of elements presented in each array, and target eccentricity. In the Orientation feature search experiment observers were asked to detect the presence of a red vertical line appearing among red tilted lines (Experiment 3; 12 observers). In the Color × Orientation and Long Color × Orientation conjunction search experiments observers were asked to detect a red vertical line appearing among red tilted and blue vertical lines. The timing that elapsed between the precue onset and the display onset was longer in the latter than in the former (Experiments 3 and 4; 12 observers each). In the Color × Shape experiment, 11 observers searched for a red V target among blue Vs and red inverted V distracters, and in the Ls experiment 9 observers searched for a red mirror image L-like target among red 180° counterclockwise rotated Ls and red 90° clockwise rotated L distracters (see discussion of Experiment 3). The results for these experiments are summarized in Fig. 4.

## Appendix B

### B.1. Nested-hypothesis test

The data from a 2-IFC experiment can be summarized as a matrix  $[n_{ij}]$ ,  $i, j = 1, 2$  where  $n_{ij}$  is the count of the trials on which the observer responded ‘Interval  $i$ ’ when the signal was actually in interval ‘ $j$ ’. The entries  $n_{11}$  and  $n_{22}$  correspond to correct classifications. We denote the probabilities of each outcome as a second matrix  $[p_{ij}]$ ,  $i, j = 1, 2$  where  $\sum_{i=1}^2 p_{ij} = 1$ . We do not know these probabilities but we can estimate them by the proportion of responses  $i$  when the signal is in interval  $j$ ,

$$\hat{p}_{ij} = n_{ij} / \sum_{k=1}^2 n_{kj} \quad (1)$$

for  $j = 1, 2$ . These estimates are maximum likelihood estimates (Mood et al., 1974, pp. 276ff). The unknown probabilities  $p_{jj}$  are the probability correct in interval  $j = 1, 2$  and we want to examine the consistency of the data with the hypothesis that  $p_{11} = p_{22}$ .

We approach the problem by first developing a nested-hypothesis test (Mood et al., 1974, p. 441ff) of the hypothesis given the data. We first compute the unconstrained log likelihood

$$\lambda_1 = \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \log \hat{p}_{ij}. \quad (2)$$

Under the null hypothesis, we have  $p_{11} = p_{22} = p_c$ : the probability correct is the same in each of the two intervals and we denote this common value by  $p_c$ . The maximum likelihood estimator of  $p_c$  is

$$\hat{p}_c = (n_{11} + n_{22}) / \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \quad (3)$$

based on “pooling” the data in the two intervals. The constrained log likelihood is

$$\lambda_0 = (n_{11} + n_{22}) \log \hat{p}_c + (n_{12} + n_{21}) \log(1 - \hat{p}_c) \quad (4)$$

and it must be less than or equal to  $\lambda_1$  (the unconstrained log likelihood). The test statistic  $X = 2(\lambda_1 - \lambda_0)$  is asymptotically distributed as a  $\chi_1^2$  under the null hypothesis (Mood et al., 1974, pp. 441ff). To complete the nested hypothesis test, we could, for example, compare  $X$  to  $\chi_1^2(1 - \alpha)$ , the  $(1 - \alpha)$ -tile of the  $\chi_1^2$  random variable

for a test of size  $\alpha$ . We would reject the null hypothesis if and only if  $X > \chi_1^2(1 - \alpha)$  where  $\alpha$  is typically set to the conventional value 0.05.

We wished instead to develop a measure of the inconsistency of the data and the null hypothesis based on the  $p$ -value of this hypothesis test. We computed the exact  $p$ -value

$$p = 1 - F^{-1}(X) \quad (5)$$

where  $F$  is the cumulative distribution function of a  $\chi_1^2$  random variable. The  $p$ -value is the probability of obtaining the value  $X$  or greater if the null hypothesis were true. We use the transformed value  $\Gamma = -\log_{10}(p)$  as a convenient measure of the incompatibility of the data and the null hypothesis. In the main text, we will truncate  $\Gamma$  to an integer for convenience in representing it graphically. A value of  $\Gamma \geq 1.3$  would lead to rejection of the null hypothesis when  $\alpha = 0.05$ . A value of  $\Gamma > 6$  indicates that the data is extremely unexpected if the null hypothesis were true.

### B.2. The difference model and 2-way data

Using the same notation as in the preceding section (Appendix B.1) we summarize the outcome of a 2-IFC experiment as a matrix  $[n_{ij}]$ ,  $i, j = 1, 2$  where  $n_{ij}$  is the count of the trials on which the observer responded ‘Interval  $i$ ’ when the signal was actually in interval ‘ $j$ ’. We interpret the 2-IFC as a signal detection experiment, designating trials where the signal is in Interval 1 as ‘signal plus noise’ trials and those where the signal is in Interval 2 as ‘noise only’ trials. Then  $\hat{p}_{11}$  is  $\hat{p}[\text{HIT}]$  and  $\hat{p}_{12}$  is  $\hat{p}[\text{FA}]$  (‘false alarm’) in the usual nomenclature of signal detection theory and we can estimate the parameters  $d'$  and the sensory criterion  $c$  of the Difference Model in the usual way,

$$\begin{aligned} \hat{c} &= \Phi^{-1}(1 - \hat{p}_{12}) \\ \hat{d}' &= \Phi^{-1}(\hat{p}_{11}) - \Phi^{-1}(\hat{p}_{12}) \end{aligned} \quad (6)$$

where  $\Phi(x)$  is the cumulative distribution function of a Gaussian random variable with mean 0 and variance 1 (Green & Swets, 1973).

The motivation for the following digression will become clear when we consider fitting 4-way data in Appendix B.3.

An alternative method for fitting  $\hat{d}'$ ,  $\hat{c}$  is to choose the values of the parameters  $d'$ ,  $c$  which maximize the log likelihood

$$\lambda(d', c) = \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \log(p_{ij}(d', c)) \quad (7)$$

where

$$\begin{aligned} p_{11}(d', c) &= 1 - \Phi(c - d') \\ p_{12}(d', c) &= 1 - \Phi(c) \end{aligned} \quad (8)$$

and the remaining two probabilities can be computed from the constraints that  $p_{11} + p_{21} = 1$  and  $p_{12} + p_{22} = 1$ . We denote these maximum likelihood estimates by  $d'$ ,  $\hat{c}$  and, as the notation suggests, they coincide with the estimates obtained through Eq. (6).

To see that the estimates of Eq. (6) are maximum likelihood estimates it is only necessary to note that there is a 1–1 mapping between the parameters  $(d', c)$  and the parameters  $(p_{11}, p_{12})$  specified by Eq. (6) and its inverse. Any choice of parameters  $(d', c)$  determines the probabilities  $(p_{11}, p_{12})$  and vice versa (through Eq. (8)). It is unusual to treat the two probabilities as a parameterization of signal detection in this way, but it is perfectly valid to do so because of the 1–1 transformation. Then it is easy to show that the estimates  $(p_{11}, p_{12})$  that maximize likelihood are the usual proportions  $(\hat{p}_{11}, \hat{p}_{12})$ . And now we use a special property of maximum likelihood estimation: under a 1–1 remapping of parameters, maximum likelihood estimates are mapped to maximum likelihood estimates in the new parameterization (Mood,

Graybill & Boes, 1974, pp. 284ff). But then the maximum likelihood estimates of  $(d', c)$  are obtained by applying the 1–1 transformation to the maximum likelihood estimates  $(\hat{p}_{1|1}, \hat{p}_{1|2})$  and that is exactly what Eq. (6) does. Therefore the outcome of Eq. (6) must be the maximum likelihood estimates obtained by maximizing Eq. (7).

### B.3. Analyzing 4-way data

In the 4-Way task, the observer views sensory activity  $S_j$  in each of two intervals  $j = 1, 2$ . A signal can be present in either interval, both intervals, or neither. When a signal is present in Interval  $j$ , the distribution of the random variable  $S_j$  is Gaussian with mean  $d'_j$  and variance 1. When a signal is absent, the distribution is Gaussian with mean 0, variance 1. On each trial, a signal is present in either interval with probability .5 and the presence of a signal in either interval is independent of the presence or absence of a signal in the other.

The observer's task is to judge which of the two intervals contained signals and which did not. If we specify the presence of a signal by Y, its absence by N, then the four possible patterns of signal present and absent across the two trials can be denoted by YN, NY, NN and YY where the first letter specifies the state of the first interval, the second, the state of the second. The observer's possible responses are then yn, ny, nn and yy. Since events in both intervals are independent, the optimal decision rule for the observer (which maximizes the probability of correct response) is to carry out two ordinary Yes–No signal detection judgments, one on each interval. If we were certain that the observer is carrying out the actual task as two independent Yes–No tasks, then the data could be fit as data from two signal detection tasks using the methods described in Appendix B.2. However, we cannot assume that, for example, the observer's judgment in one interval affects the judgment in the other and consequently we fit the data by the method of maximum likelihood. For example, the human observer may be biased against responding NN and an N response in either interval would affect the probability of an N response in the other.

Details: For simplicity we denote the four possible trial types YN, NY, NN, and YY by the numbers 1, 2, 3, 4, respectively, and the four possible responses yn, ny, nn, and yy by 1, 2, 3, 4 as well. Thus, the probability of responding correctly on a YN trial is  $p_{1|1} = p_{y|YN}$  with notation analogous to that of Appendix B.2.

The model that we fit to the 4-way data has four parameters, sensitivity and criterion measures for the two intervals, which we denote  $d'_i, c_i, i = 1, 2$ . We fit these parameters by maximizing the log likelihood,

$$\lambda(d'_1, c_1, d'_2, c_2) = \sum_{i=1}^4 \sum_{j=1}^4 n_{ij} \log(p_{ij}(d'_1, c_1, d'_2, c_2)) \quad (9)$$

where the  $n_{ij}, i = 1, 2, 3, 4; j = 1, 2, 3, 4$  are the counts of actual responses of each kind. We use a numerical optimization method in MATLAB (Math Works, Inc.) to compute the values  $\hat{d}'_i, \hat{c}_i, i = 1, 2$  that maximize the likelihood in Eq. (9) above.

For the Yes–No task there are two parameters  $d'$  and sensory criterion  $c$  and two independent pieces of data, the proportion of HITS and the proportion of correct rejections and consequently the maximum likelihood estimates can be computed directly without numerical optimization by the formulas in Eq. (6). There is no need to perform a numerical optimization. For classification tasks involving more than two alternatives such as our 4-way task, numerical optimization allows rapid and accurate maximum likelihood estimates.

On each iteration of the optimization method, the numerical optimizer selects a candidate set of parameters  $d'_i, c_i, i = 1, 2$ . These must be transformed into the 16 values of  $p_{ij}(d'_1, c_1, d'_2, c_2)$  in order to compute the likelihood value in Eq. (9). The mapping from

parameters  $d'_i, c_i, i = 1, 2$  to probabilities  $p_{ij}(d'_1, c_1, d'_2, c_2)$  is not difficult to work out by reference to Fig. 6. The optimizer tries many candidate values of the parameters and converges on the values that maximize the likelihood in Eq. (6).

We describe briefly how to transform a candidate set of parameters  $d'_i, c_i, i = 1, 2$  into the sixteen probabilities  $p_{ij}(d'_1, c_1, d'_2, c_2)$ . First we compute the HIT and False Alarm (FA) probabilities for interval  $k = 1, 2$  in terms of the four parameters  $d'_i, c_i, i = 1, 2$ :

$$\begin{aligned} q_{\text{HIT}}^k(d'_k, c_k) &= 1 - \Phi(c_k - d'_k) \\ q_{\text{FA}}^k(d'_k, c_k) &= 1 - \Phi(c_k) \end{aligned} \quad (10)$$

Note that  $q_{\text{FA}}^k$  does not really depend on  $d'_k$ . We notate it this way for simplicity. Then to compute, for example,  $p_{2|3}(d'_1, c_1, d'_2, c_2)$  we translate it to  $p_{ny|NN}(d'_1, c_1, d'_2, c_2)$  and note that it is the probability of a correct rejection in the first interval and a false alarm in the second. Thus,

$$p_{2|3}(d'_1, c_1, d'_2, c_2) = (1 - q_{\text{FA}}^1(d'_1, c_1))q_{\text{FA}}^2(d'_2, c_2). \quad (11)$$

The sixteen values of  $p_{ij}(d'_1, c_1, d'_2, c_2)$  can be computed this way. Of course,

$$\sum_{i=1}^4 p_{ij}(d'_1, c_1, d'_2, c_2) = 1 \quad (12)$$

for  $j = 1, 2, 3, 4$ . Even with this constraint there are 12 probabilities  $p_{ij}, i = 1, 3; j = 1, 4$  controlled by the four free parameters  $d'_i, c_i, i = 1, 2$ . Thus, while any choice of the four parameters  $d'_i, c_i, i = 1, 2$  determines the probabilities, the converse is not true. Unlike the 2-way case, a set of estimates of  $\hat{p}_{ij}, i = 1, 4; j = 1, 4$  based on data do not have to correspond to a unique setting of the parameters  $d'_i, c_i, i = 1, 2$ . In the 4-way case, there is no analogue of the computational trick that made it very easy to compute the maximum likelihood fits in the 2-way case (Eq. (6)). Instead we maximize log likelihood numerically by choice of the parameters  $d'_i, c_i, i = 1, 2$  to arrive at maximum likelihood estimates  $\hat{d}'_i, \hat{c}_i, i = 1, 2$  that are basis for our analyses.

## References

- Alcalá-Quintana, R., & García-Pérez, M. A. (2005). Interval bias in discrimination tasks. European Conference on Visual Perception, A Coruña, August 22–26, 2005. *Perception*.
- Berliner, J. E., & Durlach, N. I. (1973). Intensity perception. IV. Resolution in roving-level discrimination. *Journal of the Acoustical Society of America*, 53, 1270–1287.
- Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2), 673–692.
- Creelman, C. D., & Macmillan, N. A. (1979). Auditory phase and frequency discrimination: A comparison of nine procedures. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 146–156.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). NY: Wiley.
- Durlach, N. I., & Braitin, L. D. (1969). Intensity perception. I. Preliminary theory of Intensity resolution. *Journal of the Acoustical Society of America*, 46, 372–383.
- Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.
- Green, D. M., & Swets, J. A. (1973). *Signal detection theory and psychophysics*. Huntington, NY: Krieger Publishing.
- Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, 6(11), 13. 1307–1322.
- Jesteadt, W., & Bilger, R. C. (1974). Intensity and frequency discrimination in one- and two- interval paradigms. *Journal of the Acoustical Society of America*, 55, 1266–1276.
- Kaernbach, C. (1990). A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing. *The Journal of the Acoustical Society of America*, 88, 2645–2655.
- Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, 63(8), 1377–1388.
- Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, 63, 1421–1455.
- Köhler, W. (1923). Theorie des Sukcessivvergleichs und der Zeitfehler. *Psychologische Forschung*, 4, 115–175.

- Leshowitz, B. (1969). Comparison of ROC curves from one- and two-interval rating-scale procedures. *Journal of the Acoustical Society of America*, 46, 399–402.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Markowitz, J., & Swets, J. A. (1967). Factors affecting the slope of the empirical ROC curves: Comparisons of binary and rating responses. *Perception & Psychophysics*, 2, 91–100.
- McIntosh, A. R., Sekuler, A. B., Penpeci, C., Rajah, M. N., Grady, C. L., Sekuler, R., et al. (1999). Recruitment of unique neural systems supporting visual memory in normal aging. *Current Biology*, 9, 1275–1278.
- Miller, J., & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman–Kärber method. *Perception & Psychophysics*, 63(8), 1399–1420.
- Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.
- Morgan, M. J., Watamaniuk, S. N. J., & McKee, S. P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research*, 40, 2341–2349.
- Nachmias, J. (2006). The role of virtual standards in visual discrimination. *Vision Research*, 46, 2456–2464.
- Needham, J. G. (1934). The time-error as a function of continued experimentation. *American Journal of Psychology*, 46, 558–567.
- Pynn, C. T., Braid, L. D., & Durlach, N. I. (1972). Intensity perception. III. Resolution in small-range identification. *Journal of the Acoustical Society of America*, 51, 559–566.
- Schulman, A. I., & Mitchell, R. R. (1966). Operating characteristics from yes–no and forced-choice procedures. *Journal of the Acoustical Society of America*, 40(2), 473–477.
- Swets, J. A., & Green, D. M. (1961). Sequential observations by human observers of signals in noise. In C. Cherry (Ed.), *Information theory: Proceedings of the fourth London symposium* (pp. 177–195). London: Butterworth.
- Viemeister, N. F. (1970). Intensity discrimination: Performance of three paradigms. *Perception & Psychophysics*, 8, 417–419.
- Watson, C. S., Kellogg, S. C., Kawanishi, D. T., & Lucas, P. A. (1973). The uncertain response in detection-oriented psychophysics. *Journal of Experimental Psychology*, 99, 180–185.
- Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of in absolute and comparative judgments. *Journal of Mathematical Psychology*, 6, 13–61.
- Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford.
- Wolfson, S. S., & Landy, M. S. (1995). Discrimination of orientation-defined texture edges. *Vision Research*, 35, 2863–2877.
- Wolfson, S. S., & Landy, M. S. (1998). Examining edge- and region-based texture mechanisms. *Vision Research*, 38, 439–446.