Atten Percept Psychophys
DOI 10.3758/s13414-012-0269-0

Decision criteria in dual discrimination tasks estimated using external-noise methods

Ido Zak · Mikhail Katkov · Andrei Gorea · Dov Sagi

Abstract According to classical signal detection theory (SDT), in simple detection or discrimination tasks, observers use a decision parameter based on their noisy internal response to set a boundary between "yes" and "no" responses. Experimental paradigms where performance is limited by internal noise cannot be used to provide an unambiguous measure of the decision criterion and its variability. Here, unidimensional external noise is used to estimate a criterion and its variability in stimulus space. Within this paradigm, the criterion is defined as the stimulus value separating the two response alternatives. This paradigm allows the assessment of interactions between criteria assigned to different targets in dual tasks. Previous studies suggested that observers' criteria interacted or even collapsed to one (hence, nonoptimal) criterion. An alternative interpretation of those results is that observers equated their false alarm (FA) rates. The external-noise method enables the confrontation of the two hypotheses. It is shown that the variability of observers' criterion in stimulus space is about 1.6 times their measured sensory threshold, suggesting that the presence of external noise increases decision uncertainty. Observers' stimulus criterion settings are close to SDT predictions in single tasks, but not in dual tasks where the two criteria tend to "attract" each other. Observers maintain distinct FA rates even when SDT predicts equal rates. Observers trained in psychophysics or provided with basic notions of SDT exemplified with the present experimental design manage to better separate their criteria in some conditions.

Keywords Decision making · Dual-task performance · Signal detection theory

I. Zak · M. Katkov · D. Sagi (✉)
Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel
e-mail: Dov.Sagi@Weizmann.ac.il

A. Gorea
Laboratoire Psychologie de la Perception (LPP), Université Paris Descartes & CNRS, 75006 Paris, France

Published online: 14 February 2012

Introduction

Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005) is the best known normative approach to assessing human's sensitivity (d') and decision behavior instantiated by this system's decision criterion (c). Both indices are expressed in units of the internal noise of the system, σ_{IN}. While the standard application of SDT provides a means of estimating the system's noise when responding to an external stimulus, *relative* to its noise in the absence of stimulation, via the assessment of this system's receiver operating characteristic (ROC), it cannot be used to compare noise levels across different stimuli and tasks. As a consequence, the mean internal response evoked by any given stimulus and the associated decision criterion (i.e., the internal response level serving as the decision boundary) are expressed in these arbitrary units. The estimation of these internal parameters, as well as of d' and c, relies on some critical assumptions regarding the statistics of the internal events and the dynamics of the associated decision mechanism. Thus, not surprisingly, the validity of the assumptions made and their associated consequences for the obtained measure of sensitivity are under extensive research (Balakrishnan, 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2008; Treisman, 2002; Weidemann & Mueller, 2008).

Previous results of *dual* detection and *dual* discrimination tasks (Gorea & Sagi, 2000, 2001, 2002) provided evidence

for the use by observers of a unique decision criterion, *uc*, even though an optimal behavior requires (and the experimental design permits) the use of distinct criteria as actually observed in *single* (detection or discrimination) tasks. In Gorea and Sagi's (2000, 2001) terminology, this *uc* is defined as a value on the sensory continuum, measured in σ_{IN} units referenced to the mean of the noise, $\mu_{IN} = 0$. This number is but the *z*-score of observers' false alarm (FA) rate, $z\text{FA} = c/\sigma_{IN}$. However, as was pointed out by Kontsevich, Chen, Verghese, and Tyler (2002), Gorea and Sagi's experimental design could not discriminate between the use by observers of a *uc* (i.e., $\sigma_{IN} \times z\text{FA}$) or of a unique FA rate, inasmuch as the *actual* σ_{IN} is unknown and arbitrarily set to 1. The main purpose of the present work is to disambiguate the *uc* versus unique FA rate dilemma. To do this, we override σ_{IN}, with an experimentally controlled external noise $\sigma_{EX} >> \sigma_{IN}$ in a task where observers have to classify a stimulus (Gaussian luminance blob) as belonging to one of two overlapping distributions varying along a single dimension (Lee & Janke, 1964)—here, luminance (or contrast). The use of unidimensional external noise with well-defined statistical properties allows for the assessment of decision criteria in stimulus space (stimulus criterion) as the stimulus level at which observers switch from one response alternative to the other, thus largely avoiding the above-mentioned limitations inherent in the standard SDT analysis.

In the present experiments, the displayed luminance levels were drawn from normal distributions with standard deviations (σ_{EX}; i.e., external noises) at least twice observers' just noticeable (luminance) difference (JND) as assessed in separate experiments. Hence, most flashed blobs, whether they belonged to the noise (N) or signal (S) distributions, were highly suprathreshold. Because of the overlap of the N and S distributions, a given stimulus cannot be unequivocally attributed to either one of them. An ideal observer (with no internal and decisional noise) will make its decision according to the most likely distribution associated with this stimulus level. For the present experimental formats, this amounts to picking out a specific stimulus level whose likelihood of belonging to either of the two distributions is the same (a likelihood ratio of 1) and assigning any given flashed blob to either N or S depending on whether its luminance is lower or higher, respectively, than the reference one. This reference value is the observer's decision criterion *c* expressed in *known* stimulus intensity units, rather than in *unknown* internal-noise units—thus, here termed *stimulus criterion*. For a noisy observer (nonideal), it is the stimulus level yielding an equal number of responses assigned to both distributions—namely, the inflexion point of the psychometric function fitted to the proportion of "S" responses for each presented stimulus level. The standard deviation of the fitted cumulative distribution function is a measure of observer's decision criterion variability over trials (here, in luminance units). This criterion variance must be the sum of internal noise associated with the internal response evoked by the corresponding luminance level and decision noise. Inasmuch as decision noise does not depend on experimental conditions, the presently assessed criterion variability should match the discrimination threshold (JND) between two luminance levels around the criterion measured in luminance units. Importantly, the direct estimation made here of the stimulus criterion and of its variance (in stimulus space) avoids the dependence on questionable SDT tools and assumptions (see Balakrishnan, 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2008). Given that the internal response is monotonically related to stimulus contrast (for contrast > 0), a comparison between criteria expressed in stimulus luminance units maps onto the comparison of the corresponding internal evoked responses. Hence, in contrast to previous experiments that could not discriminate between criteria equality being due to a *uc* or to an FA equality (Gorea & Sagi, 2000, 2001; Kontsevich et al., 2002), the present external-noise paradigm does permit such a distinction. This is so because the performance-limiting noise, σ_{EX}, is known and can be manipulated; given that, in the present conditions, $c = z\text{FA} \times \sigma_{EX}$ (assuming $\sigma_{EX} >> \sigma_{IN}$) observers cannot equalize at the same time their FA rates and *c* when facing two tasks with a different σ_{EX}.

More specifically, the present study consists of a series of dual-task experiments where observers decide from which luminance contrast distribution (S or N) a marked target—that is, one of two simultaneously flashed Gaussian luminance blobs—was drawn. In one class of experiments, the two stimuli (presented at two symmetrical locations about fixation) were drawn from the same N and S distributions. This configuration will be hereafter referred to as dual-same (DS; see Fig. 1a); it was meant to serve as a reference baseline since it should (and actually does) yield identical d' and *c* values for the two simultaneously flashed blobs. The DS configuration was also used to assess putative decisional biases (relative to the ideal observer) under external-noise conditions where no such biases are to be expected. The critical experiments involved two classes of dual-different (DD) tasks. The first class (DD1) differed from the DS experiments in that the two S distributions had different means (but an equal σ_{EX}), thus entailing different d' and optimal *c* values (Fig. 1b). In the second class (DD2), the two S distributions differed in both their μ_{EX} and σ_{EX} (with σ_{EX} being equal for paired N and S distributions) so that each N–S pair entailed identical d' values but different optimal *c* values (Fig. 1c). The critical difference between DD1 and DD2 is that in DD1 equal criteria entail equal FA rates, while in DD2 criteria and FA rates are dissociated. The main empirical question asked is whether, as in Gorea and Sagi's (2000, 2001) threshold experiments involving only internal noise, the present external-noise DD experimental format will also reveal a *uc* decisional behavior.

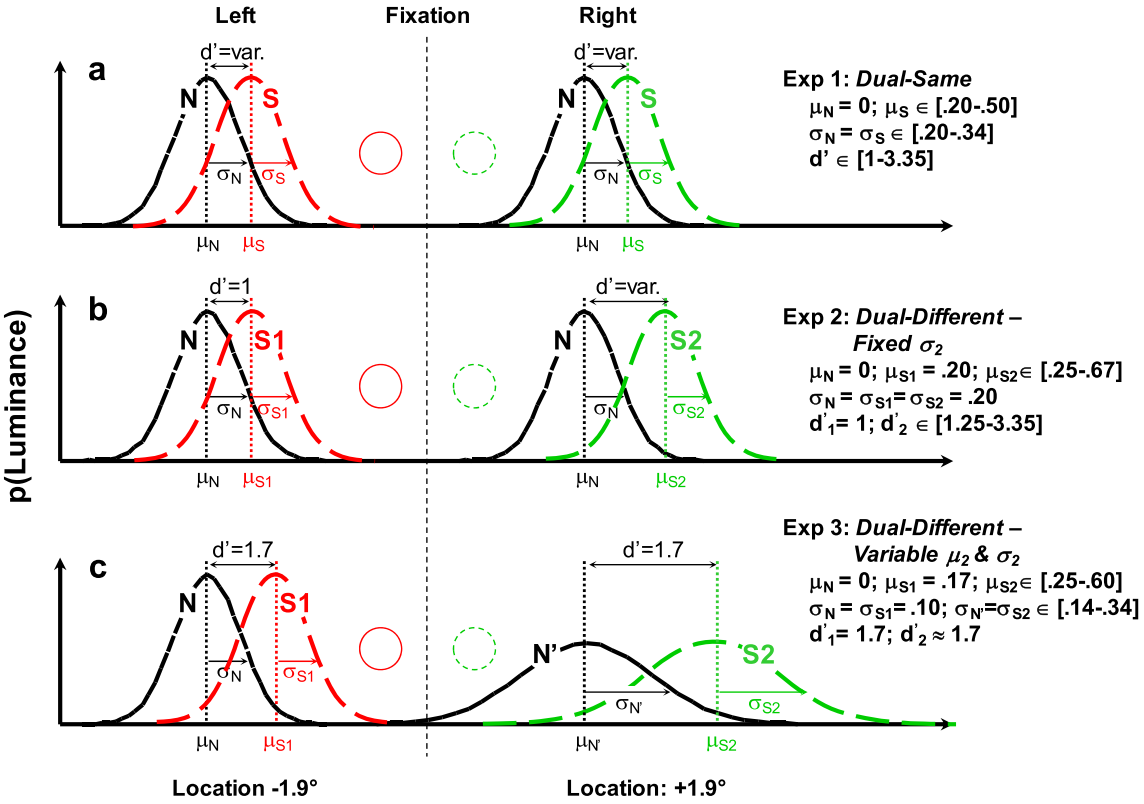

Fig. 1 Illustration of the noise (N; solid Gaussians) and signal (S; dashed Gaussians) normal distributions used in the paired tasks of all experiments. A luminance blob was presented on each side of fixation, one within a red circle, the other within a green circle (shown here with continuous and dashed outlines, respectively; see Fig. 2). The amplitude of each blob was drawn with equal probabilities from either the S or the N distribution. For any given dual task in Experiments DD1 and DD2, the red (continuous)/green (dashed) circle let the observer know that the blob within could have been drawn from the S distribution with the highest/lowest mean (green/red Gaussians), respectively, or from the associated N distribution. The means (μ) and standard deviations (σ) of each of the four distributions follow the rules shown in the insets and are detailed in Tables 1, 2 and 3

Method

Stimuli and experimental conditions

Stimuli were two Gaussian luminance blobs $\{I(x,y) = A\exp[-(x^2+y^2)/2\sigma^2)]$; $\sigma = 0.28°$, x and y are relative to stimulus location, and amplitude A set per trial according to the experimental condition as described below$\}$ presented on the left and on the right of fixation at ±1.9° eccentricity (Fig. 2), on a linearized 19-in. Philips Brilliance 109p4 monitor (85-Hz raster rate) with a GeForce 8600GT adapter at 1 m from observers' eyes. The Gaussian blobs were added to a screen of uniform luminance (29 cd/m^2). We here refer to *stimulus amplitude* as the intensity difference between the Gaussian peak intensity and the background and to *contrast* as the ratio between amplitude and background intensities. The amplitude of the Gaussian blobs was drawn randomly across trials from a normal distribution with parameters defined according to the stimulus specification in the experimental conditions. Colored cues (red and green) were presented at the two stimuli locations to mark the task assigned

Fig. 2 Spatio-temporal display of one trial in Experiments 1–3: The left stimulus sequence includes the fixed signal (marked by a red circle), while the right stimulus sequence includes the variable signal (marked by a green circle), with the postcue pointing to the latter

to each location (see below). Within each of these colored cues in any experimental condition stimuli were drawn randomly from either the "noise" or "signal" distribution: N (black Gaussians in Fig. 1), with zero mean (equal to the screen luminance), and S (red and green Gaussians in Fig. 1), with mean varied across conditions.

In the DS experiment (DS; Fig. 1a), stimuli from the same N and S distributions were used for both tasks—hence, its DS tag. The mean of S as well as the standard deviation (σ_{EX}) of both N and S were varied across sessions so as to yield theoretical d' values in between 1 and 3.35 (Appendix, Table 1). These manipulations defined 14 distinct experimental conditions.

In the DD1 experiment (Fig. 1b), there were two different S distributions, S1 and S2 (hence, its DD tag), characterized by a fixed and a variable mean, respectively, with all S and N distributions having equal standard deviations (σ_{EX}). Theoretical d' values were 1 for S1 and in between 1.25 and 3.35 for S2 (Appendix, Table 2). There was a total of six experimental conditions in Experiment DD1, corresponding to the six levels of S2.

In the DD2 experiment (Fig. 1c), N1 and S1 parameters were kept constant, but the mean and σ_{EX} of S2 (with σ_{EX} of N2 equal to σ_{EX} of S2) covaried so as to keep a roughly constant $d' \approx 1.7$ (Appendix, Table 3). The design of this experiment is such that an ideal observer would use different criteria across the paired tasks but equal FA rates. There were a total of 6 experimental conditions in Experiment DD2, corresponding to the six levels of S2.

While draws from the paired N and S distributions were randomly presented on the left and right of fixation, these specific pairings were consistently tagged by one red and one green "precue" circle (radius 1.15°) whose onset preceded by 400 ms the Gaussian blobs (80-ms presentation) and persisted for 400 ms after their offset, with one of them (randomly chosen) persisting for yet another 1,000 ms (Fig. 2). This extra duration indicated which of the two simultaneously presented blobs was the *target* to be reported as belonging to the respective (i.e., as tagged by the color of the persistent cue circle) N or S distribution.[1] This experimental design is of the *partial report* type and was chosen to prevent the onset and offset transients of the cues from interfering with the stimulus to be judged. With the exception of Experiment DS, where the stimulus blobs on the left and right of fixation were drawn from the same paired N and S distributions, the color of the cue circles denoted the fixed S1/N1 (red circle) and the variable S2/N2 (green circle) distributions.

[1] Note that this timeline does not compel observers to make their decision while the two cue circles are present (i.e., 400 ms). Decision can be (and typically is) made well after the offset of the stimulus.

Procedure

The three main experiments (DS, DD1, DD2) and the preliminary one (JND) were run in the following order: JND, DD1, DD2, DS for observers O.G. and N.E. and JND, DS, DD1, DD2 for observers O.T., D.H., and I.Z. A second JND assessment was run at either halftime or at the end of all experimental sessions.

Main experiments Observers were told that on each side of fixation, a Gaussian luminance blob would be drawn randomly and independently from either a low (N) or a high (S) mean luminance distribution of which they were shown samples during an initial training phase (10 trials/condition). For each of the three distinct experiments, observers were informed that the N and S distributions were either identical across sides, so that the color of the circles was irrelevant (DS), or that the S blobs, drawn within the green circle, were to be, on average, more intense than those drawn within the red circle (DD1, DD2). They were prompted to pay attention to the colors of the circles so as to calibrate their responses accordingly. The notion of "on average" was stressed and exemplified during the initial training trials. Observers were instructed to report, by left/right clicking the mouse, whether the blob presented in the persistent circle belonged to the N or the S distribution tagged by the persistent circle color. They were asked to maximize the number of correct responses and received auditory feedback for the incorrect ones. It was also made clear to them that the same intensity blob might belong to either the N or the S distribution, so that feedback might occasionally sound inconsistent with respect to their memory of past trials. Observers were also told that the alphabetical order of the letter presented on the screen before each DD1 and DD2 session indicated the mean contrast *rank* of S2 distributions tagged by the green circle. For the DD2 experiment (always run after DD1), observers were told that the distribution of luminances within the green circle might appear different from that in DD1 but were not given further information (i.e., that the luminance variability (σ_{EX}) of the blobs presented in the green circles was larger than that in the red circles). Finally, since the external-noise method entails trials with lower blob amplitudes than the background (contrast < 0), observers were made aware of the dark appearance of such amplitudes and were instructed that, on such trials, they should always classify the blobs as belonging to the N distribution. On such trials, they were warned that feedback would be partly inconsistent, since such darker than background blobs could belong to either N or S distributions. These trials were used to estimate the frequency of finger errors (see the Data Analysis section). Sessions were limited to 1 h a day.

The 14 distinct conditions (mean S intensities) in Experiment DS were repeated 4 times each (56 blocks

total), with each repeat/block consisting of 100 trials, totaling 400 trials per condition. One experimental session consisted of 8 blocks. Each of the first 8 conditions was repeated 4 times before continuing with the next condition. The remaining 6 conditions were randomized within and across sessions.

The six distinct conditions of Experiments DD1 (fixed σ_{EX}) and DD2 (variable σ_{EX}) were repeated 8 times each (100 trials per repeat/block), totaling 400 trials for each ("red"/"green") task. One session consisted of eight blocks (800 trials), randomly drawn from the six different conditions, with a new random draw for each session.

Each block of trials started with 12 training trials whose ending was announced by a beep. Since the three main experiments were not mixed, the instructions above were given at the beginning of an experiment and only occasionally repeated to the naive observers, who described their impressions between blocks.

Just noticeable difference JNDs were measured before the main experiments with a standard yes/no procedure using the same Gaussian blobs as in the main experiments, but this time set at a fixed reference (pedestal) luminance of one of three values (31, 34.5, 38.5 cd/m²) tested in independent blocks. Each of these three values was paired with four luminance increments (in the range of 4–8.5 cd/m²). Pedestal only and pedestal + increment stimuli were presented with equal probability (.5). The d' values were computed from at least 100 trials per pedestal and per increment value (at least 1,200 trials total per observer), yielding three psychometric functions per observer. The JND increment values yielding $d' = 1$ were derived from the linear regressions fitted to each set of data. These three values (one per pedestal) were fitted with a "threshold versus contrast" function (e.g., Gorea & Sagi, 2001) to allow the interpolation/extrapolation of the JNDs corresponding to the randomly drawn target luminances in the main experiments. Five out of 6 observers repeated all JND measurements after gaining some experience in the main experiments.

Training naive observers Three observers who were naive as to the experiments were subsequently given a detailed explanation of the experimental design and a short SDT tutorial bearing on the internal representation of signals, on the ideal placement of criteria, and on how external noise may influence their decisions (a related procedure was used by Lee & Janke, 1964). The experiments' goals were also revealed, and they were informed about their observed inability to separate their criteria (as naive observers) in the paired tasks (see the Results section). They were then given four to seven sessions of 800 trials each under condition C or D of experiment DD1 (Appendix, Table 2). Two of these observers showed an improvement in their decisional behavior (with respect to optimality), were considered "informed observers," and were run once again through all conditions of Experiments DD1 and DD2.

Observers

Six observers participated in the experiments (2 additional observers were discarded once they had completed Experiment DS on the basis of their finger error rates being higher than 10% or of too slow and unreliable responses). Two observers (E.S. and author I.Z.) were trained psychophysicists and were considered as informed. The remaining 4 observers were either undergraduate students or staff of the Weizmann Institute, 22–43 years of age, with normal or corrected-to-normal vision. Five observers were ran through all experiments (4 naive and the author), and 1 informed observer (E.S.) ran only the DD1 experiment. Two of the originally naive observers ran the DD1 and DD2 experiments once again after being given a detailed account of the experimental design and a short SDT tutorial.

Data analysis

Ideal observer Since observers' internal noise was overridden by the external noise, we posit an ideal observer in the sense of SDT, but with no internal or decision noise. For every experimental condition, the ideal observer uses a fixed criterion for all trials so as to minimize the expected number of errors. It provides a "high" (S) response for all targets with contrasts higher than the criterion and a "low" (N) response otherwise. It is in this sense that we refer hereafter to any deviation from optimality.

Due to finite sampling, it is possible that the d' of the ideal observer may differ from the one predicted based on the theoretical stimulus parameters (expected $\Delta\mu_{EX}/\sigma_{EX}$). Therefore, d' values were computed on the basis of the statistics of the actually presented stimuli. More specifically, ROC functions [i.e., p(Hit) vs. p(FA)] were constructed (Fig. 3), and d' was computed as the area under the ROC curve, multiplied by $\sqrt{2}$ (Green & Swets, 1966).

Estimating observers' criterion and criterion noise Observers' criteria and associated noises were estimated from fitting a psychometric function—namely, the function relating the proportion of "high" (S) responses to targets' contrast (see example in Fig. 4). The proportions of S responses were corrected according to each observer's finger error rate (see the Finger Errors section below). On the assumption of a locally linear transducer (the function relating input and output signals within the contrast range defined by the psychometric function) and of a normally distributed

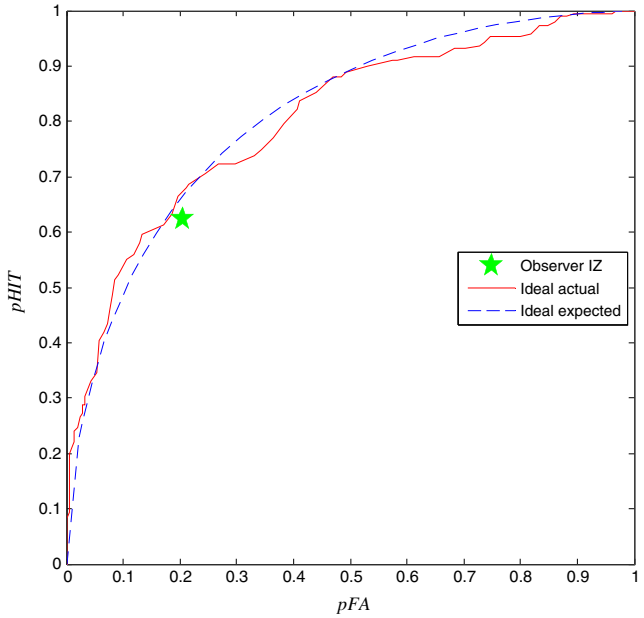

Fig. 3 ROC curves, an example: Experiment DD1, condition A. The star mark represents observer I.Z.'s performance (p_{Hit}, p_{FA}). The continuous (red) ROC curve is based on the *actual* luminance samples presented in the experiment, assuming an ideal observer. The blue (dashed) curve is the theoretical ROC curve based on Gaussian distributions with the predefined μ and σ_{EX} parameters of the N and S distributions used in this particular condition

internal noise (perceptual and decisional), the proportion of S responses for a stimulus of contrast X will be

$$p(S|X) = 1 - \Phi\left(\frac{c - X}{\sigma}\right) = \Phi\left(\frac{X - c}{\sigma}\right),$$

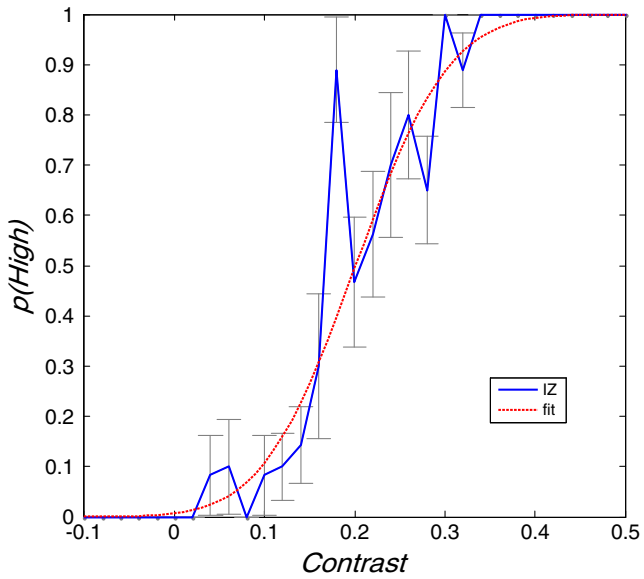

Fig. 4 An example of a criterion psychometric function (observer I.Z., Experiment DD1, condition A, 400 trials). The blue (solid) line shows the proportion of "high" (or S) responses for each stimulus contrast bin, with error bars showing ±1 binomial *SE*. The red (dotted) curve is the fitted logistic function (see the Data Analysis section)

where Φ is the standard normal cumulative distribution function, σ is the standard deviation of combined perceptual and decision noise, and c is the criterion (both measured in stimulus contrast units). However, the observed results were better fit by the logistic function

$$\frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}},$$

where $-b_0/b_1$ is the criterion c and $1/b_1$ is the criterion noise (i.e., the s1ope of the psychometric function). Fitting was done using MATLAB's generalized linear model fit, also known as logistic regression (Sokal & Rohlf, 1995). A simulation was performed to evaluate the goodness of fit of the logistic regression model. The model could not be rejected for any of the 6 observers (all ps > .25).

The standard error (*SE*) of each estimated criterion was computed using inverse regression (Neter, Wasserman, & Kutner, 1983):

$$SE = \sqrt{\frac{MSE}{b_1^2}\left[1 + \frac{1}{n} + \frac{(c - \overline{X})^2}{\sum (X - \overline{X})^2}\right]},$$

where n is the number of trials. Criteria were also independently computed with an alternative method that yielded practically identical (not statistically different) results.[2]

Finger errors Finger error is a term that includes several types of errors that are assumed to be independent of stimulus strength: clicking the wrong button, answering to the wrong target, guessing when failing to attend a trial. In the present external-noise experiments, finger error rates could be estimated directly, since observers were informed that all targets darker than the background should be given an "N" ("low") response. Therefore, an "S" ("high") response on such trials is a finger error. Observers' finger error rates (*fe*) were estimated as the proportion of "dark" trials on which they answered "S":

$$fe = \frac{\text{number of "dark" presentations reported as "S"}}{\text{total number of "dark" presentations}}$$

[2] Criteria were computed with an algorithm that maximizes the number of consecutive trials (the *criterion patch*) with a stable criterion (i.e., one whose estimate did not change for at least 10 trials). First, for every trial, a criterion was selected that accounted for the largest number of consecutive trials (i.e., lower and higher than all the following contrasts classified as signal and noise, respectively). Second, all such patches were gathered in a set of conflicting criterion patches. A dynamic programming algorithm (Cormen, Leiserson, Rivest, & Stein, 2001) was used to find the subset of nonconflicting criterion patches that included the largest amount of trials. The mean criterion was obtained by averaging the criteria over those patches, with each such criterion weighted by the number of trials included in the corresponding patch. The mean difference between criteria estimated with the two methods was 0.62% (*SE* = 0.1) of the mean screen luminance.

These rates were less than .02 for all observers but O.T. (.08) (1 observer with $fe > .1$ was excluded). The probability of an "S" response (p_i) can be derived from the percentage of "S" responses observed (p_{obs}) and the fe rate:

$$p_{obs} = p_i(1 - fe) + (1 - p_i)fe = p_i + fe - 2p_i fe \quad \Rightarrow$$

$$p_i = \frac{p_{obs} - fe}{1 - 2fe}.$$

Results

Optimal and nonoptimal criteria placing

Figure 5a displays the estimated observers' criteria versus optimal criteria in Experiment DS, with different symbols for the different observers. Criteria are shown in units of stimulus contrast, thus corresponding to the stimulus contrast level that is classified in equal proportion as noise and as signal (see Fig. 4). Ideal criteria are the luminance values (or contrasts) where the N and S luminance distributions cross over. For equal σ N and S distributions ($\sigma_N = \sigma_S$, Experiment DD1), the crossover point equals half of the theoretical d' (here derived from the actual means of the N and S samples presented in the experiments). The slope of the regression (red/continuous) line is close to unity (slope 0.92) and is not significantly different from 1 ($p > .2$), indicating that, aside from a constant positive (conservative) bias, (indicated by the positive intercept at the origin: 0.13) observers' decision behavior is close to optimality (dashed straight line). Figure 5b displays measured versus optimal criteria for the variable signals in Experiments DD1 and DD2, together with the regression line yielding now a slope of 0.41, significantly different from 1 ($p < .0001$). Hence, observers do not place their criteria optimally when handling two different tasks. This decisional behavior in the DD tasks replicates the behavior observed in previous studies with close to threshold stimuli (i.e., where performance was limited by internal noise only; Gorea & Sagi, 2000, 2001), thereby generalizing the notion of "criteria attraction" to the case where performance is limited by external noise.

To verify that these criteria shifts are due to decisional processes per se and not to responses to the unprobed blob, we measured linear correlations between P("High") and the unprobed blob amplitude, finding $r < .09$ across all conditions and observers (see the Appendix, Fig. 11). Such responses, if due to confusion between target and nontarget blobs, are independent of target amplitude and are captured by the finger errors parameter in our data analysis, thus having no effect on the estimated criteria (see the Finger Errors section).

Nonoptimal criteria placing under dual-different conditions

Figure 6 displays all observers' criteria (symbols) as derived from the psychometric function fits (see the Data Analysis section), together with ideal observer criteria (i.e., as predicted by SDT; solid lines) in Experiments DD1 and DD2 (as

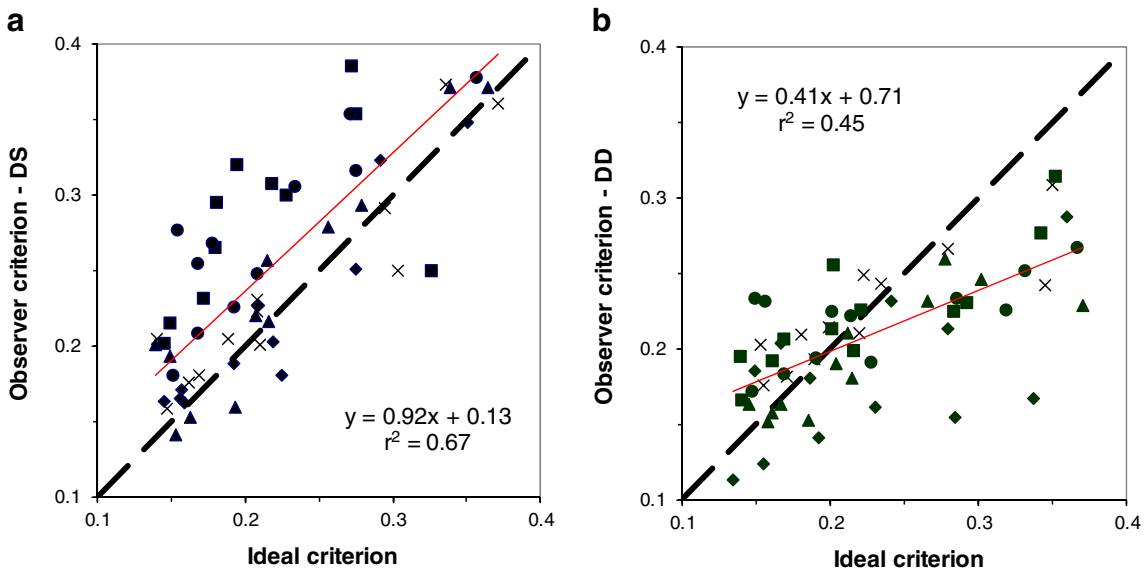

Fig. 5 Estimated observers' versus ideal criteria in the **a** dual-same (Experiment DS) and **b** dual-different (Experiments DD1 and DD2) conditions. Different symbols are for different observers (all naive except I.Z.: ×). Straight continuous lines are linear regressions over all data in the DS experiment (slope 0.92; not significantly different from the identity line; $p > .2$) and for the variable mean signal tasks in the DD experiments (slope .41; significantly different from the identity line; $p < .0001$). Dashed straight lines denote ideal-measured identity (slope 1). These results show that criteria in DS are slightly shifted from the ideal, reflecting a small constant response bias, but spread across an equal range, while in DD, observers' criteria are narrowly distributed as compared with the ideal. Criteria are given in units of stimulus contrast

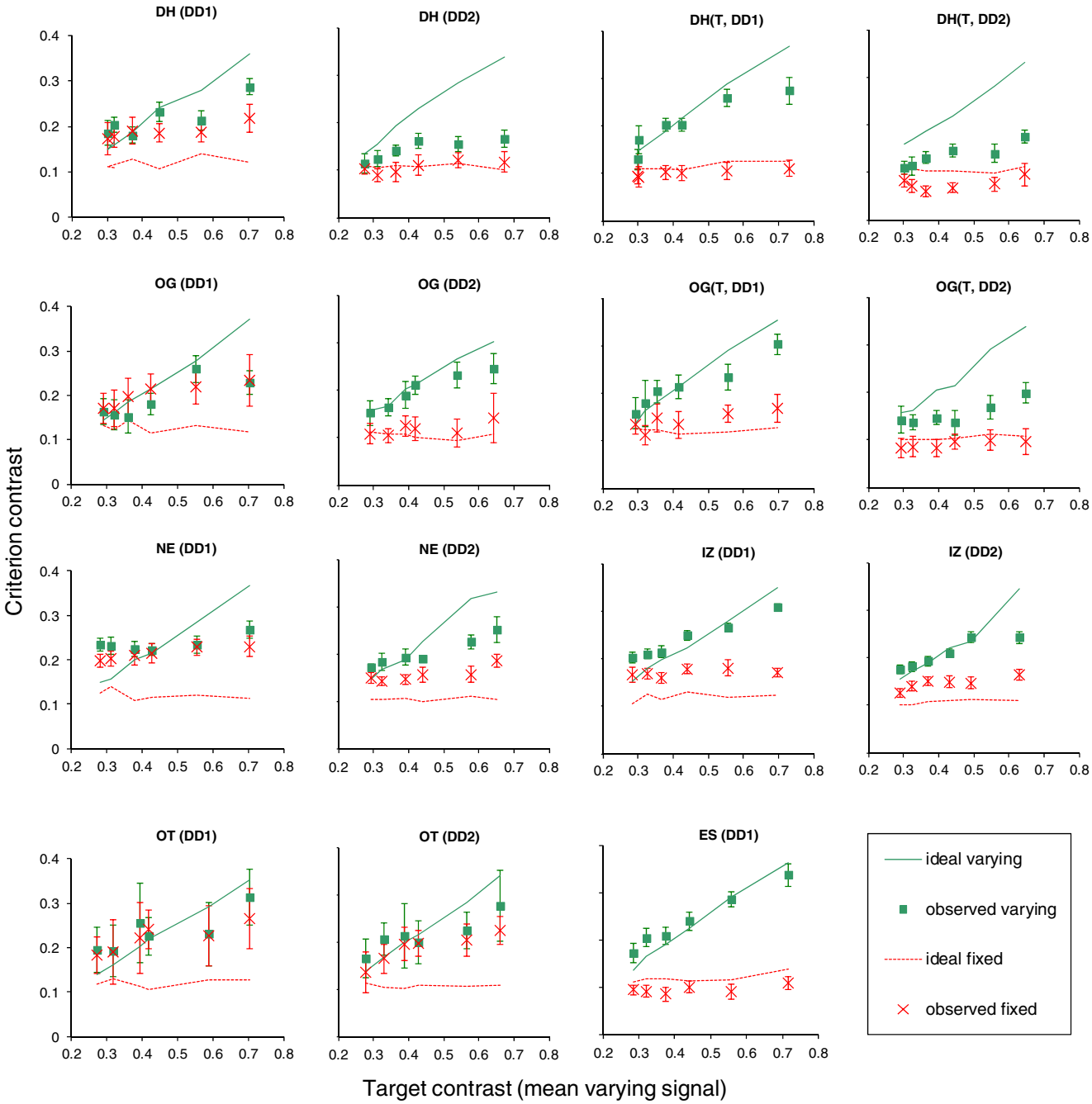

Fig. 6 Criteria for the paired tasks in Experiments DD1 and DD2 (signaled by the tags between parentheses adjoined to each observer's initials) as a function of the varying signal mean contrast. Symbols and lines show observers' and ideal observer's criteria, respectively, for the fixed-mean-signal task (red crosses, dotted lines) and for the variable-mean-signal task (green squares, solid lines). The first two (left) and last two (right) columns show data for naive and "informed" observers, respectively. Error bars are ±1 *SE*

indicated by the tags adjoined to the observers' initials) as a function of the mean contrast of the variable signal. Red crosses and red/dotted lines show criteria for the fixed signal tasks, and green squares and green/continuous lines show criteria for the paired variable signal tasks. The first two columns show data for the naive observers (D.H., O.G., N.E., and O.T.). The remaining two rightmost columns show data for the "informed" observers, considered so either because they were given training after having completed a first round of experiments [DH(T) and OG(T)] or because they were trained psychophysicists familiar with SDT (including one author, I.Z.). The main observation is that naive observers, with the exception of O.G.–DD2, tended to use very similar and sometimes identical criteria across the paired tasks (symbols at the same location on the *x*-axis) in both DD experiments. Note that an optimal behavior requires that the

difference between these criteria increase with the contrast of the variable signal (compare dotted and solid lines). The observed criteria "attraction" appears to be more or less symmetrical, since the paired criteria tend to meet at the mid-range between the optimal criteria. The decisional behavior of the "informed" observers is substantially closer to that prescribed by the optimal observer in Experiment DD1, but not in Experiment DD2.

Comparing criteria and *z*FA differences

In Experiment DD1, the external-noise variance was the same across the two N and the two S distributions. As a consequence, the "equal *c*" and the "equal *z*FA" hypotheses cannot be dissociated (see the Introduction). In Experiment DD2, this dissociation was made possible by assigning different variances to the two N and S pairs. More specifically, the ratio between σ_{EX} and the mean S–N contrast difference was kept equal across the two S and N pairs and constant across the different mean S2 values (see Fig. 1c and Table 3 in the Appendix), thereby yielding equal ideal *z*FA and d' values but distinct ideal criteria (c).

Figures 7 (Experiment DD1; fixed σ_{EX}) and 8 (Experiment DD2; variable σ_{EX}) display measured (red/horizontal stripes and green/vertical stripes bars for naive and informed observers, respectively) and ideal observer[3] (blue/dotted bars) criteria (panel a) and *z*FA (panel b) *differences* between the two paired tasks. In Experiment DD1, the decision behavior of informed observers is close to that of the ideal observer, with deviations from it increasing with mean S2 contrast (compare vertically striped and dotted bars in Fig. 7a, b). Instead, naive observers (red/horizontally striped bars) show close to zero criterion differences and hence, given the experimental design, also close to zero *z*FA differences for all conditions. In contrast, for the three conditions with the highest mean S2 (and σ_{EX}) in Experiment DD2, observers show significant differences between their *z*FAs (*t* tests yielding $p < .02$; Fig. 8b) and, accordingly, given the experimental design, reduced criteria differences relative to the ideal observer (Fig. 8a). In fact, the measured criteria differences are almost constant across conditions, unlike the ideal observer differences that increase with the mean S contrast difference by up to ~3.5 times the experimentally observed difference for the naive observers. As is shown in Fig. 9, for uninformed naive observers (except I.Z.), these criterion differences were stable across the eight testing sessions.

[3] Ideal observer performances are computed from the actual luminance distributions (i.e., as drawn for each observer and experiment); see the Method section.

Efficiency and criterion noise

Efficiency, defined as

$$\left(\frac{d'_{observed}}{d'_{ideal}}\right)^2,$$

was calculated to evaluate observers' performance relative to an ideal observer (see Footnote 3), with performance limited only by the external noise used in the experiment (Burgess, Wagner, Jennings, & Barlow, 1981). It should be noted that efficiency is not affected by the criterion setting that determines the percentage of correct responses. The efficiency measure allows us to examine whether the detected criteria changes are accompanied by an increase in internal noise (sensory and/or decisional). Averaged across the 5 observers that performed all the DS, DD1, and DD2 conditions, it was found to be .65 (.77, .68, .61, .72, and .44, for observers D.H., I.Z., N.E., O.G., and O.T., respectively). Observers' efficiency was stable across conditions: .64, .58, .59 for DS, DD1, and DD2, respectively (averaged across observers), with no significant difference between naive and trained observers (pairwise *t*-test). Efficiency increased with external-noise level from ~0.5 at $\sigma_{EX} = 0.1$ to ~0.8 at $\sigma_{EX} =$ 0.3–0.4 (Fig. 10). Since the ideal observer is assumed to be limited by external noise only, inefficiencies must be due to observer-related noise, such as internal (coding) noise and decision noise, here estimated by criterion noise (σ_c; see the Method section and Fig. 4; note that all measurements were corrected for finger errors). Criterion noise was found to correlate with the measured criterion (linear regression, all observers: $\sigma_c = 0.40\ c + 0.36$, $r^2 = .12$, $p < .004$), possibly reflecting the well-known increase of JND with increasing stimulation level (Weber law). On average, σ_c was 1.57 times larger than observers' JND (2.09, 1.51, 1.1, 2.47, 1.2, and 1.63, for observers D.H., I.Z., N.E., O.G., O.T., and E.Z., respectively).

Assuming linear transduction and statistical independence of criterial and external noises, the total noise in the system is $(\sigma^2_{EX} + \sigma^2_c)$. Therefore, efficiency is given by $\sigma^2_{EX}/(\sigma^2_{EX} + \sigma^2_c)$. Taking into account the observed correlation found between criterion noise and external noise means (linear regression, all observers: $\sigma_c = 0.48\sigma_{EX} + 0.04$, $r^2 = .17$, $p < .001$), it is possible to calculate the predicted efficiencies. The results show that the predictions match the mean measurements quite well—thus, since the model contains no free parameters, demonstrating a consistency between the different calculations and approximations made (Fig. 10).

The present data and analyses show that, at least for cases where external noise significantly exceeds internal noise ($\sigma_{EX} > 0.2$; see Fig. 10), external noise is the main limiting factor in observers' performance. At these noise levels, criterion noise (σ_{crit}) is approximately proportional to external

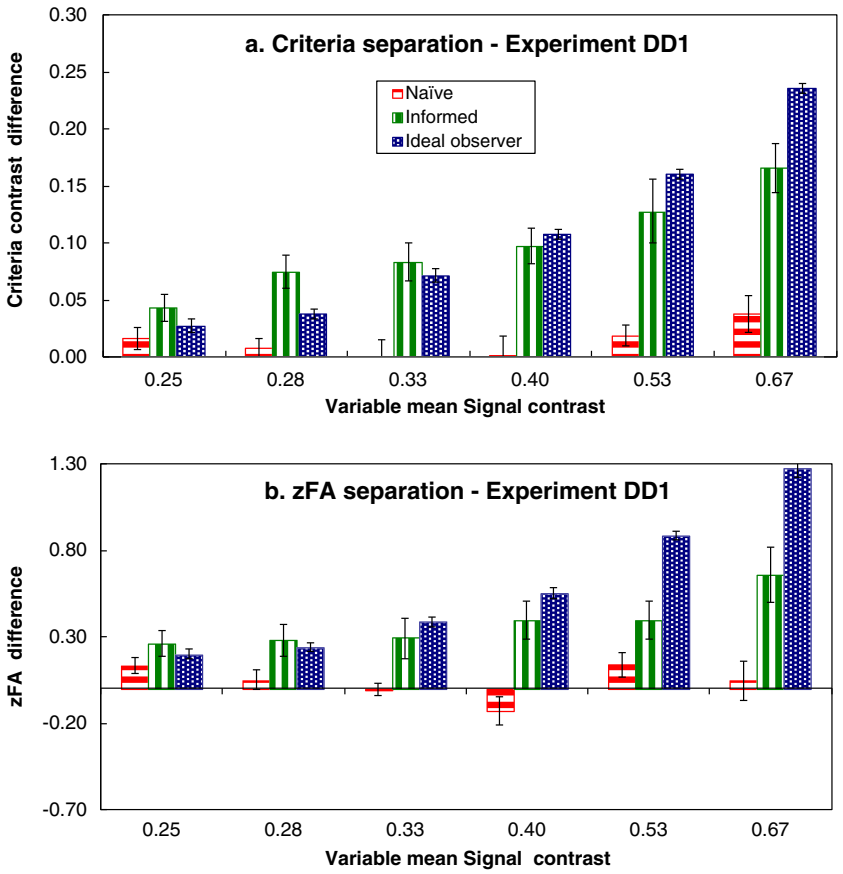

Fig. 7 Criteria (**a**) and *z*FA (**b**) differences for each of the six paired tasks of Experiment DD1 (fixed external noise, variable *d'*). Differences were separately averaged for naïve (horizontal red stripes) and informed (vertical green stripes) observers and are displayed together with the ideal observer differences (dotted blue bars). Error bars are ±1 *SE*

noise (σ_{EX}), leading to a constant efficiency of ~.8. The observed correlation between criterion noise and external noise may reflect a direct or indirect dependency. This is so because external noise in Experiment DD2 correlates by design with mean signal amplitude, which is found to correlate with the measured criteria (Fig. 6), which, in turn, correlate with criterion variability.

Discussion

The present study focused mainly on the causes of observers' nonoptimal decisional behavior when faced with two simultaneous tasks involving equally probable signal and noise trials. On the basis of their results obtained with fixed, close to threshold stimulations (where performance is exclusively limited by internal noise), Gorea and Sagi (2000, 2001, 2002) concluded that observers use a close to identical criterion across the two tasks. This conclusion was challenged by Kontsevich et al. (2002) inasmuch as Gorea and Sagi's criteria equality could not be dissociated from an FA rate equality. This is indeed so inasmuch as the standard deviation of the internal noise, σ_{IN}, is not directly accessible, so that criteria equality—that is, $\sigma_{IN1} \times zFA_1 = \sigma_{IN2} \times zFA_2$—is verified only to the extent that $\sigma_{IN1} = \sigma_{IN2}$. If this were not the case, Gorea and Sagi's equality could be explained by positing that observers equate their FA rates. The basis on which this could be achieved remains, however, obscure. The use of a known external noise significantly larger than the internal noise bypasses this problem, since it allows direct access to the performance-limiting noise. Moreover, the use of external noise also allows the direct assessment of the noise limiting observers' criterion placement, which consists of an unknown combination of internal and decision noise.

Most important, the direct assessment of criterion values in stimulus space precludes the resort to assumptions on the internal-noise distribution required by the standard (internal-noise-based) SDT analysis. There are indeed claims in the literature that the criterion shift observed in standard SDT experiments, where the internal noise is dominant, could in fact reflect an internal-noise dependence on the base rates of (or payoffs associated with) the two alternatives (Balakrishnan 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2002, 2008). In our experiments (where the base rate is always 0.5), the ideal observer's criterion is at the crossover of the external-noise (signal and noise) distributions, but our results clearly show this not to be the case for the measured criteria that show deviations from optimality of up to a factor of 4 (at the higher external-noise levels). Assuming a monotonic relationship between stimulus intensity and internal response, we can conclude that, in our experimental setup, response criteria

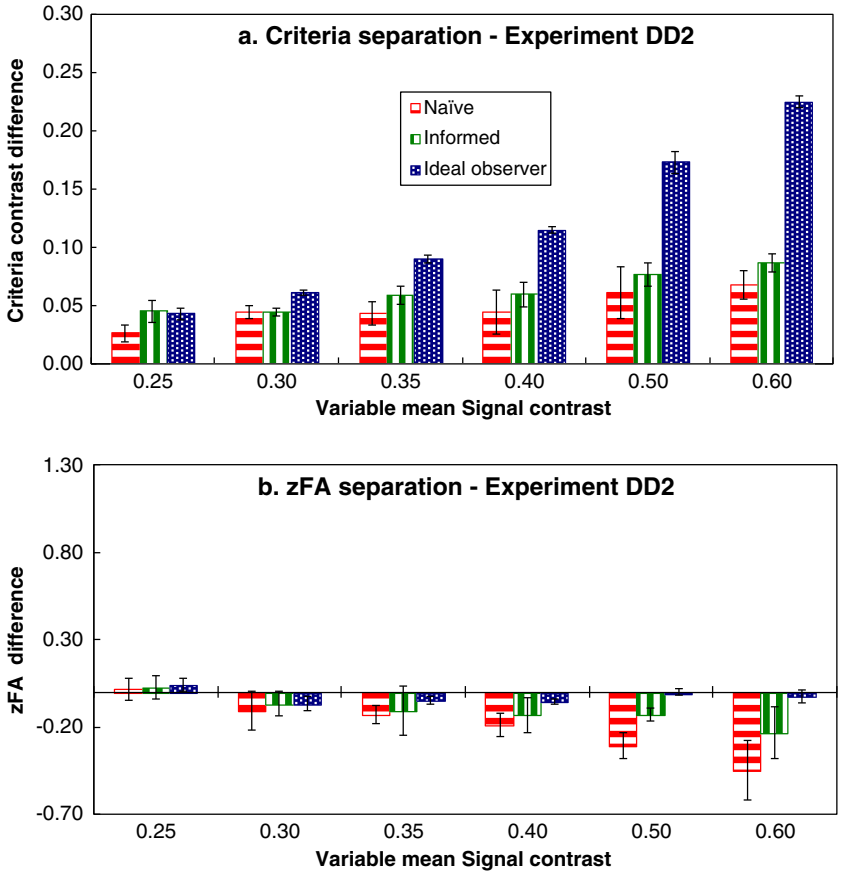

Fig. 8 Criteria (**a**) and *z*FA (**b**) differences for each of the six paired tasks of Experiment DD2 (variable external noise, fixed d'). All details are as for Fig. 7

are not always situated at the intersection of the two distributions corresponding to the two available stimulus sets.

The use of a dual task with identical S and N external distributions in Experiment DS (dual-same conditions) first confirmed that observers do, indeed, use close to optimal criteria when they deal with identical concurrent stimuli. Experiment DD1 (dual-different conditions with equal external noise across the two paired tasks) generalized Gorea and Sagi's results (for "naive" observers; see below) under external-noise conditions. By design, however, Experiment DD1 could not dissociate between observers equalizing their criteria or their FA rates. Experiment DD2 (dual-different conditions with unequal external noises across the two paired tasks) was designed to this end, with the rather clear outcome that observers tended to equalize their criteria, rather than their FA rates. This was so even though this experiment was designed to yield optimal performance for equal FA rates across the paired tasks.

The results of Experiment DD1 show marked differences between "informed" and "naive" observers. The distinction between the two subpopulations was based on their being or not being trained psychophysicists (2 observers, including one author) or on having or not having been introduced to the experimental design and to general SDT principles. Both groups were highly experienced with the task, having carried out hundreds of trials in the experiments. Observers were able to set a close to optimal criterion in the DS condition, but while naive observers practically equalized their criteria (and FA rates) throughout the S contrast range used in this experiment (DD1), "informed" observers used distinct, close to optimal criteria. In contrast, prior psychophysical experience or training of originally naive observers only marginally helped them keep their criteria apart in Experiment DD2, with 1 observer (O.G.) actually worsening his decisional performance.

The case of Experiment DD1, where experts' decisional behavior was close to optimal, is to be contrasted with Gorea and Sagi's (2000, 2001, 2004) results showing non-optimal behavior even for trained psychophysicists. If criteria are represented as internal response values separating the two (yes/no) behavioral responses, and if these values are separated by more than, say, one unit of internal noise (which was not the case in Gorea and Sagi's threshold experiments), they should be easily categorized and, hence, kept apart. This should have been the case for both DD experiments insofar as observers explicitly used such a rule-based response strategy (e.g., Ashby & Maddox, 1998, 2005; Spiering & Ashby, 2008). On this logic, trained observers' failure to keep apart their response criteria in Experiment DD2 (even though an optimal behavior requires that these criteria be significantly more separated than observers' internal noise) suggests that the limiting factor in keeping criteria apart is criterion noise per se (i.e., a combination of

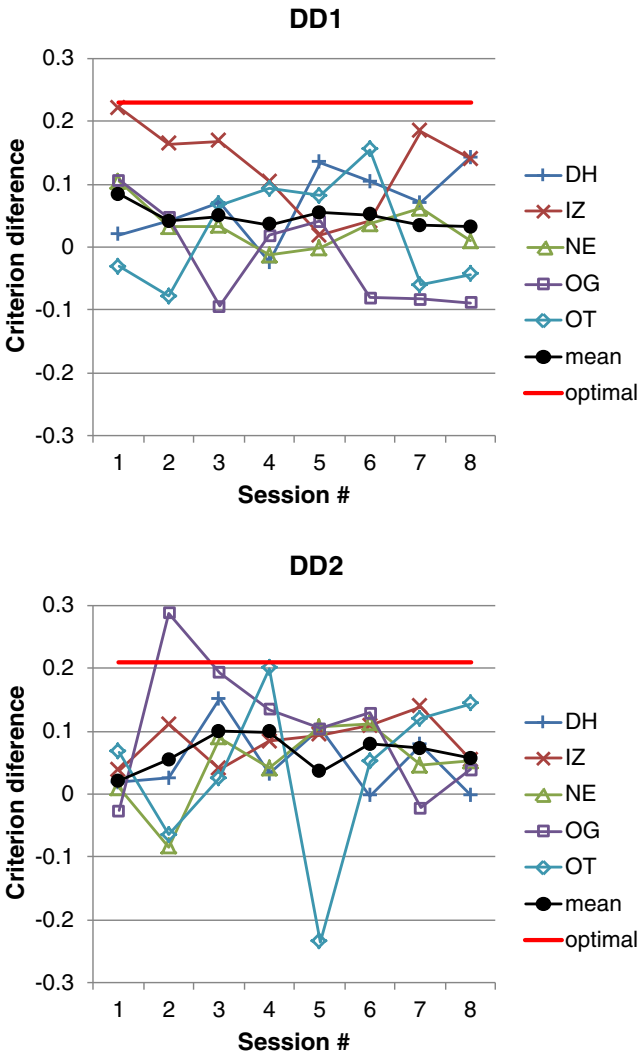

Fig. 9 Criteria differences of the 5 observers (different colors) across the eight sessions for the largest target contrast conditions (condition F in Tables 2 and 3 in the Appendix) in Experiments DD1 and DD2. Black solid circles and lines show the mean (across observers) criteria differences. The straight horizontal line denotes the expected criteria differences had observers used optimal criteria. Data for naive observers are shown before the training stage; observer I.Z. is one of the authors

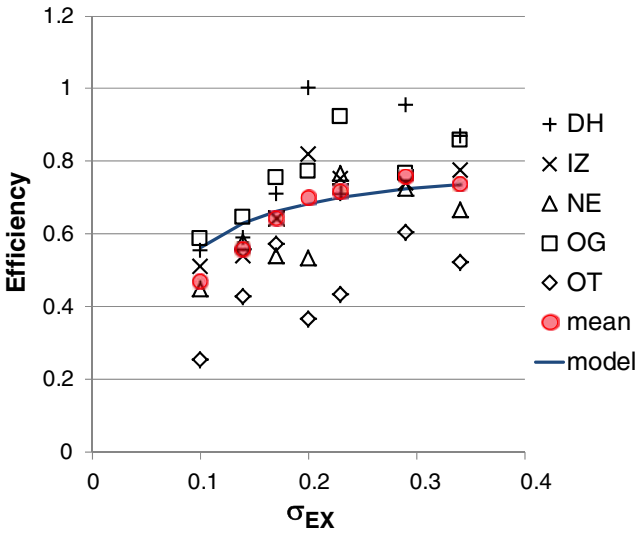

Fig. 10 Individual (5 observers; crosses, *x*s, asterisks, triangles, and diamonds), mean (solid red circles), and predicted (curve) efficiencies as a function of external-noise level (σ_{EX}). Model predictions assume that performance is limited by external noise and the measured criterion noise (see the text for details)

internal and decision noise). Because criterial noise was found here to depend on external noise (i.e., σ_{EX}) that was larger in Experiment DD2 (for the three largest σ_{EX} conditions) than in Experiment DD1, one would expect, indeed, that observers be less capable of keeping their criteria apart.

An alternative although not exclusive account of the fact that, overall, observers are less efficient in keeping track of separate criteria in larger external-noise environments is based on the scaling property of the presently used normal distributions: A given criterion shift (in stimulus units) entails fewer percent correct losses with larger than with lesser σ_{EX}. For example, the largest mean signal contrasts in Experiments DD1 and DD2 were associated with external noises of 0.2 and 0.34, respectively. For these two cases, a shift from the optimal criterion departure from optimality of 0.15 entails a correct response loss of 5% and 2.4%, respectively. Inasmuch as keeping track of more than one criterion has its own cost, subjects might more easily give up with this extra effort for a lesser cost.

The use of externally distributed stimulus samples from bivariate distributions allows the direct assessment of criterion noise (i.e., the slope of the "yes" response function of stimulus contrast). Criterion noise includes internal (or coding) noise and decisional noise per se. In the present experiments, the measured criterion noise was 1.6 times the observers' JNDs. A JND yielding a d' of 1 (as derived from yes/no experiments) represents, in theory, the sum of the internal and decision noise variances. Inasmuch as the present JND and the external-noise classification experiments have been run at similar adaptation levels, it can be assumed that they were associated with equal internal-noise levels. If so, the notable difference between the presently assessed JNDs and criterial noise should be attributed mainly to a decision noise increase in the external-noise classification experiments. This is most likely due to the spread of the external noise per se, which correlates negatively with one's capacity of extracting its statistical properties (e.g., Lee & Janke, 1964).

Conclusion

The presently used external-noise method offers a direct means of estimating observers' criteria and their associated noise in binary classification tasks. The method reveals a close to optimal decision behavior in single but not in dual classification tasks designed so as to entail different optimal criteria across

tasks. In the latter case, the method revealed that observers tend to minimize the distance between their concurrent internal-response criteria, rather than between the respective FA rates.

Prior psychophysical training, as well as information given to naive observers on the experimental design and on SDT principles, entails a decisional behavior closer to optimal under some experimental conditions. This suggests that observers can take advantage of explicit rules for placing multiple criteria in concurrent tasks.

Acknowledgments This work was supported in part by the Basic Research Foundation, administered by the Israel Academy of Sciences and Humanities (D.S.). A.G. was supported by Grant ANR-06-NEURO-042-01.

Appendix

A. Parameters used in the different experiments: signal means and standard deviations relative to background intensity (contrast units) and theoretical *d′* values.

Table 1 Experiment DS (dual-same)

Mean Signal	σ_{EX}	*d′*
0.20	0.20	1.00
0.60	0.34	1.76
0.25	0.20	1.25
0.28	0.20	1.40
0.33	0.20	1.65
0.40	0.20	2.00
0.53	0.20	2.65
0.67	0.20	3.35
0.17	0.10	1.70
0.25	0.14	1.79
0.30	0.17	1.76
0.35	0.20	1.75
0.40	0.23	1.74
0.50	0.29	1.72

Table 2 Experiment DD1 (fixed external noise)

Condition Label	Fixed Signal (red cue)			Variable Signal (green cue)		
	Mean	σ_{EX}	*d′*	Mean	σ_{EX}	*d′*
A	0.2	0.2	1	0.25	0.2	1.25
B				0.28		1.40
C				0.33		1.65
D				0.40		2.00
E				0.53		2.65
F				0.67		3.35

Table 3 Experiment DD2 (variable external noise)

Condition label	Fixed Signal (red cue)			Variable Signal (green cue)		
	Mean	σ_{EX}	*d′*	Mean	σ_{EX}	*d′*
A	0.17	0.10	1.70	0.25	0.14	1.79
B				0.30	0.17	1.76
C				0.35	0.20	1.75
D				0.40	0.23	1.74
E				0.50	0.29	1.72
F				0.60	0.34	1.76

B. Dependence of observers' responses to the test stimulus (cued location) on the contrast of the "distracting" stimulus (noncued location).

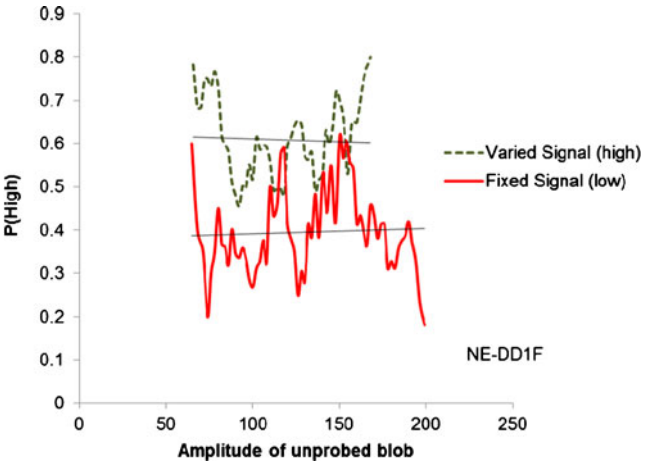

Fig. 11 Dependence of the fraction of "high" reports on the contrast of the unprobed blob illustrated for one of the naive observer's data (observer N.E.) in Experiment DD1, condition F (higher amplitude in Table 2). Two separate curves are shown, for the fixed signal (continuous curve, red) and for the varied signal (dashed curve, green). Both curves are practically flat (slope < 0.0001; $r^2 < .002$). The vertical displacement of the two curves, below and above *P*(high) = .5, corresponds to biases due to the merged criteria, as seen in Fig. 6 for observer N.E. in experiment DD1

References

Ashby, F. G., & Maddox, W. T. (1998). Stimulus categorization. In M. H. Birnbaum (Ed.), *Handbook of perception & cognition: Judgment, decision making, and measurement* (pp. 251–301). San Diego: Academic Press.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178.

Balakrishnan, J. D. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors, 40,* 601–623.

Balakrishnan, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods, 3,* 68–90.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of*

Experimental Psychology: Human Perception and Performance, 25, 1189–1206.

Balakrishnan, J. D., & MacDonald, J. A. (2002). Decision criteria do not shift: Reply to Treisman. *Psychonomic Bulletin & Review, 9,* 858–865.

Balakrishnan, J. D., & MacDonald, J. A. (2008). Decision criteria do not shift: Commentary on Mueller and Weidemann (2008). *Psychonomic Bulletin & Review, 15,* 1022–1034.

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

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to Algorithms* (2nd ed.). Cambridge, MA: MIT Press.

Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences, 97,* 12380–12384.

Gorea, A., & Sagi, D. (2001). Disentangling signal from noise in visual contrast discrimination. *Nature Neuroscience, 4,* 1146–1150.

Gorea, A., & Sagi, D. (2002). False alarms? A reply to Kontsevich et al. *Nature Neuroscience, 5,* 707–708.

Gorea, A., & Sagi, D. (2004). On decision and attention. In L. Itti, G. Rees, & J. Tsotsos (Eds.), *Neurobiology of attention* (pp. 152–159). Amsterdam: Elsevier.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Kontsevich, L. L., Chen, C. C., Verghese, P., & Tyler, C. W. (2002). The unique criterion constraint: A false alarm? *Nature Neuroscience, 5,* 707.

Lee, W., & Janke, M. (1964). Categorizing externally distributed stimulus samples for three continua. *Journal of Experimental Psychology, 68,* 376–382.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Neter, J., Wasserman, W., & Kutner, M. H. (1983). *Applied linear regression models*. Homewood, IL: Irwin.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: The principles and practice of statistics in biological research* (3rd ed.). San Francisco: Freeman.

Spiering, B. J., & Ashby, F. G. (2008). Response processes in information-integration category learning. *Neurobiology of Learning and Memory, 90,* 330–338.

Treisman, M. (2002). Is signal detection theory fundamentally flawed? A response to Balakrishnan (1998a, 1998b, 1999). *Psychonomic Bulletin & Review, 9,* 858–865.

Weidemann, C. T., & Mueller, S. T. (2008). Decision noise may mask criterion shifts: Reply to Balakrishnan and MacDonald (2008). *Psychonomic Bulletin & Review, 15,* 1031–1034.