Classification images reveal decision variables and strategies in forced choice tasks

Lisa M. Pritchett and Richard F. Murray[1]

Department of Psychology and Centre for Vision Research, York University, Toronto, ON, Canada M3J 1P3

Edited by Brian A. Wandell, Stanford University, Stanford, CA, and approved May 5, 2015 (received for review November 24, 2014)

Despite decades of research, there is still uncertainty about how people make simple decisions about perceptual stimuli. Most theories assume that perceptual decisions are based on decision variables, which are internal variables that encode task-relevant information. However, decision variables are usually considered to be theoretical constructs that cannot be measured directly, and this often makes it difficult to test theories of perceptual decision making. Here we show how to measure decision variables on individual trials, and we use these measurements to test theories of perceptual decision making more directly than has previously been possible. We measure classification images, which are estimates of templates that observers use to extract information from stimuli. We then calculate the dot product of these classification images with the stimuli to estimate observers' decision variables. Finally, we reconstruct each observer's "decision space," a map that shows the probability of the observer's responses for all values of the decision variables. We use this method to examine decision strategies in two-alternative forced choice (2AFC) tasks, for which there are several competing models. In one experiment, the resulting decision spaces support the difference model, a classic theory of 2AFC decisions. In a second experiment, we find unexpected decision spaces that are not predicted by standard models of 2AFC decisions, and that suggest intrinsic uncertainty or soft thresholding. These experiments give new evidence regarding observers' strategies in 2AFC tasks, and they show how measuring decision variables can answer long-standing questions about perceptual decision making.

vision | psychophysics | classification images | signal detection theory | decision making

Many current questions about human cognition are related to how people make decisions, including decisions based on perceptual information. For example, how do we decide whether a search target is present in a cluttered display? How do we decide when to respond in a task where both speed and accuracy are important? How do we judge which of two signals is present in a discrimination task?

Most theories of perceptual decision making rely on the notion of a decision variable, a quantity that the observer calculates from the stimulus to summarize task-relevant information, e.g., the probability that a faint signal is present in a detection task (1). Some theories of decision making are very simple, e.g., the observer gives one response if the decision variable is greater than a fixed criterion, and another response if the decision variable is less than the criterion. Other theories use more complex decision rules. Testing theories of decision making would be much easier if we had access to observers' decision variables, but these are usually thought of as theoretical constructs that cannot be measured psychophysically. Here we show that in some tasks, it is possible to estimate decision variables on individual trials, and this provides a very direct way of testing theories of perceptual decision making. We use this method to examine the long-standing question of how people make decisions in two-alternative forced choice (2AFC) tasks.

Proxy Decision Variables

Our approach relies on the linear template model that has been shown to account for performance in many simple discrimination tasks (2, 3). In this model, the decision variable is the dot product of a template with the stimulus, plus a sample of normally distributed internal noise. Previous work has shown that we can estimate an observer's template by measuring a "classification image," which is a map that shows the impact of small luminance fluctuations in each region of the stimulus on the observer's responses (4–6). A stimulus region that has a large effect on the observer's responses has a large value in the classification image, and a stimulus region that has little or no effect has a small value. Thus, a classification image shows what stimulus regions an observer uses to perform a task, and eliminates the need to make one type of assumption about how the observer uses the stimulus.

Motivated by the linear template model, we estimate observers' decision variables by taking the dot product of the classification image with the stimulus on each trial. We call the result of this dot product a "proxy decision variable." The proxy decision variable is an informative but imperfect estimate of the true decision variable, and we return to this point after discussing theories of decision making in 2AFC tasks.

Decision Making in 2AFC Tasks

In a 2AFC task, the observer views two signals in random order and judges which order was shown (7). This task has played an important role in perception research for over 60 years. It has often been used as a method of reducing observer bias, but more importantly, it has also been a testing ground for theories of perceptual decision making.

The classic model of 2AFC decisions is the "difference model." This model assumes that the observer calculates a decision variable from each of the two stimulus intervals and makes a response based on which decision variable is greater (7). A useful tool for understanding such decision strategies is the decision space, a map that shows the probability of the observer's responses for all values of the decision variables (8). The difference model implies that the observer's decision space is divided into two response regions by a diagonal line (Fig. 1A): the

Significance

Signal detection theory is a classic theory of perceptual decision making. This theory states that when people make a decision based on visual or auditory information, they calculate a "decision variable" that encodes their estimate of the probability of each possible response being correct. Until now, there has been no way to measure decision variables behaviorally. We describe a method of estimating decision variables, and we show that it is a highly effective way of revealing peoples' decision strategies. This creates a new approach to answering long-standing questions about perception and decision making.

Author contributions: L.M.P. and R.F.M. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: rfm@yorku.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1422169112/-/DCSupplemental.

observer gives one response if the first decision variable is greater, and the other response if the second decision variable is greater, so the dividing line, or decision line, is $y = x$.

Alternative models of 2AFC decisions have also been proposed. According to the "double detection model," the observer makes independent decisions about which signal was shown in the first stimulus interval and which was shown in the second interval (9, 10). When the observer judges that both intervals contain the same signal (which does not happen in a 2AFC task), the observer guesses. Under this model, the decision space is divided into four quadrants (Fig. 1B). According to the "single-interval model," the observer simply ignores one interval, and the decision space is divided by a vertical or horizontal line (Fig. 1C). In the "difference model with guessing," the observer compares decision variables from the two stimulus intervals, as in the difference model, but if the decision variables differ by less than a threshold amount, then the observer guesses (11). Here the decision space is divided into three regions by two diagonal lines (Fig. 1D).

Previous studies have tested these models by comparing observers' performance under various conditions in 2AFC tasks and other designs (9–11). However, different tasks put different demands on poorly understood factors such as attention and memory, so these comparisons can be difficult to interpret (12). A recent review of studies on the 2AFC task concludes that we know very little about how people actually make 2AFC decisions, and that the standard theory of 2AFC tasks, including the difference rule, has little experimental support (10). Here we take a new approach to the problem of understanding peoples' strategies in 2AFC tasks: We measure proxy decision variables in the two stimulus intervals over thousands of trials, and we use these measurements to reconstruct observers' decision spaces.

Proxy Decision Space

The proxy decision variable is the dot product of an observer's classification image with the stimulus. This gives an imperfect estimate of the true decision variable for at least two reasons: The classification image is an imperfect estimate of the template (3–6), and the observer has internal noise (1). In SI Text, Properties of the Proxy Decision Space, we show that both these factors imply that the proxy decision variable is equal to the true decision variable plus a normal random variable that represents measurement error. The "proxy decision space" is a map that shows the probability of the observer's responses for all values of the proxy decision variables. In SI Text, Properties of the Proxy Decision Space, we also show that the measurement error in the proxy decision variables implies that the proxy decision space does not have sharp edges, as in Fig. 1 A–D, but instead is the true, sharp-edged decision space convolved with a Gaussian kernel whose scale constant is the standard deviation (SD) of the measurement error. For instance, an observer who uses the difference rule will produce a blurred proxy decision space, as in Fig. 1E, instead of a sharp-edged space, as in Fig. 1A. For this reason, when fitting a model to a proxy decision space, we require an SD parameter σ_{blur} that controls the amount of blurring, in addition to parameters that control the position and orientation of the decision lines.

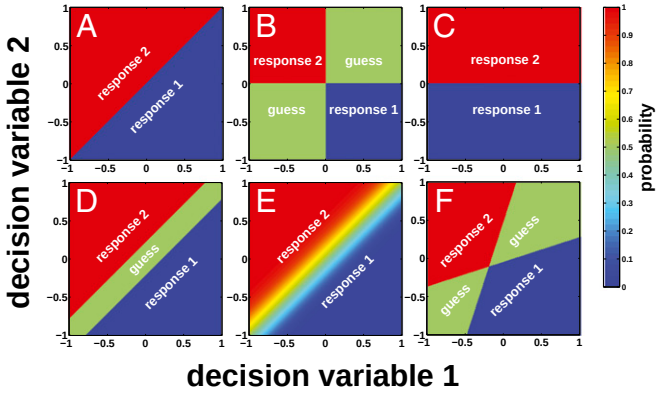

Fig. 1. Decision spaces for 2AFC decision models. Color encodes the probability of the observer's responses. (A) Difference model. (B) Double detection model. (C) Single-interval model. (D) Difference model with guessing. (E) Proxy decision space for the difference model, with blur due to internal noise. (F) GDD function.

Experiment 1

Task. In our first experiment, three observers discriminated between black and white Gaussian profile disks in noise, at fixation, with two 500-ms stimulus intervals separated by a blank 1,000-ms interstimulus interval. Fig. S1 and Movie S1 show typical stimuli. Each observer ran in 9,900 trials. We calculated each observer's classification image and took its dot product with the two stimulus intervals on every trial. We constructed each observer's proxy decision space by finding the probability of the observer responding "black disk first" for all values of the two proxy decision variables. The data from both experiments reported in this paper and MATLAB code implementing all our analyses are available at purl.org/NET/rfm/pnas2015.

Results. Fig. 2A shows the resulting proxy decision spaces, which are divided diagonally, consistent with the difference model or the difference model with guessing (Fig. 1 A and D). Model selection via 10-fold cross validation supports this observation (Fig. 3). For all observers the cross-validation error was the same for the difference model and the difference model with guessing, and significantly lower than for the double detection and single-interval models ($P < 0.05$ for within-observer, independent samples t tests between the means of the difference model and the double detection model, and between the means of the difference model and the single-interval model, Bonferroni corrected for six comparisons). The difference model performs as well as the difference model with guessing, but with one less parameter, which supports the difference model as the better model of performance on this task.

As another test of the four models (Fig. 1 A–D), we fitted each model to each observer's decision space (Fig. 2A) and calculated the Akaike information criterion (AIC) of the fits (13). AIC evaluates goodness of fit in a way that penalizes larger numbers of parameters. For all three observers, the difference model had the lowest AIC (see Table S1), showing again that it gave the best account of observers' data, consistent with our cross-validation results.

As a further test for a guessing region, we collapsed the decision spaces parallel to the fitted decision line of the difference model, which is a line approximately halfway between the two red lines in each observer's panel in Fig. 2A (to be discussed below). The difference model predicts that the transition between response regions follows a normal cumulative distribution function, whereas the difference model with guessing predicts a flattening of the curve in the guessing region where the two decision variables are approximately equal. Fig. 4 shows that the normal cumulative distribution function (blue line) gives an excellent fit, with no flattened interval apparent for any of the observers. To find confidence intervals for the width of the guessing region, we bootstrapped fits of the difference model with guessing to the data shown in Fig. 4. For the panels from left to right, the maximum likelihood estimates and 95% confidence intervals for the width of the guessing region were 0.00 (0.00, 0.29), 0.00 (0.00, 0.25), and 0.00 (0.00, 0.00). We conclude that there is little or no role for a guessing region in these observers' decision rules.

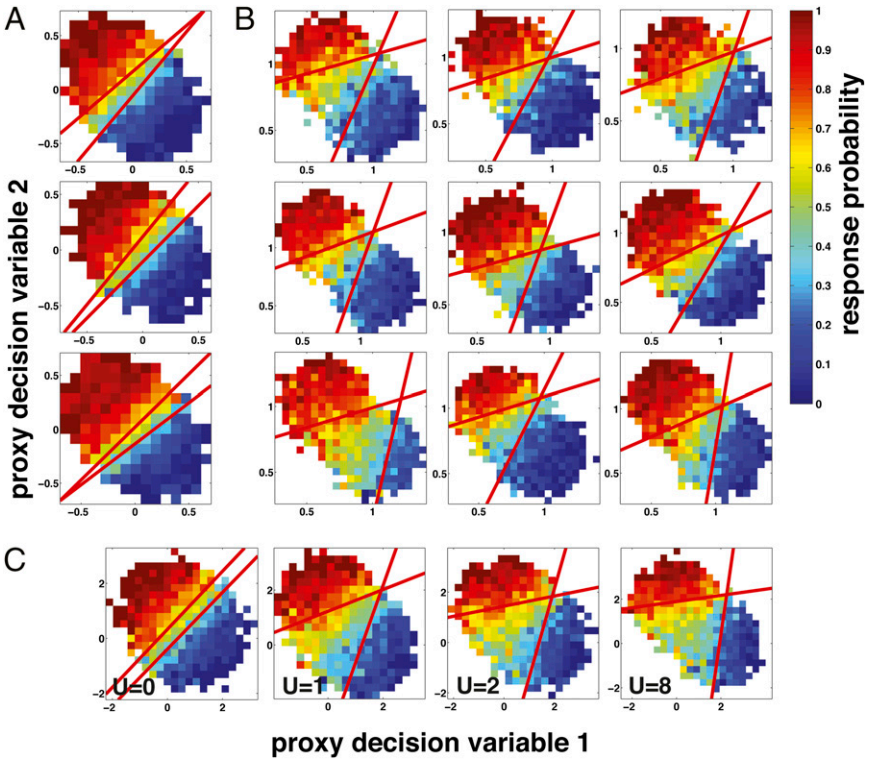

Fig. 2. Proxy decision spaces. Each panel shows results from a single observer, based on ▫10,000 trials. The axes are the proxy decision variables for the two stimulus intervals, and the plots show the response probability as coded in the color bar. Red lines are maximum likelihood fits of the GDD function. (A) Probability of a "black disk first" response in the black/white discrimination task. (B) Probability of a "right disk brighter" response in the contrast increment detection task. (C) Probability of a "signal in interval 2" response from simulated model observers with intrinsic uncertainty. U is the number of irrelevant mechanisms that the model observer monitors in each stimulus interval.

In fact, even these relatively small bootstrapping estimates overstate the evidence for a guessing region in our data. Fig. S2 (blue line) plots response probability as a function of the distance from the center of the guessing region, according to the difference model with guessing with a guessing region 0.30 units wide (the upper limit of the bootstrapped 95% confidence intervals reported above) and an SD parameter $\sigma_{blur} = 0.25$ (a typical fitted value). With a guessing region this size, the response probability smoothly transitions from one side of the decision line to the other, with no apparent flattening in the middle. This occurs because, as we explained in the Introduction, the proxy decision space is a blurred version of the true decision space, and so a guessing region that is small relative to the amount of blur is effectively blurred out of the proxy decision space. Indeed, Fig. S2 (dashed green line) shows that the difference model with an SD parameter of $\sigma_{blur} = 0.30$ gives a psychometric function that is practically identical to the psychometric function with a guessing width of 0.30 (blue line), without the need for an additional guessing width parameter.

In Monte Carlo simulations, we generated artificial data from the difference model, and we fitted the difference model with guessing to these data. We matched the parameters of the simulated difference model and the number and distribution of trials to the psychometric function shown in Fig. 4, Left. These simulations assigned the guessing region width a 95% confidence interval of (0.00, 0.30), simply due to the randomness in the simulated observer's responses. This confidence interval is similar to those we found for two of three human observers, and the third observer's confidence interval was even smaller. (We include MATLAB code for this simulation in the code posted at purl.org/NET/rfm/pnas2015.)

In summary, our bootstrapping results show that Fig. 2A is consistent with a guessing region up to 0.30 units wide, but our cross-validation and AIC tests show that the difference model gives a better account of the data, and our Monte Carlo simulations show that the 95% confidence intervals we found for human observers' guessing regions are no larger than one would expect

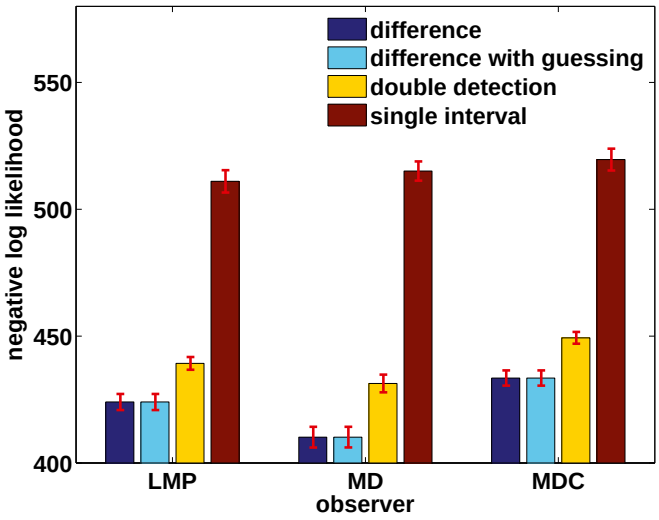

Fig. 3. Results of 10-fold cross-validation. The y axis shows the negative log likelihood of the observer's responses on validation trials, averaged across the 10 cross-validation blocks. Error bars show the SEM. The single-interval bars show results averaged across a model that used only the first interval and a model that used only the second interval; the results for these two single-interval models were practically identical.

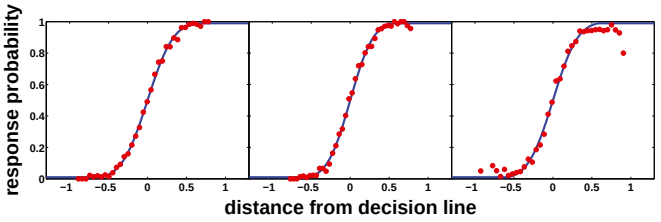

Fig. 4. A test of the difference model with guessing. Each panel shows the probability of an observer making response 2 as a function of the distance of the proxy decision variable pair from the decision line of the difference model fitted to the data in Fig. 2A. The decision lines of the difference model are not shown in Fig. 2A but are approximately halfway between the two red GDD lines in each panel. Here, there is no flattening of the curve near the zero point on the x axis, indicating little or no guessing region. The blue lines are maximum likelihood fits of the normal cumulative distribution function. We have grouped the distances from the decision line into 50 bins, so there are 50 data points in each plot. When fitting models to this data, we used the raw, unbinned distances from the decision line on individual trials, which are continuously distributed.

from fitting the difference model with guessing to approximately 10,000 trials from an observer who follows the difference model.

Do all four models discussed above (Fig. 1 A–D) fail to describe important features of the decision spaces? To test this possibility, we defined a new function that flexibly divides decision spaces into response regions, and that has the four models in Fig. 1 A–D as special cases. The "generalized double detection (GDD) function" divides the decision space with two decision lines at arbitrary orientations and positions (Fig. 1F). When the two decision lines indicate the same response, the decision space has probability 1.0 for that response, and when they indicate different responses, the decision space has a response probability near 0.5, indicating a random guess. Suitably arranged, the two decision lines can produce any of the decision spaces illustrated in Fig. 1 A–D. We prefer not to call the GDD function a model, because we do not think it is a plausible account of how observers make decisions, but rather a flexible tool for exploring decision spaces, much as one might fit a spline to data in the xy plane. Maximum likelihood fits of the GDD function to the proxy decision spaces (Fig. 2A, red lines) support the difference model, possibly with a small guessing region, as the best explanation of observers' responses. As we showed above, cross-validation, AIC tests, and bootstrapped confidence intervals show that the guessing region is small or nonexistent (Fig. 3).

These findings rely on a standard method of calculating classification images that has been developed assuming the difference model (6), and this may seem inappropriate given that here we are questioning the difference model. In SI Text, Properties of the Proxy Decision Space, we show that this method of calculating classification images gives unbiased estimates of the template for a much wider range of decision rules than has previously been shown, including the four decision rules illustrated in Fig. 1 A–D.

Experiment 2

Task. Many 2AFC experiments show stimuli separated in space instead of time, which places different demands on memory and attention. To test the robustness of our findings, we examined a second experiment where nine observers detected a contrast increment in one of two disks located to the left and right of fixation and shown in noise (4). Fig. S1 and Movie S2 show typical stimuli. Each observer ran in ▫10,000 trials.

Results. Fig. 2B shows the resulting proxy decision spaces. Surprisingly, the GDD fits (red lines) do not correspond to any of the four models discussed so far. Instead, they consistently show a triangular guessing region, narrow at high values of the decision variables and wide at low values. Across observers, the average orientation of the bisector line (the line, not shown, midway between the two GDD lines) is 44° relative to the x axis, with an SD of 4°. The average angle between the two GDD lines is 49°, with an SD of 8°.

What can this mean? These fits show that the decision spaces are again divided diagonally, but now the transition between response regions is more gradual at lower values of the decision variables. That is, the difference between two low-valued decision variables must be relatively large before the observer can make a reliable response. This is consistent with an intrinsic uncertainty model where the observer monitors the relevant stimulus locations and also a number of irrelevant locations (14). In this model, the noise from irrelevant locations interferes with weak signals more than with strong signals. To test this explanation of the triangular GDD fits, we simulated a template-matching model observer that follows the difference model but is uncertain about the signal location. The model observer's decision variable in each stimulus interval is the maximum of the template response at the stimulus location and at some number U of nonoverlapping irrelevant locations (U = 0, 1, 2, or 8; see SI Text, Uncertain Observer Simulations for details). Fig. 2C shows that even a small amount of uncertainty produces triangular GDD fits much like those from human observers. Other possible explanations of poor discrimination at low decision variables are thresholding of weak signals (15) or higher internal noise for weak signals. Our results are qualitatively consistent with the difference model plus any of these mechanisms, which illustrates the fact that although decision spaces show the relationship between stimuli and responses, they do not always uniquely identify the mechanism that underlies this relationship.

Discussion

Proxy decision variables are useful for testing theories of visual processing because they measure the task-relevant information that is available to the observer on individual trials. A traditional analysis of experiments like ours would record the signal contrast and the observer's response on each trial, so, for example, one could plot the observer's proportion of correct responses at each signal level. The present method exploits trial-to-trial fluctuations in the external noise to calculate two proxy decision variables on each trial, one from each stimulus interval, and finds the probability of the observer giving one response or the other for all combinations of the proxy decision variables. Instead of just the nominal signal contrast, we record a more precise estimate of the task-relevant information in each stimulus interval. Thus, we obtain a 2D function (the proxy decision space) that describes the observer's decision strategy, instead of a one-dimensional function such as a traditional psychometric function. As we have shown, this 2D function can be highly effective for testing theories of decision making.

Our findings clearly rule out the single-interval model and the double detection model as theories of 2AFC discrimination in experiment 1. The difference model is a special case of the difference model with guessing, so the choice between these two models cannot be as decisive; it is always possible that observers have a guessing region that is too small to be detected with the available data. A more useful approach is to test what size of guessing region is consistent with the data. For all three observers in experiment 1, the maximum likelihood estimate of the width of the guessing region was zero, and the 95% confidence intervals were relatively small. Furthermore, cross-validation and AIC tests showed that the additional parameter in the difference model with guessing gave it no significant benefit over the difference model. Altogether, the difference model gives the best account of observers' behavior in experiment 1, as it is the simpler model and there is little or no need to assume a guessing region.

The proxy decision spaces we measured in experiment 2 did not match the predictions of any of the four classic 2AFC models, and this is one of the most interesting findings of our experiments. The proxy decision spaces were roughly symmetric around the 45° diagonal line, similar to the decision spaces of the difference model and the difference model with guessing, which suggests that observers followed some variant of these models. Model observer simulations showed that a similar proxy decision space is produced by a difference model that is limited by a factor such as intrinsic uncertainty that worsens performance at low signal levels. The difference model with guessing, limited by intrinsic uncertainty, would presumably produce a similar decision space. In experiment 2, it is difficult to choose between the difference model and the difference model with guessing, because uncertainty worsens performance in a triangular region that covers the diagonal region where the difference model with guessing predicts that a guessing region will appear (Fig. 1D).

Furthermore, some care is necessary in interpreting experiment 2, as intrinsic uncertainty does not fit neatly into our modeling tools. The linear observer model does not capture spatial uncertainty, because an uncertain observer monitors many signal detection mechanisms simultaneously, e.g., a spatially uncertain observer monitors many spatial locations for the signal. This means that the observer effectively uses multiple templates (one for each location), and one consequence of this is that classification images give a blurred estimate of the observer's template (16). Experiment 1 led to clear and easily interpretable findings. Experiment 2 is less decisive, but it shows that the proxy decision variable method is open-ended and can reveal unexpected properties of observers' decision strategies, much like the classification image method that it builds on.

The literature on visual perception shows that in many tasks, observers are unable to use an optimal strategy despite extensive practice (e.g., refs. 14 and 17), and we should similarly expect that observers' decision spaces sometimes will be optimal and sometimes will not. The difference model is the optimal strategy in a 2AFC task with Gaussian noise (1). Experiment 1 gave observers an excellent chance to use or learn this optimal strategy: Observers ran in thousands of trials, in a simple foveal task, with auditory feedback on every trial. We found that in this experiment, observers did use the optimal strategy. However, observers did not use the optimal strategy in experiment 2, which was different in seemingly minor ways: The stimuli were separated in space instead of time, they were just 0.5 degrees to the left and right of the fovea, and observers judged which stimulus had a brightness increment instead of which was black or white. Thus, even small changes in a task can produce qualitative differences in behavior. There are currently no general rules for predicting decision strategies, and they will need to be investigated case by case until generalizations become possible.

Beyond testing models of 2AFC decisions, the proxy decision variable method should be useful whenever it would be informative to have trial-by-trial estimates of observers' decision variables. For example, most models of response times are based on decision variables that fluctuate over time, accumulating information until they reach a state that causes the observer to make a response (18). If we show stimuli in dynamic noise, take the dot product of the observer's classification image with each frame of noise, and sum these dot products over successive frames, we may be able to estimate how an observer's decision variable evolves over time on individual trials (compare the approach in ref. 19). To take another example, one influential theory of visual search states that the observer calculates a decision variable from each target or distractor element in the search display and responds "target present" if the maximum of these decision variables exceeds a criterion (20). Proxy decision variables could give estimates of the decision variable from each search element, and these measurements could be used to test the maximum rule model of visual search. These examples illustrate how methods like the one we have presented can be used to test theories that are based on decision variables. Decision variables have previously been thought of as inaccessible theoretical constructs, but our experiments show that, in some circumstances, they are measurable, and these measurements lead directly to new tests of sensory processing and decision making.

Materials and Methods

Experiment 1: Black/White Discrimination Task. Each observer ran in 33 blocks of 300 trials. Each trial showed two 500-ms stimulus intervals separated by a blank 1,000-ms interstimulus interval. Fig. S1 and Movie S1 show typical stimuli. The signals were black and white Gaussian profile disks with scale constant σ = 0.055 degrees of visual angle (°). The black and white signals were randomly assigned to the two stimulus intervals. The observer pressed a key to indicate the order of the signals, and received auditory feedback. On each trial, the white disk had peak luminance +L above the background luminance of 65 cd/m^2, and the black disk had peak luminance −L below the background luminance. The luminance perturbation L was adjusted across trials according to a one-up two-down staircase converging on 71% correct performance (21). The disks were shown in Gaussian white noise: The luminance of each stimulus pixel was randomly perturbed by a value drawn from a normal distribution with mean zero and SD 16.25 cd/m^2. A faint fixation point was shown continuously before and after the stimulus intervals. To minimize spatial uncertainty, a thin white square surrounding the stimulus location was always present on the screen, and there was a small tick in the middle of each side to indicate the center of the square. The stimuli were 0.81° square (31 × 31 pixels). Viewing distance was 1.65 m. In each block, trials 101–150 were repeated as trials 151–200 (that is, the staircase was suspended for trials 151–200, and these trials were exact repetitions of trials 101–150), but we do not examine response consistency in this paper. Stimuli were shown on a Sony Trinitron G520 monitor (512 × 384 resolution, pixel size 0.755 mm, refresh rate 75 Hz). We show results for three observers. A fourth observer's thresholds rose sharply over the course of the experiment, probably due to a loss of motivation; it is not meaningful to analyze all of this observer's trials in a single decision space, so we discarded this data.

Experiment 2: Contrast Increment Detection Task. The data for this experiment are taken from Murray et al.'s (4) experiment 2 (Fig. 2B, Top) and experiment 3 (Fig. 2B, Middle and Bottom). Each observer ran in 50 blocks of 200 trials. Fig. S1 and Movie S2 show typical stimuli. The stimulus showed two disks of radius 0.11° positioned 0.50° to the left and right of a fixation point. The baseline luminance of the disks was 3 cd/m^2 above the background luminance of 30 cd/m^2, and on each trial, one of the disks had a luminance increment. The luminance increment was set to each observer's 70% threshold, based on pilot trials. Threshold luminance increments ranged from 1.2 cd/m^2 to 2.1 cd/m^2 across observers. The observer pressed a key to indicate which disk was brighter, and received auditory feedback. In Murray et al.'s (4) experiments 2 and 3, observers gave rating responses on a six-point scale, and we converted these to two-alternative responses by grouping ratings 1–3 to mean "left disk brighter" and grouping ratings 4–6 to mean "right disk brighter." The stimuli were shown in Gaussian white noise: The luminance of each pixel of the stimulus was randomly perturbed by a value drawn from a normal distribution with mean zero and SD 6 cd/m^2. The stimuli were 1.0° vertical × 2.0° horizontal (38 × 76 pixels). The stimulus duration was 200 ms. Viewing distance was 1.0 m. Stimuli were shown on an AppleVision monitor (640 × 480 resolution, pixel size 0.467 mm, refresh rate 67 Hz). Although the stimulus contrast was set to an estimate of each observer's 70% threshold, performance ranged from 69% to 80% correct across observers.

Proxy Decision Spaces. We calculated each observer's classification image using the weighted sum method for 2AFC tasks (6). To reduce measurement noise, we radially averaged each classification image around the center of the signal (6) and set the classification image to zero at locations far from the center. To calculate proxy decision variables p_1 and p_2 for the two stimulus intervals, we took the dot product of the observer's classification image with the two stimuli on each trial.

We constructed three 20 × 20 matrices, $K = (k_{ij})$, $N = (n_{ij})$, and $P = (p_{ij})$, to represent the proxy decision space, as follows. We chose 20 evenly spaced values, $x_1, \ldots, x_{20}$, that spanned the range of the proxy decision variables in increments of Δx. Each element n_{ij} was set to the total number of trials for which $x_i - \Delta x/2 \le p_1 < x_i + \Delta x/2$ and $x_j - \Delta x/2 \le p_2 < x_j + \Delta x/2$. Each element k_{ij} was set to the number of trials in this range where the observer gave response 2 ("black disk first" in experiment 1, "right disk brighter" in

PSYCHOLOGICAL AND COGNITIVE SCIENCES

experiment 2). Each p_{ij} was set to $k_{ij}=n_{ij}$. In Fig. 2, we show the matrix P rotated 90° counterclockwise.

We used the same trials to calculate classification images and proxy decision spaces, but there is little danger of overfitting the classification images: We used ▫10,000 trials to estimate each small, radially pooled classification image, which had effectively just six or seven free parameters. In any case, we show in SI Text, Properties of the Proxy Decision Space, that any measurement error in the classification images simply increases the blur in the proxy decision spaces.

Modeling. The GDD function has five parameters: θ_1 and δ_1 that control the orientation and position, respectively, of the first decision line; θ_2 and δ_2 that control the second decision line; and γ that controls the probability of response 2 in the guessing regions. According to the GDD function, the probability of response 2 when the decision variables are d_1 and d_2 is

$$D_{GDD}(d_1, d_2; \theta_1, \theta_2, \delta_1, \delta_2, \gamma) = P(R = 2|d_1, d_2) \quad [1]$$

$$D_{GDD}(d_1, d_2; \theta_1, \theta_2, \delta_1, \delta_2, \gamma) = \begin{cases} 1 & \text{if } (d_1, d_2)\bullet(\cos(\theta_1), \sin(\theta_1)) \geq \delta_1 \text{ and } (d_1, d_2)\bullet(\cos(\theta_2), \sin(\theta_2)) \geq \delta_2 \\ 0 & \text{if } (d_1, d_2)\bullet(\cos(\theta_1), \sin(\theta_1)) < \delta_1 \text{ and } (d_1, d_2)\bullet(\cos(\theta_2), \sin(\theta_2)) < \delta_2. \\ \gamma & \text{otherwise} \end{cases} \quad [2]$$

Here, $\bullet$ is the vector dot product. The other models we considered are special cases of this function. The difference model is

$$D_D(d_1, d_2; \delta) = D_{GDD}(d_1, d_2; 3\pi/4, 3\pi/4, \delta, \delta, 0). \quad [3]$$

The double detection model is

$$D_{DD}(d_1, d_2; \delta_1, \delta_2, \gamma) = D_{GDD}(d_1, d_2; \pi, \pi/2, \delta_1, \delta_2, \gamma). \quad [4]$$

The single-interval model for the first interval is

$$D_{S1}(d_1, d_2; \delta) = D_{GDD}(d_1, d_2; \pi, \pi, \delta, \delta, 0). \quad (5)$$

The single-interval model for the second interval is

$$D_{S2}(d_1, d_2; \delta) = D_{GDD}(d_1, d_2; \pi/2, \pi/2, \delta, \delta, 0). \quad (6)$$

The difference model with guessing is

$$D_{DMG}(d_1, d_2; \delta_1, \delta_2, \gamma) = D_{GDD}(d_1, d_2; 3\pi/4, 3\pi/4, \delta_1, \delta_2, \gamma). \quad (7)$$

In Eqs. 3, 5, and 6, we set $\gamma = 0$ because there are no guessing regions in these models.

We fit these models to the proxy decision space matrices K and N described in Proxy Decision Spaces as follows. We will use the double detection model D_{DD} for illustration. This model has three parameters: δ_1, δ_2, and γ. Observers' responses had only small biases, so we fixed the guessing parameter to $\gamma = 0.5$. As we show in SI Text, Properties of the Proxy Decision Space, the proxy decision space is the true decision space convolved with a Gaussian kernel, so we added another parameter σ to account for this blurring effect. Thus, we optimized parameters $\Theta = (\delta_1, \delta_2, \sigma)$. For a given choice of parameters, we first calculated a 20 × 20 matrix $M(\Theta) = (m_{ij})$ representing the predicted decision space, setting $m_{ij} = D_{DD}(x_i, x_j; \delta_1, \delta_2, \gamma)$ where x_i and x_j are the values used to construct the proxy decision space matrices as described in Proxy Decision Spaces. Next, we blurred $M(\Theta)$ by a Gaussian kernel $G(\sigma)$ with scale constant σ to obtain a 20 × 20 matrix representing the predicted proxy decision space, $Q(\Theta) = (q_{ij}) = M(\Theta) * G(\sigma)$, where * is 2D convolution. To avoid edge effects, we extended $M(\Theta)$ by 3σ on each side. To make the model fitting more robust against occasional keypress errors or lapses in attention, we made the probabilities in the proxy decision space saturate at 0.01 and 0.99, by defining $\tilde{Q}(\Theta) = (\tilde{q}_{ij}) = (\max(\min(q_{ij}, 0.99), 0.01))$. We used MATLAB's fminsearch function to find the parameter values that minimized the negative log likelihood of the observed proxy decision space matrices N and K,

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} - \sum_{ij} \log b\left(k_{ij}, n_{ij}, \tilde{q}_{ij}(\Theta)\right). \quad [8]$$

Here, $b(k, n, p)$ is the binomial probability mass function. To make the fitting routines more reliable, we took the best fit of 20 fits with randomly chosen starting points. In Fig. S3, we show the results of 20 independent fits of the GDD function to each observer's proxy decision space, to show that the fitting algorithm reliably converged to the global minimum.

The data from both experiments and MATLAB code that implements all our analyses are available online at purl.org/NET/rfm/pnas2015.

Cross-Validation. For cross-validation, we randomly divided each observer's trials into 10 equally sized blocks. On each cross-validation run, we used nine blocks for training and one block for validation. We measured the observer's classification image, decision variables, and proxy decision space on the training blocks, using the methods described in Proxy Decision Spaces. To find the cross-validation error for a given model (e.g., the difference model), we fitted the model to the proxy decision space from the training trials, using the methods described in Modeling. We then used the fitted model to find the negative log likelihood of the observer's responses in the validation block. We did this by taking the dot product of the classification image (measured from the training blocks) with the two stimulus intervals of each validation trial, producing two proxy decision variables (p_1, p_2). We then calculated the probability p of the observer making response 2 given the proxy decision variables (p_1, p_2), according to the fitted model being tested. If we let r = 1 on trials where the observer gave response 2 and r = 0 where the observer gave response 1, then the negative log likelihood of the observer's response is $-\log(rp + (1-r)(1-p))$. The negative log likelihood of all responses in the validation block is the sum of this negative log likelihood over all validation trials.

ACKNOWLEDGMENTS. We thank Minjung Kim, Yaniv Morgenstern, the editor, and two anonymous reviewers for comments on the manuscript, and James Elder and Laurie Wilcox for helpful discussions. This work was funded by grants to R.F.M. from the Natural Sciences and Engineering Research Council and the Canada Foundation for Innovation.

1. Green DM, Swets JA (1966/1974) Signal Detection Theory and Psychophysics (Kreiger, Malabar, FL).
2. Murray RF, Bennett PJ, Sekuler AB (2005) Classification images predict absolute efficiency. J Vis 5(2):139–149.
3. Murray RF (2011) Classification images: A review. J Vis 11(5):2.
4. Murray RF, Bennett PJ, Sekuler AB (2002) Optimal methods for calculating classification images: Weighted sums. J Vis 2(1):79–104.
5. Ahumada AJ, Jr (2002) Classification image weights and internal noise level estimation. J Vis 2(1):121–131.
6. Abbey CK, Eckstein MP (2002) Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. J Vis 2(1):66–78.
7. Tanner WP, Jr, Swets JA (1954) A decision-making theory of visual detection. Psychol Rev 61(6):401–409.
8. Ashby FG, Gott RE (1988) Decision rules in the perception and categorization of multidimensional stimuli. J Exp Psychol Learn Mem Cogn 14(1):33–53.
9. Treisman M, Leshowitz B (1969) The effects of duration, area, and background intensity on the visual intensity difference threshold given by the forced-choice procedure: Derivations from a statistical decision model for sensory discrimination. Percept Psychophys 6(5):281–296.
10. Yeshurun Y, Carrasco M, Maloney LT (2008) Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. Vision Res 48(17):1837–1851.
11. García-Pérez MA, Alcalá-Quintana R (2010) The difference model with guessing explains interval bias in two-alternative forced-choice detection procedures. J Sens Stud 25(6):876–898.
12. Wickelgren WA (1968) Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. J Math Psychol 5(1):102–122.
13. Akaike H (1974) A new look at statistical model identification. IEEE Trans Automat Contr 19(6):716–723.
14. Pelli DG (1985) Uncertainty explains many aspects of visual contrast detection and discrimination. J Opt Soc Am A 2(9):1508–1532.
15. Pelli DG, Farell B, Moore DC (2003) The remarkable inefficiency of word recognition. Nature 423(6941):752–756.
16. Tjan BS, Nandy AS (2006) Classification images with uncertainty. J Vis 6(4):387–413.
17. Gold JM, Murray RF, Bennett PJ, Sekuler AB (2000) Deriving behavioural receptive fields for visually completed contours. Curr Biol 10(11):663–666.
18. Ratcliff R, Smith PL (2004) A comparison of sequential sampling models for two-choice reaction time. Psychol Rev 111(2):333–367.
19. Wilder JD, Aitkin C (2014) Saccadic timing is determined by both accumulated evidence and the passage of timeJ Vis 14(10), 753 (abstr).
20. Palmer J, Verghese P, Pavel M (2000) The psychophysics of visual search. Vision Res 40(10-12):1227–1268.
21. Wetherill GB, Levitt H (1965) Sequential estimation of points on a psychometric function. Br J Math Stat Psychol 18(1):1–10.
22. Neri P (2010) How inherently noisy is human sensory processing? Psychon Bull Rev 17(6):802–808.

Movie S2. A typical stimulus sequence in experiment 2.

Movie S2