NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2016, 1–17

doi: 10.1093/nc/niw002
Research article

The signal processing architecture underlying subjective reports of sensory awareness

Brian Maniscalco[1,2,*] and Hakwan Lau[3]

[1]Department of Psychology, Columbia University, 1190 Amsterdam Ave., MC 5501, New York, NY 10027, USA; [2]National Institute of Neurological Disorders and Stroke, National Institutes of Health; [3]UCLA Psychology Department 1285 Franz Hall, Box 951563 Los Angeles, CA 90095-1563

*Correspondence address. National Institute of Neurological Disorders and Stroke, National Institutes of Health, 10 Center Drive, Building 10, Room 1D51A, MSC 1065, Bethesda, MD 20892-1065, USA, Tel: 301 451 7744; Fax: 301 480 2558, E-mail: bmaniscalco@gmail.com

Abstract

What is the relationship between perceptual information processing and subjective perceptual experience? Empirical dissociations between stimulus identification performance and subjective reports of stimulus visibility are crucial for shedding light on this question. We replicated a finding that metacontrast masking can produce such a dissociation (Lau and Passingham, 2006), and report a novel finding that this paradigm can also dissociate stimulus identification performance from the efficacy with which visibility ratings predict task performance. We explored various hypotheses about the relationship between perceptual task performance and visibility rating by implementing them in computational models and using formal model comparison techniques to assess which ones best captured the unusual patterns in the data. The models fell into three broad categories: Single Channel models, which hold that task performance and visibility ratings are based on the same underlying source of information; Dual Channel models, which hold that there are two independent processing streams that differentially contribute to task performance and visibility rating; and Hierarchical models, which hold that a late processing stage generates visibility ratings by evaluating the quality of early perceptual processing. Taking into account the quality of data fitting and model complexity, we found that Hierarchical models perform best at capturing the observed behavioral dissociations. Because current theories of visual awareness map well onto these different model structures, a formal comparison between them is a powerful approach for arbitrating between the different theories.

Key words: awareness; consciousness; contents of consciousness; theories and models; perception; psychophysics

Introduction

Humans and some nonhuman animals are able to assess the dependability of evidence associated with perceptual decisions by giving subjective ratings of confidence or visibility (Metcalfe and Shimamura, 1996; Kepecs, 2008; Smith, 2009; Fleming *et al.*, 2010). Conceptually, such subjective ratings are distinct from the associated perceptual decision; perceptual decisions are about states of the world, whereas subjective ratings are about the quality, quantity, or overall dependability of internal evidence associated with perceptual decisions. We can call perceptual decisions about the stimulus 'objective' judgments, and confidence and visibility ratings about one's own perceptual proccessing 'subjective' judgments.

Subjective and objective judgments are empirically dissociable. For instance, blindsight patients can objectively discriminate visual stimuli in their "blind" fields at above chance levels, and yet they deny having subjective perceptual experience (Weiskrantz, 1986; Azzopardi and Cowey, 1998; Davidson *et al.*, 2010). Under specific experimental manipulations, healthy human observers (Lau and Passingham, 2006; Wilimzig *et al.*, 2008; Rounis *et al.*, 2010; Rahnev *et al.*, 2011; Rahnev *et al.*, 2012;

Received: 15 July 2015; **Revised:** 24 November 2015. **Accepted:** 20 January 2016

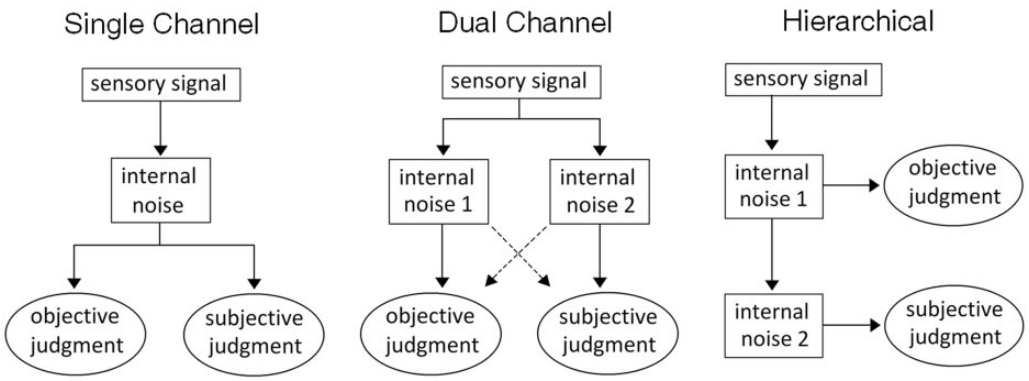

Figure 1. Schematic diagram for the three categories of models. (Left) According to a Single Channel model, the same process gives rise to both objective judgments (e.g. perceptual decisions about the identity of a stimulus) and subjective judgments (e.g. confidence ratings or visibility ratings). The model can still support some independence between task performance and subjective reports by supposing that the sensory evidence is a continuous variable that can be evaluated by setting various decision criteria (Fig. 2; Macmillan and Creelman, 2005). (Middle) An alternative model is that objective and subjective judgments are driven by two parallel processes, each influenced by independent sources of noise. Differential contribution of the two channels to objective and subjective judgments can lead to dissociations between the two kinds of responses. Note that the model can allow that each channel can contribute both kinds of judgments to some extent. In particular, one would expect that the channel which primarily influences one's subjective ratings would also heavily influence one's objective task response. For instance, when an observer subjectively reports clearly and vividly seeing squares, this should strongly correlate with objective judgments that the stimuli on the current trial are squares. (Right) Another alternative is that objective and subjective judgments are driven by different processes that are organized in a serial hierarchy, such that an early stage of processing generates the objective judgment and a later stage of processing generates the subjective judgment, as if the latter evaluates the quality of the former. Note that on this model, the second stage inherits the noise of the first stage, and thus the two are not entirely independent. However, the influence is one sided; the "subjective" stage does not influence the "objective" stage of processing

Vlassova *et al.*, 2014) and animals (Komura *et al.*, 2013; Fetsch *et al.*, 2014; Lak *et al.*, 2014) also exhibit some dissociations between subjective and objective judgments.

What are the mechanisms that drive subjective and objective judgments, and how are they related? The most parsimonious account would hold that subjective and objective judgments, though distinct, are generated from the same underlying process (Single Channel models, Fig. 1, left panel). For instance, on a common signal detection theory (SDT) account, perceptual decisions result from a binary comparison between an internal signal and a criterion, whereas subjective judgments of the quality of evidence are made by evaluating some transformation of the signal, such as its distance from the criterion (Clarke *et al.*, 1959; Galvin *et al.*, 2003). According to this kind of model, subjective and objective judgments are just different ways of evaluating the same underlying evidence (Fig. 2).

Alternatively, even if subjective and objective judgments are based on the same evidence, the 'quality' of evidence available for each kind of judgment might differ. For instance, a Hierarchical model (Fig. 1, right panel) might suppose that evidence is first used to generate objective perceptual decisions, and subsequently undergoes further processing to make subjective judgments (Cleermans *et al.*, 2007; Fleming *et al.*, 2010; Lau and Rosenthal, 2011). On such an account, the evidence might become degraded by the time it is processed by subjective judgment mechanisms, due to a decaying signal and/or the accrual of noise (Pleskac and Busemeyer, 2010).

A third possibility is a Dual Channel model (Fig. 1, middle panel) in which subjective and objective judgments are based on separate cognitive or neurophysiological processes (Jacoby, 1991; Jolij and Lamme, 2005; Del Cul *et al.*, 2009; Morewedge and Kahneman, 2010). For instance, perhaps there are two independent visual processing routes, one of which supports conscious vision and another whose visual processing is entirely unconscious. On such an account, subjective and objective judgments access different sources of information (and noise).

In the current work, we capitalize on a psychophysical paradigm that dissociates changes in objective perceptual decision performance from changes in subjective visibility ratings (Lau and Passingham, 2006) to evaluate SDT implementations of the model categories described above.

Materials and Methods

In the metacontrast masking procedure, stimulus identification performance varies with stimulus-mask onset asynchrony (SOA) in a U-shaped fashion (Fig. 3). Visibility judgments follow a similar U-shape that is asymmetrical with respect to the objective performance curve, thus yielding similar levels of performance associated with different levels of subjective stimulus visibility. We compared the ability of various implementations of the Single Channel, Dual Channel, and Hierarchical models to capture the relative dissociation between subjective and objective judgments found in this data set.

Participants

A total of 59 students from the Columbia University undergraduate population participated in the experiment and were paid $10 for approximately 1 h of participation. All subjects were naive regarding the purpose of the experiment, had normal or corrected-to-normal vision, and signed an informed consent statement. The research was approved by the Columbia University's Committee for the Protection of Human Subjects.

Experimental procedure

Subjects were seated in a dim room, 60 cm away from the computer monitor. Stimuli were generated using Psychophysics

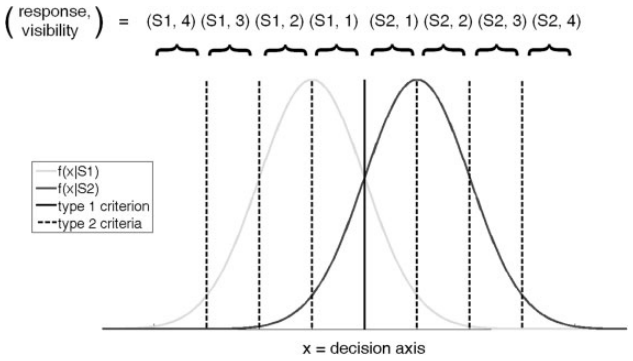

Figure 2. The standard SDT model. All models under consideration are built upon the foundation of the standard SDT model. This model assumes that stimulus categories S1 and S2 each generate normal distributions of perceptual evidence along an internal decision axis. The observer segments the decision axis into discrete regions using a type 1 criterion (for making a stimulus classification response) and a set of type 2 criteria (for rating subjective levels of decision confidence or percept visibility). The stimulus classification and subjective rating reported by the observer on any given trial are determined by which region of the decision axis contains the perceptual evidence observed on that trial, as illustrated in the figure. The probability with which the observer produces a given (response, visibility) pair upon being shown stimulus SN is equal to the area under the curve f(x|SN) in the region of the decision axis corresponding to that response pair. Note that it need not be the case that the type 1 criterion is located at the intersection of the distributions, or that the type 2 criteria are symmetrically distributed around the type 1 criterion

Toolbox (45) in MATLAB® (MathWorks, Natick, MA) and were shown on an iMac monitor (19 inch monitor size, 1680 × 1050 pixel resolution, 60 Hz refresh rate).

On each trial, a ring of eight shapes was presented around a central fixation point (4° radius). (A ring of stimuli was used with potential extension to fMRI in mind; to facilitate efficient retinotopic delineation of visual areas it is useful to present stimuli outside of the fovea. However, behavioral results similar to those reported here were also found with foveal presentation of single stimuli in Lau and Passingham, 2006.) Within each trial, each of the eight shapes was identical. The shapes could be either squares or diamonds with sides measuring 1.5° of visual angle. The shapes were presented for 33 ms on a gray background. Shapes were darker than the background, with the precise darkness determined separately for each subject by a thresholding procedure described below. A set of metacontrast masks designed to trace the outline of the square and diamond stimuli without physically overlapping with them (line width 0.025°) was subsequently displayed for 50 ms. Stimulus onset asynchrony (SOA) between stimulus and mask was determined randomly on each trial and counterbalanced among eight possible durations, ranging from 0 ms to 116.7 ms in increments of 16.7 ms.

Following each stimulus presentation, subjects provided two responses. First, they made a forced choice objective judgment about the shapes of the stimuli (squares or diamonds). Next, they rated how subjectively visible the shape of the stimulus appeared using a four-point scale. Specifically, subjects were asked to rate how clearly they had perceived the stimuli. Subjects were encouraged to use the entire rating scale while still accurately characterizing what they had visually experienced. Stimulus presentation for the next trial commenced 1050 ms after subjects entered the visibility rating. However, if subjects failed to enter both the stimulus identity judgment and the visibility rating within 5 s of stimulus offset, the current trial was aborted and the next trial commenced automatically.

After receiving task instructions, subjects completed two blocks of 28 practice trials. Following practice, subjects completed a block of 120 trials to determine the Weber contrast of the stimuli at which threshold performance across all SOAs could be obtained. Because performance in this task is close to maximal with an SOA of 0 ms (Lau and Passingham, 2006), all trials in the thresholding procedure had the minimum stimulus-mask SOA of 0 ms. We reasoned that if near maximal performance at 0 ms could be controlled to be at threshold levels, performance at other SOA values would also be near threshold. Stimuli were initially set to a Weber contrast of −0.15 and were subsequently adjusted online using a QUEST procedure (Watson and Pelli, 1983). Three separate QUEST tracks were recorded (40 trials each). Each QUEST track provided an independent estimate of the stimulus contrast needed to produce threshold performance (84% correct) at the minimum SOA. Trials for each track were interleaved randomly. Among the three resulting QUEST estimates, the median stimulus contrast was selected as the contrast to be used throughout the remainder of the experiment.

In the main experimental block, subjects completed 800 trials (100 trials for each of the 8 SOAs). SOAs were distributed across trials randomly. Every 100 trials, subjects received a self-terminated break lasting up to 60 s.

Subject selection

To maximize the suitability of the data for model fitting, we omitted from analysis all subjects who performed below chance levels at any of the SOAs ($n = 16$), any who performed perfectly at any of the SOAs ($n = 3$), and any whose mean visibility rating was lower than 5% of the maximum possible value at any SOA ($n = 1$). Most subjects were excluded due to having at least one SOA with below chance levels of performance, which is perhaps not surprising given that we performed the thresholding procedure on only the 0 ms SOA and subjects had many chances at each of the other SOAs to perform considerably worse, potentially recording average performance below chance. Nonetheless, we kept strict inclusion criteria to optimize model fitting.

For the remaining 39 subjects, we quantified the extent to which each subject exhibited a dissociation between objective task performance and subjective visibility ratings across SOA as follows. We made the qualitative observation that, when mean visibility is plotted as a function of mean task performance, the function is roughly linear, with a single outlying point corresponding to SOA = 16.7 ms for which visibility is lower than other SOAs with similar task performance (Fig. 3). Thus, for each subject, we ran a least squares regression between task performance (as assessed by the SDT measure d'; Macmilla and Creelman, 2005) and mean visibility rating at all but one SOA. The measured value of mean visibility at the left-out SOA was then subtracted from the "expected" visibility predicted by the regression on the other SOAs. We defined the absolute value of this difference between observed and expected visibility for the left-out SOA as the "dissociation score" for that SOA. We calculated the dissociation score for each SOA and defined each subject's "dissociation index" as the maximum dissociation score across all SOA from that subject's data. Each subject's

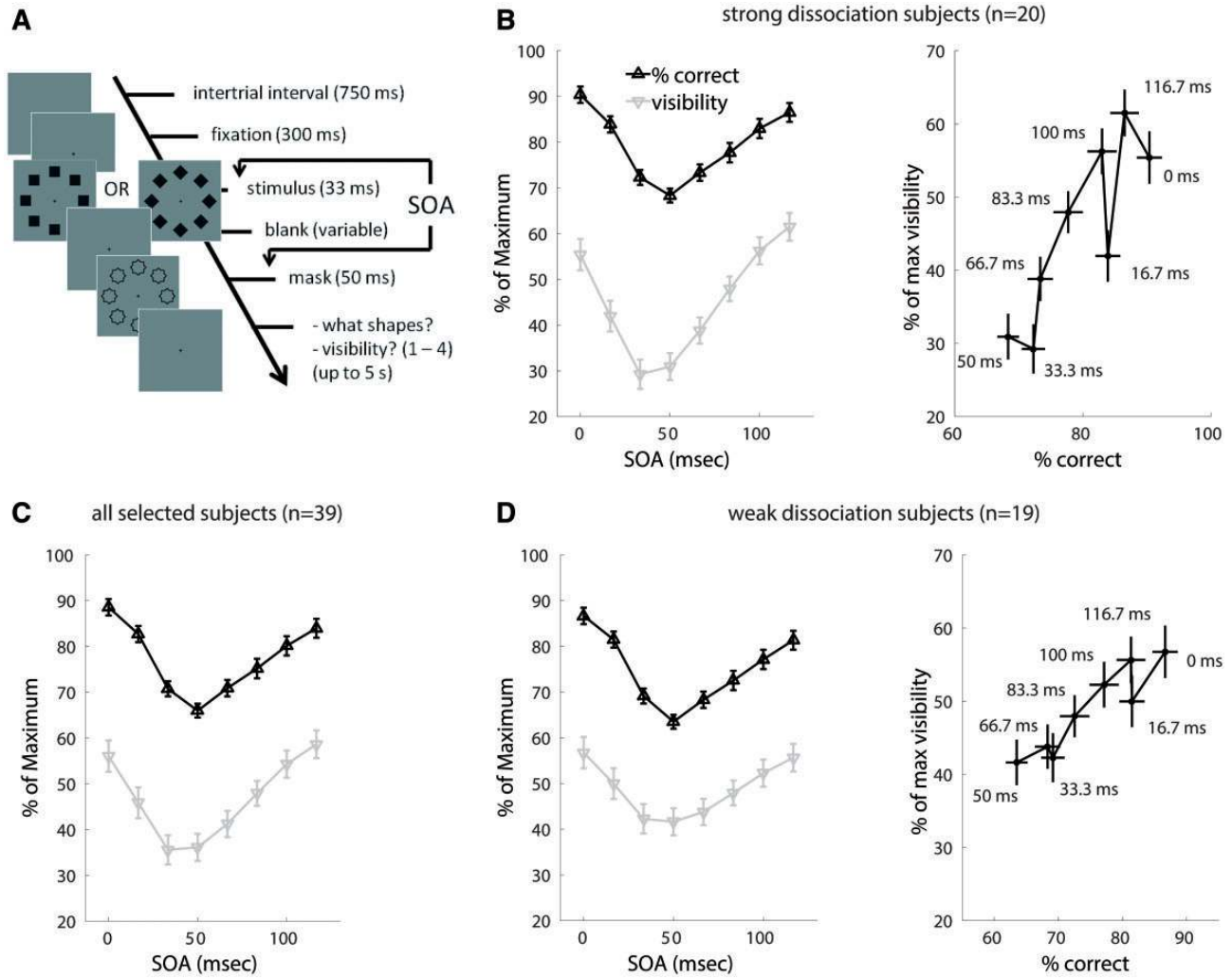

Figure 3. Experimental design and behavioral results. (**A**) We used a paradigm based on metacontrast masking, similar to the one used in a previous study (Lau and Passingham, 2006). In every trial, the subject was presented with a set of squares or diamonds (i.e. tilted squares). After a varying SOA (the temporal gap between the two sets of stimuli), a metacontrast mask was presented at the location of each shape. The masks were not drawn "on top of" the targets, but rather only traced the spatial contours where the targets had been presented. Nevertheless, the masks were successful in impairing target visibility. In each trial, subjects first decided whether the targets were squares or diamonds, and then gave subjective visibility ratings (four levels) to indicate how clearly they saw the identity of the targets. (**B**) Replicating previous findings (Lau and Passingham, 2006), this masking procedure gives rise to a U-shaped masking function when stimulus identification performance is plotted against SOA. The average level of subjective visibility ratings across SOAs, however, did not take the same shape, and reflected a bias toward giving lower ratings at lower SOAs. Shown here are data from a group of subjects ($n = 20$) who exhibited this dissociation particularly strongly. Right panel plots the same data as left panel, now showing visibility as a function of percent correct, to illustrate the dissociation more clearly. Data points correspond to SOA from the left panel. Visibility is an approximately linear function of percent correct, with the notable exception of SOA = 16 ms. This SOA had similar percent correct as SOA = 100 ms and 116.7 ms, but much lower visibility. (**C**) Data from all 39 subjects selected for model comparison analysis. (**D**) Data from the 19 subjects exhibiting a relatively weaker dissociation

dissociation index provides a measure of the extent to which that subject exhibited a dissociation between task performance and visibility ratings.

The logic of computing the dissociation index in this way presumes that visibility is linearly related to d' at all SOA other than the "left-out" one. This assumption is qualitatively born out by visual inspection of the data (Fig. 3). We tested this assumption more rigorously as follows. For each subject, after excluding the data from the SOA used to compute that subject's dissociation index, we performed separate linear and quadratic regressions of visibility onto d'. We then used a similar model comparison methodology as is used in the main data analyses to investigate whether the linear or quadratic regressions provided a better characterization of the data. We computed

$$\mathrm{AIC_c} = n \log \mathrm{MSE} + 2\mathrm{K}\left(\frac{n}{n - \mathrm{K} - 1}\right),$$

where MSE is the mean squared error of the regression, n is the number of data points, and K is the number of parameters in the regression model (Burnham and Anderson, 2002).

We then computed Akaike weights for the regression models using

$$w_i = \frac{e^{-\frac{1}{2}\left(\mathrm{AIC}_{c_i} - \mathrm{AIC}_{c_{\min}}\right)}}{\sum_{m=1}^{M} e^{-\frac{1}{2}\left(\mathrm{AIC}_{c_m} - \mathrm{AIC}_{c_{\min}}\right)}}.$$

The Akaike weights quantify the evidence in favor of each model by rewarding closer fits to the data and punishing greater model complexity. The mean Akaike weight for the linear regression model across all 39 subjects was 0.98 (out of a maximum possible value of 1), providing strong quantitative support for the qualitative observation that the relationship between visibility and task performance is roughly linear, and thus supporting our method for quantifying the dissociation index.

We performed a median split on the dissociation index, selecting the 20 subjects who exhibited the highest such value for further analysis and model fitting (Fig. 3B). All main analyses reported in the manuscript focus on this subset of 20 subjects. For these 20 subjects, the mean dissociation index was 0.57 and was greater than zero, $P < 0.001$. For all 39 subjects, the mean dissociation index was slightly weaker but still evident at 0.39 ($P < 0.001$). Without excluding any subjects at all ($n = 59$), a similar mean value of 0.41 obtains ($P < 0.001$).

Note that these procedures were performed to improve the quality and informativeness of the model comparison analysis. Omitting subjects with noisy data reduces the noisiness of model fits. Selecting the subjects who show the strongest dissociations between task performance and stimulus visibility provides a more stringent test for the models and thus provides a sharper way to compare their efficacy in characterizing the dissociation. Importantly, data that do not exhibit a dissociation between task performance and visibility can be straightforwardly captured by all the SDT models we consider here, and thus is not informative with respect to model comparison. It is only by examining the more interesting cases where task performance and visibility dissociate that the models considered here can be effectively differentiated in their ability to capture the data parsimoniously. Crucially, all subject selection procedures were performed *a priori*, prior to any model fitting analysis.

Although we focus on a selected subset of subjects for our main analysis, in the Supplementary Materials section titled "Expanded model comparison results" we report model selection results for different ways of selecting subjects. These supplementary analyses demonstrate that (i) as expected, model selection results are more equivocal for weak dissociation subjects, supporting our selection of strong dissociation subjects to more sharply discriminate the models; (ii) the main conclusions of our analyses remain the same when including all subjects rather than only strong dissociation subjects, as well as when analyzing all 59 subjects rather than the subset of 39 selected for having cleaner data. Thus our main conclusions are robust against the details of subject selection.

Model assumptions

In each model, we made standard SDT assumptions, as summarized in Fig. 2: (i) the two stimuli used in the experiment gave rise to internal signals normally distributed along some decision axis; (ii) perceptual decisions were made by comparing the signal on some decision axis to a criterion; and (iii) visibility judgments were made by comparing the signal on some decision axis to multiple criteria, corresponding to the multiple visibility ratings available to subjects in this experiment.

To further constrain model fitting, we made one further assumption: (iv) criteria for perceptual decisions and visibility ratings were set in the same way for each stimulus-mask SOA. That is, we assumed that subjects did not use different standards for deciding a stimulus's identity or visibility across the different SOAs. This assumption is justified by previous psychophysical findings. Gorea and Sagi (2000) found that when stimuli that are easy and difficult to perceive are interleaved randomly, subjects do not judge stimulus classes with separate criteria, but rather use a single, nonoptimal criterion for both. In our experiment, task difficulty varied across SOA, but SOAs were presented randomly, and thus task difficulty changed randomly across trials as it did in Gorea and Sagi (2000). If subjects cannot maintain separate sets of criteria for only two classes of randomly interleaved stimuli, it is highly unlikely that they could maintain seven distinct sets of criteria corresponding to the seven SOAs used in the current experiment.

Furthermore, in a study on the dynamics of criterion shifting, Brown and Steyvers (2005) found that criterion shifting is a slow process. In their experiment, task difficulty changed every 40 trials, requiring subjects to shift their decision criteria to maintain optimal task performance. However, even with this predictable block design, and even when subjects were forewarned that task difficulty would change during the experiment, subjects required about 8–22 trials (each trial lasting about 3.2 s) to change their decision criteria. In the current experiment, task difficultly changed randomly and rapidly from trial to trial. The results of Brown and Steyvers suggest that this rapid and random shift in stimulus difficulty would far outstrip subjects' ability to slowly adjust their decision criteria. Taken together, these experimental results suggest that it is highly implausible that subjects could have used different sets of decision criteria for each SOA, thus justifying our fourth modeling assumption.

An alternative way to implement constant criterion setting across different task conditions would be to define criteria based on the likelihood ratio of the evidence distributions rather than based on values of the decision axis (Rouder *et al.*, 2008). However, this approach requires assuming a much more cognitively taxing and difficult decision making strategy, as it would require that subjects (i) accurately estimate the evidence distributions and compute their ratio, and (ii) are able to do so separately for each SOA. Thus, here we opted to implement a simpler form of constant criterion setting, one that is more in the spirit of Gorea and Sagi's findings and has been successfully implemented in the modeling of other visual psychophysics tasks (Rahnev *et al.*, 2011, 2012). In future work, it may be fruitful to explicitly compare these alternative accounts of criterion setting to see which best accounts for a given data set.

Model descriptions

All models conformed to the broad specifications listed above, but differed from each other in the overall model structure (Single Channel, Dual Channel, or Hierarchical). Because there are many ways each model structure can be implemented, we compared multiple kinds of implementations for each model type. In total we fit 4 Single Channel models, 10 Dual Channel models, and 12 Hierarchical models. In the following, we give brief descriptions of each model tested. The names of the models in this section correspond to the model names used in Table 1.

Table 1. Complete model comparison results

Class	Model name	log L	# param	Average Akaike weight
Single channel	Single Channel	−1243.4519	15	0.0011
Single channel	**Single Channel CV**	**−1212.9612**	**23**	**0.1251**
Single channel	Decision Noise	−1316.5711	23	0
Single channel	Decision Noise CV	−1322.9558	31	0
Dual channel	Independent Dual Channel	−1233.6541	23	0
Dual channel	Independent Dual Channel CV	−1205.8477	31	0
Dual channel	Modulated Dual Channel 1	−1242.3511	23	0.0031
Dual channel	Modulated Dual Channel 1 CV	−1213.0657	31	0
Dual channel	Modulated Dual Channel 2	−1271.9686	23	0
Dual channel	Modulated Dual Channel 2 CV	−1242.3228	31	0
Dual channel	Modulated Dual Channel 3	−1299.1715	23	0
Dual channel	Modulated Dual Channel 3 CV	−1267.1083	31	0
Dual channel	**Weighted Dual Channel**	**−1222.6293**	**23**	**0.1445**
Dual channel	Weighted Dual Channel CV	−1201.961	31	0.0364
Hierarchical	Decay Only	−1216.9266	23	0.0037
Hierarchical	Decay Only CV	−1209.4122	31	0.0006
Hierarchical	Noise Only	−1222.1802	23	0.0002
Hierarchical	Noise Only CV	−1202.6069	31	0.0156
Hierarchical	Noise + Decay	−1242.2019	31	0
Hierarchical	Noise + Decay CV	−1197.3804	39	0
Hierarchical	Noise + Constant Decay	−1222.4464	24	0.0004
Hierarchical	Noise + Constant Decay CV	−1197.189	32	0.2023
Hierarchical	**Constant Noise + Decay**	**−1211.1514**	**24**	**0.2619**
Hierarchical	Constant Noise + Decay CV	−1201.5188	32	0.0199
Hierarchical	Constant Noise + Constant Decay	−1233.1541	17	0.0458
Hierarchical	Constant Noise + Constant Decay CV	−1204.0798	25	0.1391

"Class" denotes model category (see Fig. 1). Descriptions of each model listed under "Model name" are available in "Materials and Methods" section, Model descriptions. "log L" is the quantitative measure of goodness of fit for each model, the log of the likelihood of the observed empirical data given the model structure and optimal parameter values. Larger values indicate better fit. "# param" lists the number of parameters for each model, a measure of model complexity. "Akaike weight" is a measure of overall model quality, taking into account goodness of fit and model complexity. Larger values indicate better models, and the weights are scaled such that they sum to 1. For more details on these measures see "Materials and Methods" section, Formal model comparison. The best models in each model class are highlighted in boldface.

Single Channel models

Single Channel. parameters: μ_{diff} (8), c (7).

The simplest model we tested was this basic SDT model. We suppose that the distance between the evidence distributions, μ_{diff}, changes for each of the eight stimulus-mask SOAs. The observer must set seven decision criteria to partition the decision axis into eight regions, which correspond to the eight kinds of responses the observer can give on a given trial (2 stimulus classifications × 4 levels of subjective visibility). For all models, we suppose that the decision criteria are constant across SOA.

Single channel CV ("changing variance"). parameters: μ_{diff} (8), σ (8), c (7).

This is a modification of the Single Channel model which supposes that SOA affects not only the absolute distance between the stimulus distributions μ, but also their common standard deviation σ.

Note that the SDT measure of task performance, d', is simply the ratio of μ_{diff}/σ. Thus, one might worry that μ_{diff} and σ are redundant here and could instead be captured by a single parameter, d'. However, recall that our SDT model for the metacontrast masking task also supposes a single set of decision criteria which is held constant across SOA. μ_{diff} and σ pairings at different SOAs that have the same ratio (i.e. yield the same value of d') will nonetheless have different relationships to these constant criteria, and thus such pairings are not redundant in the behavioral data they generate. For instance, suppose that at SOA 1, $\mu_{diff} = 2$ and $\sigma = 1$, whereas at SOA 2, $\mu_{diff} = 4$ and $\sigma = 2$. In this scenario, d' will be the same for SOA 1 and SOA 2, but average visibility will not; SOA 2 will have higher visibility since the stimulus distributions are farther apart and are more variable, and thus more probability mass in the distributions will exceed the decision criteria, resulting in higher visibility ratings.

Other CV models. For every model described below, we analyzed versions which did and did not allow the standard deviation of the stimulus distributions σ to vary across SOA. Every model following the naming format "Model X CV" is identical to the simpler model "Model X" with the exception that it has eight added parameters to allow σ to vary with SOA.

Decision noise. parameters: μ_{diff} (8), σ_c (8), c (7).

This model supposes that the type 2 criteria (the six decision criteria used to evaluate subjective visibility) are not constant from trial to trial, but in fact are drawn from a normal distribution with some standard deviation σ_c, where σ_c can vary with SOA. This model is based on Mueller and Weidemann (2008).

Dual Channel models

Dual Channel models suppose that two separate information processing streams accruing noise from independent sources contribute to the perceptual decision-making process. In SDT terms, these models posit the existence of two decision axes, one of which corresponds to conscious processing and the other, unconscious processing. The versions of these models

considered here differ on how they suppose information from the conscious and unconscious processing channels are combined.

Independent Dual Channel. parameters: $\mu_{diff\ C}$ (8), $\mu_{diff\ U}$ (8), c_C (6), c_U (1).

The distance between stimulus distributions is modulated by SOA for both the conscious (μ_C) and unconscious (μ_U) decision axes. The conscious decision axis is only used to categorize stimuli that have a visibility of at least 2 or higher, i.e. it is not used to classify stimuli with visibility = 1. For this reason, only six decision criteria c_C are set on the conscious decision axis. For stimuli whose visibility is only rated as 1, the stimulus classification is made by doing signal detection on the unconscious decision axis using the criterion c_U. This model is based on Del Cul *et al*. (2009).

See Supplementary Material section titled "Comparison of Dual Channel SDT models in present paper and Del Cul et al (2009)" for an explicit comparison between our Independent Dual Channel model and the model used in Del Cul *et al*. See Supplementary Material section titled "Model comparison results using median split on visibility ratings" for a demonstration that model comparison results are not appreciably affected if we use a median split on each subject's visibility ratings to define conscious and unconscious trials, rather than defining unconscious trials as trials where visibility = 1.

Modulated Dual Channel N (N = 1, 2, 3). parameters: $\mu_{diff\ C}$ (8), $\mu_{diff\ U}$ (8), c_C (6), c_U (1).

These models are identical to the Independent Dual Channel model, with one exception. Modulated Dual Channel *N* has a provision for altering subjective reports of visibility made from the conscious decision axis when its stimulus classification conflicts with the stimulus classification provided by the unconscious channel. Specifically, if visibility > 1 and visibility ≤ *N* + 1, and if the stimulus classification of the conscious and unconscious channels disagree, then the classification from the conscious channel is used but the report of subjective visibility is reduced to 1.

Weighted Dual Channel. parameters: $\mu_{diff\ C}$ (8), $\mu_{diff\ U}$ (8), c_C (6), c_{TOT} (1).

Rather than treat information from the conscious and unconscious channels separately, the observer combines them into a new decision axis by computing a weighted average. The weight given to evidence arising from the conscious channel is $w_C = d'_C / (d'_C + d'_U)$, where $d' = \mu_{diff}/\sigma$ and $\sigma = 1$ for the non-CV models. This formula can give results outside of [0, 1] if negative d' values are entered. As a correction for this possibility, if the computation yields $w_C < 0$ then w_C is set to 0, and if it yields $w_C > 1$ then w_C is set to 1.

If visibility = 1, the stimulus is classified using the combined channel. If visibility > 1 and the conscious channel and combined channel agree on stimulus classification, then stimulus classification is given with the level of visibility dictated by the conscious channel. But if visibility > 1 and the conscious channel and combined channel disagree on stimulus classification, then the classification from the conscious channel is used but the report of subjective visibility is reduced to 1.

(Although it would be optimal to always use the stimulus classification provided by the combined channel, implementing this in the model would allow the nonsensical result that reports of stimulus classification could conflict with reports of subjective visibility, e.g. "the stimuli were squares, and I very clearly saw that the stimuli were diamonds.")

Hierarchical models

Hierarchical models suppose that stimulus classification occurs according to Single Channel SDT principles, but that the perceptual evidence used to do stimulus classification changes before it is used to report subjective visibility, becoming weaker and/or noisier.

Decay only. parameters: μ_{diff} (8), k (8), c (7).

The perceptual evidence used for performing stimulus classification is multiplied by a factor of k before it is used for reporting subjective visibility, where $0 \leq k \leq 1$. k varies across SOA. We constrained k to be less than or equal to 1 to be in line with previous empirical and theoretical SDT demonstrations that in visual psychophysics tasks like the one used here, the information content of subjective ratings is constrained by task performance (Galvin *et al*., 2003; Maniscalco and Lau, 2012; Maniscalco and Lau, 2015).

Noise only. parameters: μ_{diff} (8), σ_h (8), c (7).

The perceptual evidence used for performing stimulus classification accrues noise before it is used for reporting subjective visibility. The noise is sampled from a normal distribution with mean 0 and standard deviation σ_h. σ_h varies across SOA.

Noise + Decay. parameters: μ_{diff} (8), σ_h (8), k (8), c (7).

A combination of the Decay Only and Noise Only models.

Noise + Constant Decay. parameters: μ_{diff} (8), σ_h (8), k (1), c (7).

Same as Noise + Decay, but the signal decay parameter k is constrained to be constant across SOA.

Constant Noise + Decay. parameters: μ_{diff} (8), σ_h (1), k (8), c (7).

Same as Noise + Decay, but the hierarchical noise parameter σ_h is constrained to be constant across SOA.

Constant Noise + Constant Decay. parameters: μ_{diff} (8), σ_h (1), k (1), c (7).

Same as Noise + Decay, but the hierarchical noise parameter σ_h and signal decay parameter k are constrained to be constant across SOA.

Model fitting

Past efforts to fit SDT parameters to rating data have used the following approach (Dorfman and Alf, 1969). First, we make two simplifying assumptions: (i) responses on each trial are independent from one another; (ii) the probability of each response type associated with each stimulus class is constant across trials. When these assumptions are met, the likelihood of a set of signal detection model parameters given the observed data can be calculated using the multinomial model. Formally,

$$L(\theta|\text{data}) \propto \prod_{i,\ j} \text{Prob}_\theta(\text{Resp}_i|\text{Stim}_j)^{n_{\text{data}}(\text{Resp}_i|\text{Stim}_j)},$$

where each Resp_i is a behavioral response (stimulus classification and visibility rating) a subject may produce on a given trial, and each Stim_j is a type of stimulus that may be shown on a given trial. $\text{Prob}_\theta(\text{Resp}_i|\text{Stim}_j)$ denotes the probability with which the subject produces the response Resp_i after being presented with Stim_j, according to the signal detection model

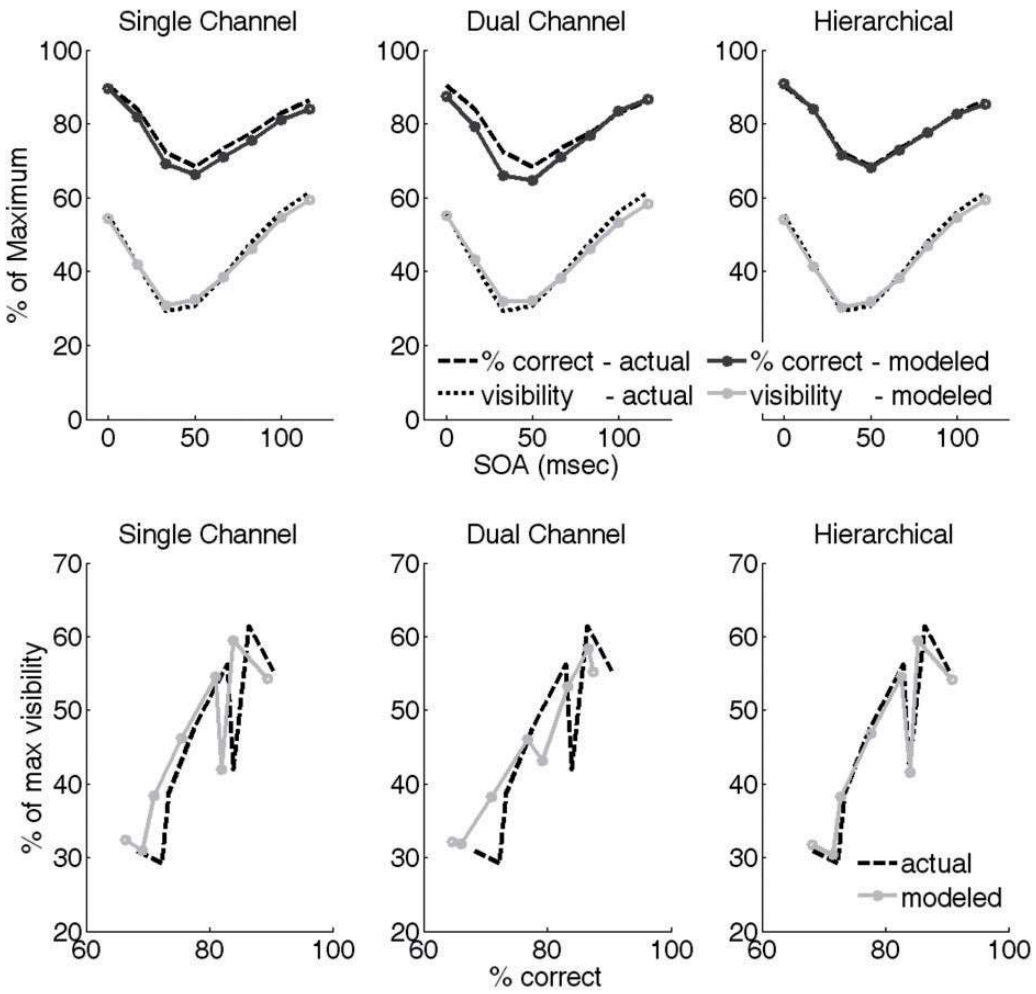

Figure 4. Model fits for task performance and reported visibility. Three categories of models (Single Channel, Dual Channel, and Hierarchical) were fitted to the behavioral data from the metacontrast masking paradigm. We tested multiple versions of each category of model (see "Materials and Methods" section for details). Shown here are the best-fitting models from each category, selected according to formal model comparison techniques (Fig. 6). The Hierarchical model performed best at capturing the dissociation between task performance and reported levels of stimulus visibility. This dissociation is made readily apparent by plotting visibility reports against task performance, as depicted in the bottom row of figures; the relationship is not monotonic, but exhibits a sharp spike at around 80–85% correct, reflecting that short SOAs had lower visibility than long SOAs in spite of having similar task performance

specified with parameters θ. $n_{data}(Resp_i|Stim_j)$ is a count of how many times a subject actually produced $Resp_i$ after being shown $Stim_j$.

In the current study, the subject has 8 possible responses from which to select (2 stimulus classification options $\times$ 4 levels of visibility rating), and there are 16 stimulus types (2 stimulus shapes $\times$ 8 levels of stimulus-mask SOA). The set of parameters θ for each model is listed above in the section titled "Model descriptions."

The set of parameters θ that is most likely given the observed data is the maximum likelihood parameter estimate. The signal detection models under consideration in this study differ in the distributions of $Prob_\theta(Resp_i|Stim_j)$ values they can produce, which in turn determines the extent to which they can fit the data well and achieve a high maximum likelihood in the multinomial model.

Note that models were not fit to summary statistics of performance such as percent correct or average visibility. Rather, models were fit to the full distribution of probabilities of each response type contingent on each stimulus type. From this full behavioral profile of stimulus-contingent response probabilities, we can derive various summary statistics such as percent correct and average visibility (Fig. 4), as well as type 2 performance (Fig. 5). Thus, the behavioral data shown in these figures are not the data upon which the models were explicitly fit, but rather different ways of highlighting aspects of the model fit to the full set of response counts for every stimulus type.

We fit all models under consideration to the observed data by finding the maximum likelihood parameter values θ. Maximum likelihood fits were found using a simulated annealing procedure (Kirkpatrick *et al.*, 1983). Model fitting was conducted separately for each subject's data.

Formal model comparison

The maximum likelihood associated with each model characterizes how well that model captures patterns in the empirical data. However, comparing models directly in terms of likelihood can be misleading; greater model complexity can allow for tighter fits to the data but can also lead to undesirable levels of

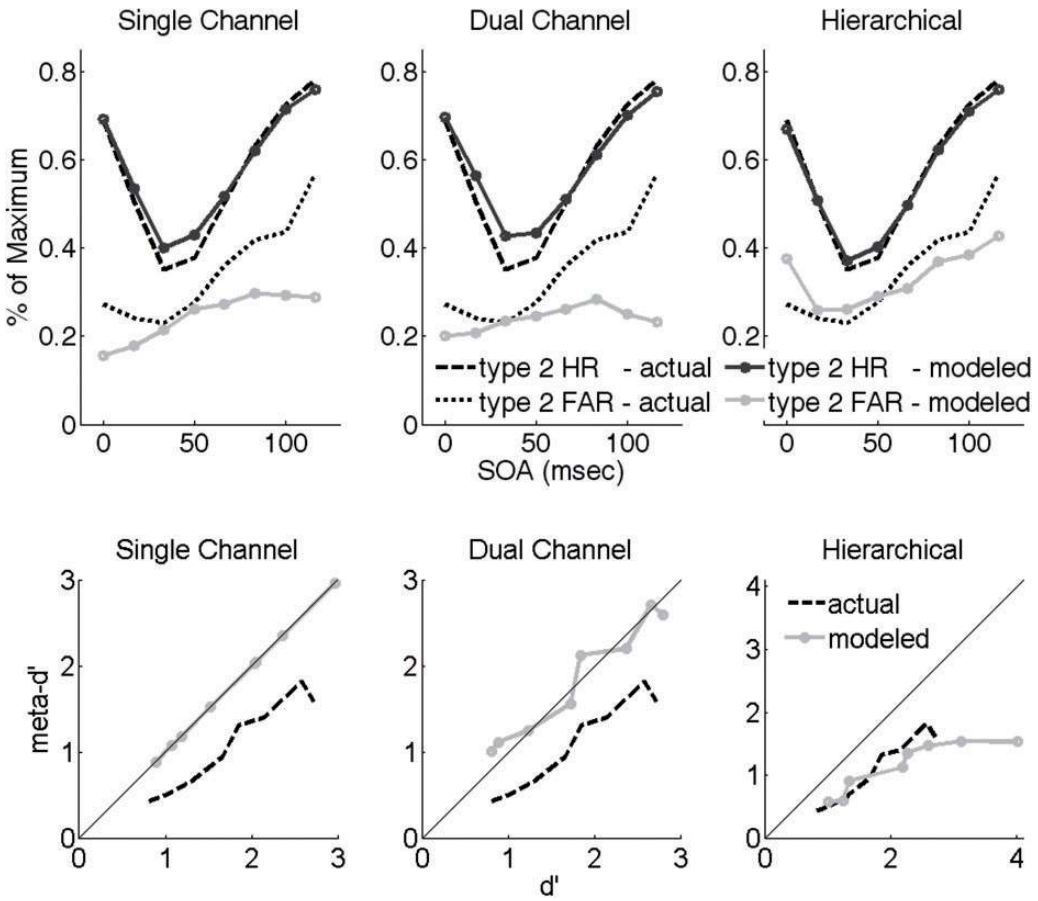

Figure 5. Model fits for type 2 data. In addition to the distinctive dissociation between task performance and visibility (Fig. 4), the behavioral data also included a set of type 2 data that provided a challenge for model fitting. By "type 2 data," we refer to the probability of giving different levels of visibility ratings conditional upon task performance. (Top panel) Type 2 hit rate (HR; probability of high visibility for correct responses) and type 2 false alarm rate (FAR; probability of high visibility for incorrect responses) as a function of SOA. Note that for ease of visualization, here we plot only a single (type 2 HR, type 2 FAR) pair based on a median split of the visibility rating scale, although in principle a four-point rating scale yields three such pairs, since there are three ways to collapse the four-point scale into a binary distinction between "high" and "low" visibility. (Bottom panel) meta-d' as a function of d', where each data point corresponds to one SOA. SDT models predict that meta-d' = d', i.e. that increases in task performance manifest as increases in the discriminability of correct and incorrect trials by subjective ratings of confidence or awareness. This improved discriminability of task performance manifests as a divergence of type 2 hit rate and type 2 false alarm rate. Although meta-d' was lower than d' at all SOA in this dataset, the Single Channel model could not reproduce this pattern as it is constrained such that it can only yield values of meta-d' equal to d' (Maniscalco and Lau, 2012; Maniscalco and Lau, 2014). The best performing Dual Channel model is not rigidly constrained in this way, but still produced meta-d' values close to d'. In contrast, the Hierarchical model produced a meta-d' vs d' curve that lies well below the line of unity, providing a closer fit to the data. Note that meta-d' was computed using all three available (type 2 HR, type 2 FAR) pairs for each SOA, i.e. it was not computed based on a median split of visibility

overfitting, i.e. the erroneous modeling of random variation in the data. The Akaike Information Criterion (AIC), motivated by considerations from information theory, provides a means for comparing models on the basis of their maximum likelihood fits to the data while correcting for model complexity (Burnham and Anderson, 2002). We used AIC_c, a variant of AIC which corrects for finite sample sizes:

$$AIC_c = -2\log L(\theta|data) + 2K\left(\frac{n}{(n-K-1)}\right),$$

where K is the number of parameters in the model and n is the number of observations being fit (i.e. $n = 800$ trials). Lower values of AIC_c are desirable because they indicate a higher model likelihood and/or a lower model complexity (lower number of parameters).

We use the likelihood of each model, given the data, as a basis for model comparison:

$$L(\text{model}_i|\text{data}) \propto e^{-\frac{1}{2}\left(AIC_{c_i} - AIC_{c_{min}}\right)}.$$

$AICc_i$ is the AIC_c for model i and $AICc_{min}$ is the lowest AIC_c exhibited by all models under consideration. These model likelihoods can be scaled to sum to 1, and the resulting "Akaike weights" reveal the relative weight of evidence for each model as evaluated by its fit to the data, correcting for model complexity.

$$w_i = \frac{e^{-\frac{1}{2}\left(AIC_{c_i} - AIC_{c_{min}}\right)}}{\sum_{m=1}^{M} e^{-\frac{1}{2}\left(AIC_{c_m} - AIC_{c_{min}}\right)}}.$$

The Bayesian Information Criterion (BIC) is one alternative measure to AIC that can be used for model selection (Burnham and Anderson, 2002). To assess the relative merits of AIC_c and BIC for the present data, we simulated data for the best-performing models in each model class (see "Results" section) and then fit these models to the simulated data to compare how well AIC_c and BIC could accurately detect the model that had generated the simulated data (Supplementary Materials, "Model comparison analysis recovery of model-generated data"). This analysis suggested that AIC_c is the more suitable measure of the two for the present data, as it exhibited less bias in its model recovery than BIC, and BIC exhibited a particular weakness for accurately classifying Hierarchical models (Supplementary Fig. S1). We also present full model comparison results using both AIC_c and BIC in the Supplementary Materials ("Expanded model comparison results").

Results

We replicated previous findings that objective task performance and subjective ratings of stimulus visibility can dissociate in the metacontrast masking paradigm (Lau & Passingham, 2006; Fig. 3). Both task performance and visibility follow U-shaped functions of SOA between stimulus and mask. The dissociation manifests as an asymmetry in the two curves, such that there exist SOAs where task performance is similar and yet stimulus visibility differs. The dissociation is made more plainly visible by plotting visibility ratings as a function of task performance, as in the right panel of Fig. 3B. Although visibility is a roughly linear function of task performance across most SOAs, visibility at the SOA of 16.7 ms is markedly lower than would be expected from the other SOAs exhibiting similar levels of task performance.

This dissociation is precisely the feature of the data from the metacontrast masking paradigm that we hoped to use to leverage a decisive model comparison analysis for the SDT models under consideration. Thus, we selected a subset of subjects who most markedly exhibited the dissociation ($n = 20$; Fig. 3B) and used this group of subjects for all model comparison analyses reported below. The remaining subjects did not exhibit a strong dissociation (Fig. 3D) and thus did not provide suitable informative data for model comparison. In the "Expanded model comparison results" section of the Supplementary Material, we show that our main conclusions are robust against particular choices for subject selection.

Complete model comparison results are listed in Table 1. To simplify analysis, we focus on comparing the best-performing models in each model class (i.e. the models with the highest average Akaike weight within each model class). These are the models titled "Single Channel CV," "Weighted Dual Channel," and "Constant Noise + Decay." Details of model specifications can be found in Materials and Methods section under the heading "Model Descriptions."

Figure 4 displays the fits of these models to stimulus classification accuracy and mean visibility ratings at each SOA. In the top panel, we plot average percent correct and visibility across subjects at each SOA, as well as the average model fit for these same data across subjects. The same data are replotted in the bottom panel to show mean visibility as a function of accuracy, so as to emphasize the strong dissociation between the two found in the behavioral data. Visual inspection suggests that the best Single Channel model qualitatively captures the performance/visibility dissociation, yet systematically underestimates task performance at all SOAs. The best Dual Channel model is not successful at capturing the dissociation. In contrast, the Hierarchical model provides a close fit to both the task performance and visibility curves.

Another way of probing the relationship between objective task performance and subjective visibility rating is to analyze the behavior of subjective ratings conditioned on stimulus classification accuracy, what has been called "type 2" analysis to distinguish it from the "type 1" analysis of basic stimulus identification performance (Clarke *et al.*, 1959; Galvin *et al.*, 2003). In the top panel of Fig. 5, we show model fits to type 2 hit rate [HR; p(high visibility | correct)] and type 2 false alarm rate [FAR; p(high visibility | incorrect)], where "high visibility" is defined for each subject as a visibility rating greater than that subject's median visibility rating across all trials. In the bottom panel we plot meta-d', a measure of how well subjective rating discriminate between correct and incorrect trials (Maniscalco and Lau, 2012; Maniscalco and Lau, 2014), as a function of d'. Each point in the curve corresponds to the meta-d' and d' values from one of the eight SOA conditions. Meta-d' is defined such that it equals d' for an observer whose behavior conforms perfectly to traditional SDT predictions. Thus, the line meta-d' = d' displayed in these plots depicts the SDT prediction, and indeed the Single Channel model produces a meta-d' curve lying exactly along this line, thus systematically overestimating subjects' actual meta-d'. The Dual Channel model allows for some deviation from the Single Channel prediction, yet not in such a way that captures the patterns in the data; the Dual Channel model also systematically overestimates meta-d'. The Hierarchical model is unique among the models considered here in its ability to capture the fact that meta-d' in this data set is systematically below traditional SDT prediction.

The results reported in Fig. 5 are easy to intuit. For the Single Channel model, the strong relationship between type 1 performance (d') and type 2 performance (meta-d') is due to the fact that they are based on the same underlying information; there is no additional process by means of which the quality of information available to type 1 and type 2 mechanisms could differ. Thus, this fundamental assumption of the Single Channel models makes them somewhat inflexible in capturing variation in the relationship between type 1 and type 2 performance. In principle, Single Channel models can reduce type 2 performance without affecting stimulus classification accuracy by supposing that type 2 criterion setting is a noisy process, such that the placement of the criteria varies randomly from trial to trial (Mueller and Weidemann, 2008), but this class of models gave poor overall fits to the current data set (Table 1).

One may expect the Dual Channel model to fare better because it postulates two different processes. However, this was not the case. The reason is that the "conscious" channel essentially acts like a Single Channel model, supposing a tight relationship between task performance and subjective visibility, and the "unconscious" channel is limited in the extent to which it can interfere with fully "conscious" processing. In the original implementation of the Dual Channel model (Del Cul *et al.*, 2009), the processing of the "unconscious" channel only manifests in behavior on trials where the subject reports not seeing the stimulus, whereas trials where the subject does report seeing the stimulus are only affected by the "conscious" channel. Thus, the unconscious channel can only alter the relationship between type 1 and type 2 performances by allowing for extra degrees of freedom in type 1 performance for "unconscious" trials. For "conscious" trials, task performance and visibility rating are still controlled by the same underlying source of information and thus are still constrained in how they may covary,

similar to Single Channel models. Thus, in many instances Dual Channel models may be limited in their ability to produce type 2 Receiver Operating Characteristic (ROC) curves that deviate strongly from Single Channel predictions. This constraint can be seen in Fig. 5, in which the Dual Channel model produces meta-d' values that are close to the Single Channel expectation of meta-d' = d'.

It is possible that Dual Channel models featuring more extensive and complicated interactions between the two channels could fare better, but such models would potentially constitute a departure from the fundamental dichotomy between "conscious" and "unconscious" processing streams that arguably is the main conceptual motivation for proposing the Dual Channel class of models. As it stands, the best Dual Channel model we tested already posits that in cases of high conflict, the "unconscious" channel can modulate visibility ratings made by the "conscious" channel; simpler Dual Channel models that better respected the distinction between "conscious" and "unconscious" processing performed worse (Table 1).

In contrast, the dissociation between type 1 and type 2 performance is more naturally captured by Hierarchical models, as they stipulate a less restrictive relationship between the quality of information available for type 1 and type 2 decision making. Changing the degree to which the evidence becomes degraded at the second stage of processing provides a means of changing the patterns of subjective rating without affecting basic task performance, which is determined by the first stage of processing. Degradation of signal quality at the second stage of processing also provides a natural mechanism for degradation of type 2 performance, as manifested in levels of meta-d' below the traditional SDT expectation (Fig. 5).

One feature of the average Hierarchical model fit to the meta-d' and d' data that may appear puzzling at first is that it seems to overestimate d' for certain SOA (Fig. 5) while not overestimating percent correct at any SOA (Fig. 4). This is likely due to the fact that for some subjects at certain SOAs, the Hierarchical model fit produced very high d' values ($\geq$6). At near-ceiling levels of task performance, large changes in d' correspond to small changes in percent correct. For instance, for a subject with unbiased responding, d' = 4 corresponds to 97.7% correct, whereas d' = 6 corresponds to 99.9% correct. Thus it is likely that for subjects at SOAs with near-ceiling task performance, the best overall fit to the data was one that slightly overestimated percent correct (and thus largely overestimated d') to capture other features of the overall data set, such as mean visibility rating.

As models become more complex, in general they become better able to capture real patterns in data, but also become more prone to erroneously capture noise in the data (overfitting). Thus, we conducted formal model analysis using the AIC, which rewards models for closeness of fit to observed data while punishing them for complexity (number of parameters). In Fig. 6, we present model comparison results based on a finite-sample correction of AIC, AIC_c (Burnham and Anderson, 2002). Overall, the hierarchical category of models collectively outperformed the Single Channel and Dual Channel models (top panel), and this pattern held up when comparing only the best models in each category (bottom panel). Thus, the superior goodness of fit for the Hierarchical model evident in Figs 4 and 5 cannot be attributed to overfitting. In fact, the three best models in each model category, though visibly differing in quality of data fitting, had essentially the same number of parameters (Single Channel and Dual Channel, 23; Hierarchical, 24).

One caveat for this model comparison analysis is that our three model classes had unequal numbers of models. We tested

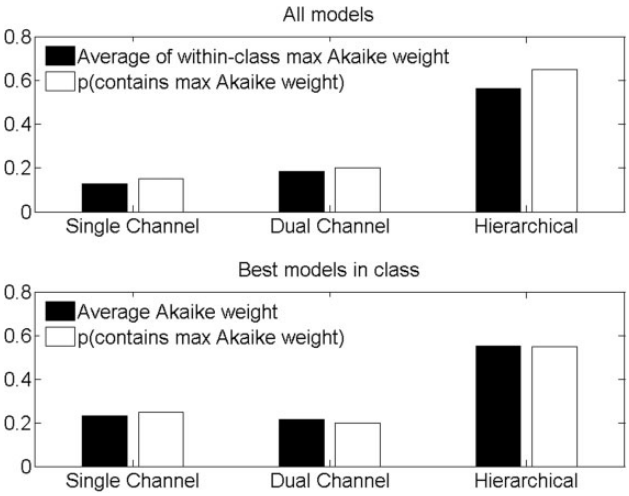

Figure 6. Model selection results. Formal model comparison was conducted using a finite sample size correction of the Akaike Information Criterion (AIC_c), which rewards models for closely fitting observed data while punishing models for the degree of complexity (i.e. number of free parameters; for list of free parameters for all models please see "Materials and Methods" section). For ease of interpretation, we display a transformation of AIC_c values into Akaike weights, which quantify the information theoretic evidence in favor of each model such that the weights sum to 1 (Burnham and Anderson, 2002). (Top panel) Model selection on all 26 models. Black bars plot the across-subject average of the Akaike weights that were maximal in each model class for each subject. White bars plot the probability that the maximum Akaike weight across all 26 models belonged to a given model class. For both measures, the value for the Hierarchical model class was roughly three times higher than that of the Dual Channel model class and roughly four times higher than that of the Single Channel model class. (Bottom panel) Similar results were found when restricting the analysis to the best models in each model class. Best models were defined by computing the average Akaike weight for all 26 models, and then selecting the models in each model class with the maximum average Akaike weight. Full model comparison results can be found in Table 1

4 Single Channel models, 10 Dual Channel models, and 12 Hierarchical models. All else being equal, the model class containing the greater number of models will tend to be favored. We attempted to mitigate the impact of this caveat in two ways. First, we found the maximum Akaike weight within each class for each subject and compared the model classes in their average maximum weight (Fig. 6, top panel). Second, we selected the one model from each class that had the largest average Akaike weight across subjects, and performed a new model comparison analysis on this restricted subset of models (Fig. 6, bottom panel). Future work may incorporate alternative strategies, such as weighting each model class by a prior probability based on the number of models being considered for each model class (Donkin *et al.*, 2015).

We can derive further insight into the way the best models in each category captured the data by investigating their parameter values (Fig. 7).

The fit for the best-performing Single Channel model indicates a U-shaped curve for σ, the standard deviation of the perceptual evidence distributions, such that σ takes on higher values at longer SOAs. When criteria are held constant across SOA (a stipulation for all models, see "Materials and Methods" section), larger values of σ entail higher levels of mean visibility

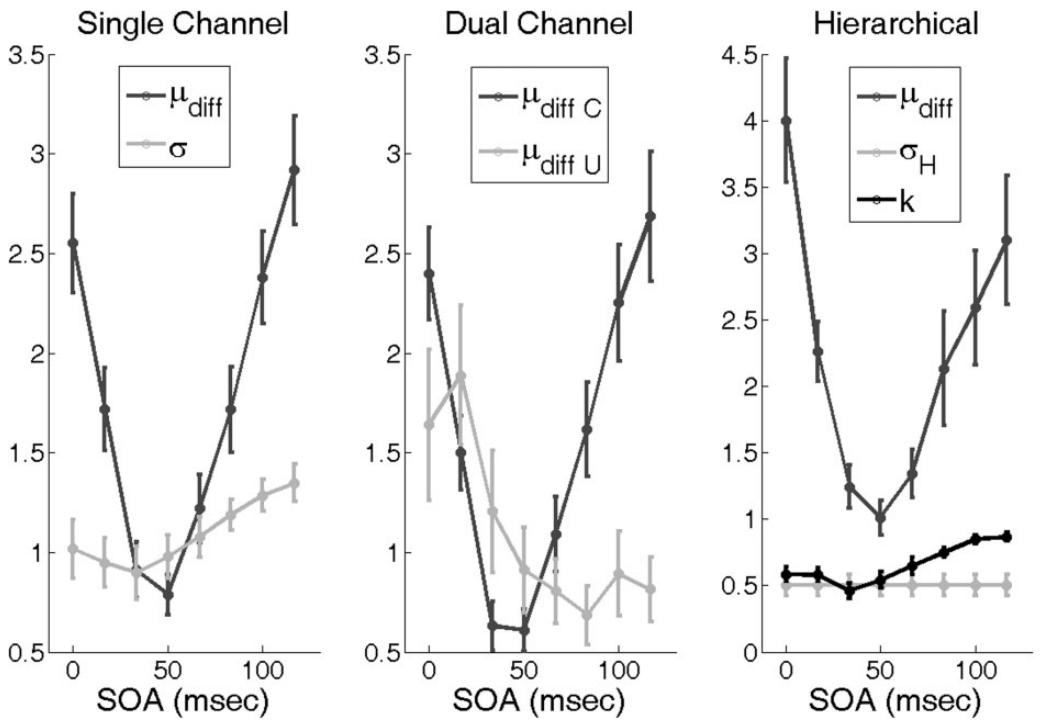

Figure 7. Parameter values from model fits. Parameters for the best fitting models from each model class—"Single Channel CV," "Weighted Dual Channel," and "Constant Noise + Decay"—are plotted. In addition to the depicted parameters, each model fit also specified values for seven decision criteria (data not shown). For descriptions of model structure and parameters, see "Materials and Methods" section

rating (see e.g. Fig. 2), and yet the model can predict similar levels of task performance at short and long SOAs since task accuracy depends on $d' = \mu_{diff}/\sigma$. In this way, provided that the standard deviations of the evidence distributions can vary independently from their distance, the Single Channel model can capture the accuracy/visibility dissociation in the behavioral data (Fig. 4). Thus, in order for the Single Channel model to capture this data, it must assume that the variance of the internal signal is highest at long SOAs where task performance and visibility are maximal. Although such a Poisson-like correlation of signal and noise is not in itself implausible, the specific patterns predicted are some causes for doubt. For instance, the model predicts that on average, the signal μ_{diff} at SOAs 0 ms and 100 ms is roughly equal, and yet the simultaneous presentation of stimulus and mask gives rise to perceptual evidence that is less noisy than when their presentation is separated by a full 100 ms. It seems more likely that, controlling for the magnitude of the absolute signal, stimulus representations should be noisier when the mask is presented simultaneously than when the mask is presented 100 ms later.

The best-performing Dual Channel model predicts that perceptual sensitivity is greater in the "unconscious" channel than in the "conscious" channel for several short SOAs. Because this model resets visibility ratings to 1 when the two channels disagree on stimulus classification, setting the sensitivity of the "unconscious" channel higher at the short SOAs has the effect of increasing the frequency of disagreements between the two channels, thus reducing visibility at those SOAs without having a drastic effect on task performance. This allows visibility to be lower at shorter SOAs than at longer ones even though task performance at those SOAs is similar. However, the model only manages to produce a somewhat weak dissociation between task performance and stimulus visibility (Fig. 4). Furthermore, it seems unlikely that processing in an unconscious channel could be so robustly high and consistently superior than conscious processing across several SOAs.

The best-performing Hierarchical model predicts that perceptual evidence decays in the second stage of "subjective" processing more readily at short than at long SOAs, thus leading to lower overall levels of visibility at the short SOAs in spite of similar stimulus discrimination sensitivity. In contrast, the model supposes that noise at the late processing stage is independent of SOA. This seems plausible if we imagine that signal transmission depends in part on the processing that occurs in early sensory areas, whereas the noise intrinsic to later processing stages is independent of the noise in earlier stages.

The structure and parameter values of the Hierarchical model are also consistent with previous empirical findings from experiments focusing on the dissociation between objective task performance and subjective ratings of visibility. For instance, Lau and Passingham (2006) used a similar metacontrast masking paradigm as in the present study, and in the fMRI scanner they focused on a short and a long SOA where task performance was matched, and yet the subjective ratings of visual awareness differed. Higher subjective ratings of visibility at the long SOA were associated with higher level of activity in the dorsolateral prefrontal cortex. Interestingly, no significant difference in level of fMRI activity was found in the posterior sensory areas. This is compatible with the Hierarchical model if we assume that the prefrontal activity reflects the late stage process. Indeed, according to the parameter values of the best Hierarchical model (Fig. 7), reported visibility was higher at the long SOAs than it was at the earlier, performance-matched SOAs due to a superior transmission of perceptual evidence to the late processing stage (i.e. higher values for the parameter k). This corroborates well with the fMRI result.

In another study, Rounis *et al.* (2010) found that transcranial magnetic stimulation (TMS) to the prefrontal cortex selectively reduced the type 2 performance of participants' visibility ratings without affecting stimulus classification accuracy. Similar findings have been observed in neuropsychological patients with lesions to the prefrontal cortex (Fleming *et al.*, 2014). These results are compatible with the view that objective task performance is largely driven by sensory signals in posterior regions, which may reflect the early stage process in the Hierarchical model, whereas subjective ratings of stimulus visibility may depend on downstream mechanisms located in the prefrontal cortex.

One might worry that the design of the current experiment is biased in favor of the Hierarchical model. We required subjects to report stimulus visibility after they reported stimulus identity, with a second key press. Perhaps signal degradation did occur between the "objective" and "subjective" decisions, in a fashion predicted by the Hierarchical model, but only because the design forced subjects to report visibility after reporting their perceptual decisions. This timing difference between the two key presses could trivialize our findings.

However, the implicit reasoning behind this argument is that signal degradation could be artificially introduced by increasing response time. The longer the subject takes to respond, the more degraded a signal presumably becomes. If this deflationary account of the modeling results were true, we might expect that the Hierarchical model's estimated values of signal decay and late processing noise should correlate with the time separating the stimulus classification key press from the subjective rating key press (henceforth, "rating RT"). However, the across-SOA correlation between estimated signal decay and rating RT was not significant for any subject ($ps > 0.15$), and the average correlation did not differ from zero (Fisher's r-to-z transform, $P = 0.4$). We plot rating RT and the k parameter of the Hierarchical model as a function of SOA in Supplementary Fig. S5. This figure suggests that, if anything, the direction of the relationship between rating RT and signal decay is in the direction opposite to that posited by the trivializing account. The model fits predict the highest degrees of signal degradation when rating RT is smallest, rather than when it is largest. Since the parameter for late processing noise was constant across SOA for the best Hierarchical model, we cannot compute within-subject correlations of this parameter with rating RT. We did find that across subjects, the estimated amount of late noise correlated with average rating RT, $r = -0.48$, $P = 0.03$. However, this result is in the opposite direction of that proposed by the trivializing critique regarding two separate key presses. That is, longer rating RTs were associated with lower estimates of type 2 noise, rather than greater estimates.

Finally, we note that rating RT was not modulated by SOA ($P = 0.4$) and that the average rating RT was relatively small (426 ms). This suggests that the time between the first and the second key presses was mainly for motor preparation and execution, i.e. it is unlikely that subjects' decision making follows a linear process in which the decision about what visibility rating to produce is forestalled until after the button indicating the stimulus identity is pressed. Rather, the decision process for what visibility rating to produce is likely well underway even before the initial key press indicating stimulus identity. In our subjective experience, this is how one would perform the task as well. Taken together, these results suggest that the success of the Hierarchical model in fitting the data cannot be trivially attributed to the two key press design of the task. However, further experimental work is needed to shed more light on these issues.

Discussion

To compare models of how subjective reports of visibility relate to objective perceptual processing, we collected data from a metacontrast masking paradigm that has been shown to induce dissociations between stimulus classification accuracy and reported levels of visibility across different levels of stimulus-mask onset asynchroncy (SOA) (Lau and Passingham, 2006; for further discussion on using this paradigm to dissociate task performance and visual awareness, see Supplementary Materials, "Viability of the metacontrast masking paradigm for dissociating objective and subjective processing"). We reasoned that the dissociation between accuracy and visibility across SOA (Figs 3 and 4) would pose a challenge to models of perceptual decision making, and thus prove useful for distinguishing among them. The data contained another feature that also proved difficult for the models to capture: visibility ratings were not as predictive of task performance as would be expected under the traditional SDT model (Fig. 5). Overall, the Hierarchical model provided the best and most parsimonious fit to the data. The model parameters it used to fit the data also seem plausible (Fig. 7), and overall the model seems compatible with the previous empirical findings (Lau and Passingham, 2006; Rounis *et al.*, 2010; Fleming *et al.*, 2014; Maniscalco and Lau, 2015).

Why was the Hierarchical model successful where the Single Channel and Dual Channel models were not? The best-performing Hierarchical model (Constant Noise + Decay) was able to accommodate the relative dissociation between task performance and visibility ratings by supposing that early-stage perceptual processing is better transmitted to late-stage processing at long than at short SOAs. Because the early stage governs task performance and the late stage governs subjective reports, this allows for long SOAs to have higher subjective visibility than short SOAs in spite of having similar task performance. The best-performing Single Channel model (Changing Variance) was able to accommodate this pattern to some extent by supposing that perceptual processing becomes more variable at long SOAs, thus producing sensory signals that more frequently exceed the observer's criteria for producing high visibility ratings. However, although this model captured the gist of the performance–visibility dissociation, it sometimes produced too-high estimates of visibility ratings or too-low estimates of task performance (Fig. 4, lower left panel).

By comparison, none of the Dual Channel models we considered appeared to capture the performance–visibility dissociation particularly well. Our SDT implementation of the Independent Dual Channel model [which most closely followed the model of Del Cul *et al.* (2009); see Supplementary Material] essentially acts like a Single Channel model with added flexibility for adjusting task performance at the lowest level of subjective visibility. This provides only a relatively limited mechanism for adjusting the relationship between task performance and visibility; holding the parameters of the "conscious" channel constant, changes in the "unconscious" channel can only influence task performance to the extent that subjects report the lowest level of subjective visibility. Thus, this model can accommodate only relatively small differences in task performance for conditions with similar mean levels of reported visibility. Additionally, because task performance at higher (presumably conscious) visibility levels cannot be affected by changing parameters of the "unconscious" channel, this model makes the relatively strong prediction that whatever differences in task performance do occur for visibility-matched conditions, they should arise purely from differences in task performance for

trials with the lowest visibility rating. The best-performing Dual Channel model (Weighted Dual Channel) was somewhat more flexible, but still did not adequately capture the dissociation (Fig. 4, bottom center panel).

In addition to the performance–visibility dissociation across SOA, we also found that the models differed in their ability to capture the degree to which visibility ratings were diagnostic of accuracy on a trial to trial basis. Visibility ratings for incorrect responses at short and long SOAs were generally higher than the model fits (Fig. 5, top row), and visibility ratings' ability to predict accuracy was generally lower than the model fits (Fig. 5, bottom row). The Hierarchical model performed best at capturing these data because it posits that the sensory signal accrues additional noise at late processing stages. This reduces the information that such sensory signals carry regarding task performance on the trial level, which manifests as higher visibility for incorrect trials (type 2 FAR, Fig. 5, top row) and lower values for meta-d', an index of metacognitive performance (Fig. 5, bottom row). In contrast, the Single Channel model posits that the same sensory evidence is used to make both the objective response and the visibility rating, and thus is considerably less flexible in the relationships it allows between task performance and type 2 accuracy (Maniscalco and Lau, 2012; Maniscalco and Lau, 2014). Dual Channel models behaved similarly to Single Channel models in this respect, as they primarily differed with respect to processing at low levels of visibility.

One recent study (Scott *et al.*, 2014) suggests that in artificial grammar learning tasks subjects can even exhibit above-chance metacognitive accuracy when task performance is at chance, a phenomenon the authors named "blind insight." This suggests an added degree of freedom in the relationship between task performance and metacognition that is challenging for any SDT model to capture, including the Hierarchical model structure, as typical SDT formulations entail that metacognitive performance is constrained by task performance (Galvin *et al.*, 2003; Maniscalco and Lau, 2012, 2014). However, it should be noted that blind insight has thus far only been observed in AGL and not in perceptual tasks of the sort investigated here. In previous investigations of the relationship between metacognitive sensitivity and task performance in visual psychophysics tasks like the one used here, the information content of subjective ratings has been shown to be constrained by task performance in the sense that meta-d' $\leq$ d'(Maniscalco and Lau, 2012; Maniscalco and Lau, 2015), consistent with the findings of the present study (Fig. 5).

All models we tested were constructed using SDT as a basis (Fig. 2; Macmillan and Creelman, 2005). In this work, SDT provided an ideal basis to compare overall model architectures in a simple but powerful framework. SDT is sufficiently powerful to be able to dissociate perceptual sensitivity from response bias—essential for the study of perceptual decision making and subjective reports of visibility—while also being sufficiently general as to be readily adapted to different model architectures. Using the same SDT framework for all models also facilitated direct model comparison by minimizing idiosyncratic computational differences between the models. Because our SDT models captured the core computational principles lying behind broadly divergent theories of how perceptual decision making and subjective visibility are related, the model comparison analysis sheds light on these broad conceptual issues.

One limitation to this approach is that the conclusions we have drawn may be somewhat specific to the particular SDT implementations we have used. [However, see the Supplementary Material for evidence that our SDT implementation of the Independent Dual Channel model behaves similarly to the Dual Channel accumulation model in Del Cul *et al.* (2009).] Nonetheless, the relative simplicity of the SDT models we have chosen, in conjunction with the broad differences in the model classes being compared (Fig. 1), would seem to mitigate such concerns. We have also endeavored to perform an unusually comprehensive analysis that directly compares a wide range of models' ability to account for the data, rather than simply demonstrating that a single model can produce reasonable fits to the data.

We also acknowledge that this analysis is driven by the current data set and is thus limited by its generalizability. For instance, it is possible that a Dual Channel model may perform better for other kinds of phenomena (Del Cul *et al.*, 2009; Charles *et al.*, 2013; Charles *et al.*, 2014). Though the important moral is, in order to make claims that a certain empirical finding support a particular model, we need to compare the fit against alternative models. Future work employing similar formal comparison strategies needs to be performed in these cases.

Are the models biologically realistic?

On the face of it, the models we considered depict a purely feedforward style of information processing. What of the fact that anatomically, the most related brain regions are linked by both feedforward and feedback connections? For instance, for the Hierarchical model it is perhaps natural to think of the first stage as representing processing in the early sensory regions in the brain, and the second stage as representing processing in higher regions such as the prefrontal cortex. In this sense, the model ignores the presence of top-down modulations from prefrontal cortex to early sensory areas. However, formally the model does not necessarily commit to such anatomical identifications. Strictly speaking, the model is agnostic as to whether the late stage is mediated by a feedforward or feedback process; late stage simply means it is late in the stream of information processing and thereby inherits the noise of earlier stages.

Even on the plausible and intuitive interpretation that in the Hierarchical model the first stage reflects early sensory processes and the second stage fronto-parietal processes, the model does not deny the existence of feedback connections. Nor does it deny the existence of parallel pathways as intuitively depicted by the Dual Channel model. The Hierarchical model suggests that "with respect to explaining" the relationships and potential dissociations of objective stimulus responses and subjective visibility ratings, the essential relevant structure of processing is hierarchical. This does not mean that the Hierarchical model explains all facts regarding brain processes or subjective experience. It is for the same reasons that the Single Channel model cannot be rejected on the grounds that the brain is clearly more complex than a single-stage processor.

Implications for theories of visual awareness

One currently popular theory suggests that feedback, and specifically feedback from extrastriate to primary visual cortex, is essential for visual awareness (Lamme, 2006; Block, 2007). One might take the point of view that the feedforward wave of processing from primary visual cortex to extrastriate areas represents an early stage of processing, and that feedback represents a second stage of processing, such that together they form a hierarchy.

Another dominant theory of visual awareness is the global workspace theory (Dehaene *et al.*, 2003; Dehaene *et al.*, 2006),

according to which early sensory processing itself does not support conscious experience. To enter consciousness, the early perceptual signal must propagate into a second stage of processing mediated by a global workspace structure located in prefrontal and parietal cortices. Considerations like these may give the impression that both theories of visual awareness discussed above are compatible with the Hierarchical model.

However, it is important to emphasize that the present work focuses on the dissociation between objective task performance and subjective reports. According to the Hierarchical model, manipulation of the second stage of processing changes subjective reports but not task performance. But the feedback model and the global workspace model would not make such predictions. In these models, the supposed second stage of processing supports both subjective experience as well as amplification of the perceptual signal itself, which is essential for objective task performance. Thus, according to these theories, if the second stage of processing (feedback to striate cortex, or global workspace activity) is disrupted, both objective task performance and subjective reports will be affected. Therefore, these models bear more functional resemblance to the Single Channel models than the Hierarchical models. In order for such theories to obtain a reduction in subjectively reported level of awareness while keeping task performance constant, one natural solution would be that the perceptual signal from a separate, unconscious channel (e.g. a subcortical route) would need to be increased to compensate for the signal loss in the "conscious" channel. In other words, a Dual Channel model would need to be stipulated.

Therefore, as far as dissociations between task performance and subjective reports are concerned (e.g. when we are specifically trying to explain the kind of performance-matched difference in subjective rating and type 2 performance depicted in Figs 4 and 5), both aforementioned theories are more congenial with Single Channel and Dual Channel models than with Hierarchical models (Del Cul *et al.*, 2009; Lau, 2011). The present results are thus surprising, or maybe even problematic, for these theories.

Potential relations to the memory literature

It has been proposed that there are two distinct and dissociable memory systems, one supporting explicit, "conscious" recollection, and the other more relevant for vaguer judgments of familiarity or feelings of knowing, or unconscious priming behavior (e.g. Jacoby, 1991; Hintzman and Curran, 1994). However, it has also been argued that a single system view is more parsimonious (Squire *et al.*, 2007; Wixted, 2007; Berry *et al.*, 2008), and that the apparent dissociation between conscious recollection and unconscious memory is due to different levels of activation within the same system. Our study may contribute to this controversy, because the paradigms used in some of these memory studies are conceptually very similar to the paradigm used here: subjects make an objective judgment about the state of the world (identity of visual stimulus, or whether an item has been presented previously or not), and then make a subjective judgment about how they subjectively feel about the first-order process (high versus low visibility, or "Remember" vs "Know" in some memory studies). Here we offer a third alternative to this debate between a single system versus two dissociated systems: it could be that there are two processes that work in hierarchy. Future studies may employ the same model comparison method to arbitrate which is the best model for memory function by fitting the models to experimental data where the objective memory performance and the subjective reports of recollection experience dissociate.

Conclusion

Here we introduce a distinction between different signal processing architectures supporting the generation of subjective reports of visual awareness. Above we discussed some limitations of this approach, such as that it depends on the specific fitted dataset. Regardless of whether these results hold true, one important message is that we can go beyond the traditional assumption that perception depends on a single decision-making process (Macmillan and Creelman, 2002). These simple single process models have enjoyed great success in explaining many aspects of perception, and remain powerful contenders because of their simplicity, as shown in our model comparison analysis (which punishes complex models). But in cases where objective task performance and subjective reports dissociate, it may be important to consider perceptual decision models that postulate more than a single process, at least as possibilities. Our investigation suggests that, of the two models which postulate two processes, the Hierarchical model is superior to the Dual Channel model.

Supplementary data

Supplementary data is available at *Neuroscience of Consciousness Journal* online.

Acknowledgements

This work is partially supported by a grant from the Templeton Foundation (6-40689) and the National Institutes of Health (NIH R01 NS088628-01). We thank Anil Seth and Alex Pouget for useful discussion.
Data available upon request.

Conflict of interest statement. None declared.

References

Azzopardi, P, Cowey, A. Blindsight and visual awareness. *Conscious Cogn* 1998;**7**:292–311.

Berry, CJ, Shanks, DR, Henson, RNA. A unitary signal-detection model of implicit and explicit memory. *Trends Cogn Sci* 2008;**12**: 367–73.

Block, N. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav Brain Sci* 2007;**30**: 481–99; discussion 499–548.

Brainard, DH. The psychophysics toolbox. *Spat Vis* 1997;**10**:433–6.

Brown, S, Steyvers, M. The dynamics of experimentally induced criterion shifts. *J Exp Psychol Learn Memory Cogn* 2005;**31**:587–99.

Burnham, KP, Anderson, DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Berlin; Heidelberg, Germany; New York: Springer, 2002.

Charles, L, King, JR, Dehaene, S. Decoding the dynamics of action, intention, and error detection for conscious and subliminal luli. *J Neurosci* 2014;**34**:1158–70.

Charles, L, Van Opstal, F, Marti, S *et al.* Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage* 2013;**73**:80–94.

Clarke, FR, Birdsall, TG, Tanner, J. Two types of ROC curves and definitions of parameters. *J Acoust Soc America* 1959;**31**:629–30.

Cleeremans, A, Timmermans, B, Pasquali, A. Consciousness and metarepresentation: a computational sketch. *Neural Net* 2007;**20**:1032–9.

Davidson, M, Persaud, N, Maniscalco, B *et al*. Awareness-related activity in prefrontal and parietal cortices reflects more than superior performance capacity: a blindsight case study. *J Vis* 2010;**10**:897.

Dehaene, S, Changeux, J-P, Naccache, L *et al*. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci* 2006;**10**:204–11.

Dehaene, S, Sergent, C, Changeux, J-P. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc Natl Acad Sci USA* 2003;**100**:8520–5.

Del Cul, A, Dehaene, S, Reyes, P *et al*. Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain* 2009;**132**:2531–40.

Dorfman, DD, Alf, E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–rating-method data. *J Math Psychol* 1969;**6**:487–96.

Donkin, C, Newell, BR, Kalish, M *et al*. Identifying strategy use in category learning tasks: a case for more diagnostic data and models. *J Exp Psychol Learn Memory Cogn* 2015;**41**:933–48.

Fetsch, CR, Kiani, R, Newsome, WT *et al*. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* 2014;**83**:797–804.

Fleming, SM, Ryu, J, Golfinos, JG *et al*. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* 2014;**137**:2811–22.

Fleming, SM, Weil, RS, Nagy, Z *et al*. Relating introspective accuracy to individual differences in brain structure. *Science* 2010;**329**:1541–3.

Galvin, SJ, Podd, JV, Drga, V *et al*. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon Bull Rev* 2003;**10**:843–76.

Gold, JI, Shadlen, MN. The neural basis of decision making. *Ann Rev Neurosci* 2007;**30**:535–74.

Gorea, A, Sagi, D. Failure to handle more than one internal representation in visual detection tasks. *Proc Natl Acad Sci* 2000;**97**:12380–4.

Hintzman, DL, Curran, T. Retrieval dynamics of recognition and frequency judgments: evidence for separate processes of familiarity and recall. *J Memory Lang* 1994;**33**:1–18.

Jacoby, LL. A process dissociation framework: Separating automatic from intentional uses of memory. *J Memory Lang* 1991;**30**:513–41.

Jolij, J, Lamme, VAF. Repression of unconscious information by conscious processing: evidence from affective blindsight induced by transcranial magnetic stimulation. *Proc Natl Acad Sci USA* 2005;**102**:10747–51.

Kepecs, A, Uchida, N, Zariwala, HA *et al*. Neural correlates, computation and behavioural impact of decision confidence. *Nature* 2008;**455**:227–31.

Kirkpatrick, S, Gelatt, CD, Jr, Vecchi, MP. Optimization by simulated annealing. *Science* 1983; **220**:671–80.

Komura, Y, Nikkuni, A, Hirashima, N *et al*. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat Neurosci* 2013;**16**:749–55.

Lak, A, Costa, GM, Romberg, E *et al*. Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* 2014;**84**:190–201.

Lamme, VAF. Towards a true neural stance on consciousness. *Trends Cogn Sci* 2006;**10**:494–501.

Lau, H. Theoretical motivations for investigating the neural correlates of consciousness. *Wiley Interdis Rev Cogn Sci* 2011;**2**:1–7.

Lau, H. Volition and the functions of consciousness. In: Gazzaniga, M (ed). The *Cognitive Neurosciences*, 4th edn. New York, NY: MIT Press, 2009, 1191–200.

Lau, HC, Passingham, RE. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc Natl Acad Sci USA* 2006;**103**:18763–8.

Lau, H, Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci* 2011;**15**:365–73.

Macmillan, NA, Creelman, CD. *Detection Theory: A User's Guide*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum, 2005.

Maniscalco, B, Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn* 2012;**21**:422–30.

Maniscalco, B, Lau, H. Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neurosci Conscious* 2015:niv002.

Maniscalco, B, Lau, H. Signal detection theory analysis of type 1 and type 2 data: meta-d', response-specific meta-d', and the unequal variance SDT model. In: Fleming, SM, Frith, CD (eds), *The Cognitive Neuroscience of Metacognition*. Berlin; Heidelberg, Germany; New York: Springer, 2014, 25–66.

Metcalfe, J, Shimamura, AP. *Metacognition: Knowing about Knowing*. Cambridge, MA: The MIT Press, 1996.

Milner, AD, Goodale, MA. *The Visual Brain in Action*. Cambridge, MA: Oxford University Press, 1996.

Morewedge, CK, Kahneman, D. Associative processes in intuitive judgment. *Trends Cogn Sci* 2010; **14**:435–40.

Mueller, ST, Weidemann, CT. Decision noise: an explanation for observed violations of signal detection theory. *Psychon Bull Rev* 2008; **15**:465–94.

Pleskac, TJ, Busemeyer, JR. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol Rev* 2010; **117**:864–901.

Pollack, I, Hsieh, R. Sampling variability of the area under the ROC-curve and of d'e. *Psychol Bull* 1969;**71**:161–73.

Rahnev, D, Maniscalco, B, Graves, T *et al*. Attention induces conservative subjective biases in visual perception. *Nat Neurosci* 2011;**14**:1513–5.

Rahnev, DA, Maniscalco, B, Luber, B *et al*. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J Neurophysiol* 2012; **107**:1556–63.

Ratcliff, R. A theory of memory retrieval. *Psychol Rev* 1978; **85**:59–108.

Ratcliff, R, McKoon, G. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput* 2008; **20**:873–922.

Ratcliff, R, Starns, JJ. Modeling confidence and response time in recognition memory. *Psychol Rev* 2009; **116**:59–83.

Rouder, JN, Morey, RD, Cowan, N *et al*. An assessment of fixed-capacity models of visual working memory. *Proc Natl Acad Sci USA* 2008; **105**:5975–9.

Rounis, E, Maniscalco, B, Rothwell, JC *et al*. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn Neurosci* 2010; **1**:165.

Scott, RB, Dienes, Z, Barrett, AB *et al*. Blind insight: metacognitive discrimination despite chance task performance. *Psychol Sci* 2014; **25**:2199–208.

Smith, JD. The study of animal metacognition. *Trends Cogn Sci* 2009; **13**:389–96.

Squire, LR, Wixted, JT, Clark, RE. Recognition memory and the medial temporal lobe: a new perspective. *Nat Rev Neurosci* 2007; **8**:872–83.

van Gaal, S, Ridderinkhof, KR, Fahrenfort, JJ *et al*. Frontal cortex mediates unconsciously triggered inhibitory control. *J Neurosci* 2008; **28**:8053–62.

van Gaal, S, Ridderinkhof, KR, van den Wildenberg, WPM *et al*. Dissociating consciousness from inhibitory control: evidence for unconsciously triggered response inhibition in the stop-signal task. *J Exp Psychol Hum Percept Perform* 2009; **35**:1129–39.

van Gaal, S, Ridderinkhof, KR, Scholte, HS *et al*. Unconscious activation of the prefrontal no-go network. *J Neurosci* 2010; **30**:4143–50.

Vlassova, A, Donkin, C, Pearson, J. Unconscious information changes decision accuracy but not confidence. *Proc Natl Acad Sci* 2014; **111**:16214–8.

Watson, AB, Pelli, DG. QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys* 1983; **33**:113–20.

Weiskrantz, L. *Blindsight: A Case Study and Implications*. Oxford: Oxford University Press, 1986.

Wilimzig, C, Tsuchiya, N, Fahle, M *et al*. Spatial attention increases performance but not subjective confidence in a discrimination task. *J Vis* 2008; **8**:7.1–10.

Wixted, JT. Dual-process theory and signal-detection theory of recognition memory. *Psychol Rev* 2007;**114**:152–76.