# Modeling categorization of scenes containing consistent versus inconsistent objects

**Michael L. Mack** Department of Psychology, Vanderbilt University, Nashville, TN, USA

**Thomas J. Palmeri** Department of Psychology, Vanderbilt University, Nashville, TN, USA

How does object perception influence scene perception? A recent study of ultrarapid scene categorization (O. R. Joubert, G. A. Rousselet, D. Fize, & M. Fabre-Thorpe, 2007) reported facilitated scene categorization when scenes contained consistent objects compared to when scenes contained inconsistent objects. One proposal for this consistent-object advantage is that ultrarapid scene categorization is influenced directly by ultrarapid recognition of particular objects within the scene. We instead asked whether a simpler mechanism that relied only on scene categorization without any explicit object recognition could explain this consistent-object advantage. We combined a computational model of scene recognition based on global scene statistics (A. Oliva & A. Torralba, 2001) with a diffusion model of perceptual decision making (R. Ratcliff, 1978). This model is sufficient to account for the consistent-object advantage. The simulations suggest that this consistent-object advantage need not arise from ultrarapid object recognition influencing ultrarapid scene categorization, but from the inherent influence certain objects have on the global scene statistics diagnostic for scene categorization.

Keywords: categorization, computational modeling, scene recognition, structure of natural images

Citation: Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision, 10*(3):11, 1–11, http://journalofvision.org/10/3/11/, doi:10.1167/10.3.11.

## Introduction

What is the relationship between scene recognition and object recognition? Past research has examined how objects are recognized in semantically consistent or inconsistent scenes (e.g., Biederman, Mezzanotte, & Rabinowitz, 1982; Davenport & Potter, 2004; Palmer, 1975). The general finding is that it is easier to recognize objects in semantically consistent scenes, such as recognizing a toaster in a kitchen compared to recognizing a toaster in a bedroom (Davenport & Potter, 2004; Henderson & Hollingworth, 1999; Palmer, 1975).

One proposed mechanism for facilitated recognition of objects contained in consistent scenes is an interacting, dual-system account (Davenport, 2007; Davenport & Potter, 2004). At the same time that the object recognition system is extracting information for an object categorization, the scene recognition system is extracting evidence for a scene categorization. Recognition of the scene activates representations of objects typically found in that kind of scene thereby facilitating categorizations of scene-consistent objects. While the interacting, dual-system account is supported by evidence for scene recognition facilitating object recognition, the converse should be true as well. Indeed, Davenport and Potter (2004) found that scene categorization was facilitated when the scene contained a consistent object (e.g., a football field with a football player) compared to an inconsistent object (e.g., a football field with a priest). Object and scene recognition operates in parallel activating probable object and scene representations to converge on a full description of the environment.

A recent study (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007) reported a similar advantage for scenes containing consistent objects versus inconsistent objects in ultrarapid scene categorization. Participants were presented with scenes for 26 ms and performed a speeded go/no-go decision about the scene's superordinate category (*natural* versus *man-made*). Most of the scenes in their stimulus set contained no salient objects, but a portion of the scene images (see Figure 1) either contained salient objects consistent with the scenes' category (e.g., an urban street scene with a parked car categorized as man-made) or contained salient objects inconsistent with the scenes' category (e.g., an urban street scene with a large tree categorized as man-made).[1] A post-hoc analysis of these different types of scenes revealed that scenes containing an inconsistent object were categorized less accurately and more slowly than scenes containing a consistent object.

The dual-system account explained above provides a straightforward explanation of Joubert et al.'s finding of a disadvantage for categorizing scenes with inconsistent objects: The object recognition system activates semantically related scene category representations that influence

doi: 10.1167/10.3.11 Received June 5, 2009; published March 26, 2010 ISSN 1534-7362 

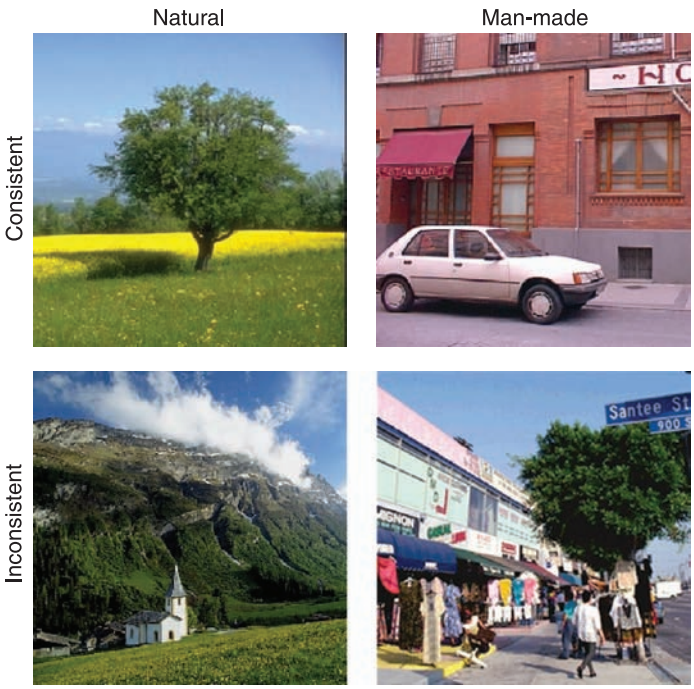


Figure 1. Examples of scene stimuli. Natural scenes (left) or man-made scenes (right) are shown containing consistent objects (top) or inconsistent objects (bottom).

the rapid processing and categorization decision made by the scene recognition system. For a scene containing an inconsistent object, the semantic information from the object conflicts with the evidence for the scene's category leading to more errors and slower reaction times. Perhaps the most intriguing aspect of this explanation is the relative timing of object and scene processing; object information is extracted and mapped onto conceptual representations fast enough to influence ultrarapid decisions about a scene's naturalness.

Joubert et al. used real scenes selected from a large database that happened to contain consistent objects, inconsistent objects, or no obvious objects. While using real scenes is a far better representation of everyday visual experience, a potentially unfortunate consequence of this realness is that potentially important visual information across scenes containing consistent or inconsistent objects cannot be controlled. Consider other previous studies (Davenport, 2007; Davenport & Potter, 2004) that tested the semantic influence of object information on scene categorization. Contrasting objects (e.g., a football player or a priest) were pasted into contrasting scenes (e.g., a football stadium or a church), forgoing a bit of reality for a full factorial combination of objects and scenes. With this level of control, any effect of object consistency on scene categorization is likely caused by semantic influences (Davenport, 2007; Davenport & Potter, 2004). The influence of object consistency on scene categorization found by Joubert et al. could also be based on semantic information from object recognition as proposed by the dual-system account. However, with uncontrolled real scenes, it could arise from differences in visual information driving scene categorization without any object recognition, as we demonstrate in this article.

Previous work has shown that ultrarapid scene categorization is largely determined by coarse, global scene properties (Oliva & Schyns, 1997; Schyns & Oliva, 1994). With just a glance at a scene, global properties, such as naturalness, are perceived rapidly even before a scene's basic-level category is determined or its objects are recognized (Greene & Oliva, 2009b). Furthermore, computational models that represent scenes based on their global spatial structure are sufficient for ultrarapid scene categorization (Oliva & Torralba, 2001). Such models capture only the diagnostic global features of scenes without explicitly representing any local content of the scene, such as the location, presence, or identity of particular objects (Greene & Oliva, 2009a). The feature set used by these models is based on global image statistics calculated across the entire scene, such as the scene's spatial frequency content. Linear combinations of certain spatial frequency components correlate with subjective judgment of scene characteristics such as openness and naturalness (Oliva & Torralba, 2001). Other proposals for global representations of scenes based on visual features including texture elements (Renninger & Malik, 2004) and SIFT descriptors (Lazebnik, Schmid, & Ponce, 2006) have also successfully accounted for scene categorization.

In the current study, we asked whether the consistent-object advantage in ultrarapid scene categorization reported by Joubert et al. could be predicted entirely by scene categorization mechanisms proposed by Oliva and colleagues without any explicit object recognition whatsoever.

Consider a forest scene. A small shed in that scene would be considered an inconsistent object (see Footnote 1). We could replace that shed with a consistent object, say a large bush. The global image statistics of a forest scene with a small shed will only be slightly different from those of a forest scene with a large bush. But they will not be identical. And that's the key. While perhaps quite small, is the difference in image statistics between scenes containing consistent objects versus scenes containing inconsistent objects sufficient to account for the consistent-object advantage? If so, then the consistent-object advantage in ultrarapid scene categorization can be explained by scene categorization alone, without any explicit object recognition. Critically, instead of semantic information activated by an object recognition system, the influence of objects on ultrarapid scene categorization arises from subtle changes in a scene's global representation of visual information.

To explore this possibility, we combined a computational model of scene recognition based solely on global scene statistics (Oliva & Torralba, 2001) with a diffusion model of perceptual decision making (Ratcliff, 1978).

Interpretation of global scene statistics provides evidence that drives a stochastic diffusion of perceptual evidence to a decision threshold. The model aims to explain both response probabilities and reaction time distributions for categorizing scenes containing consistent versus inconsistent objects. The model includes no explicit object recognition.

This paper is organized as follows: We first attempt a replication of the consistent-object advantage in scene categorization with both go/no-go and two-alternative, forced-choice paradigms. We then analyze the behavioral data using the pure diffusion model, for reasons that will be made apparent. Finally, we present fits to observed data of our computational model combining a scene categorization front-end with the diffusion model of decision making.

# Behavioral experiments

The consistent-object advantage in the Joubert et al. study was discovered through a post-hoc analysis of data from a subset of the experimental trials. The current experiments attempted to replicate Joubert et al. (2007), but with an explicit manipulation of object consistency. Following Joubert et al., we used natural scenes selected from a larger database of scenes. We focused on the specific comparison of scenes with consistent objects versus scenes with inconsistent objects. We first attempted a replication of Joubert et al.'s go/no-go scene categorization paradigm. With the goal of applying the diffusion model (Ratcliff, 1978) to this behavioral data, we also conducted a two-alternative, forced-choice (2AFC) version of the scene categorization task. While recent work (Gomez, Ratcliff, & Perea, 2007) has clarified how the diffusion model can account for go/no-go perceptual decisions, the diffusion model was developed for and is commonly applied to 2AFC decision making paradigms (Ratcliff & Rouder, 1998). Due to the similarity in methods and results between these two experiments, we present both experiments together.

## Methods

### Participants

Fifty Vanderbilt University undergraduate students (twenty-six female; age 18–23 years) participated in the go/no-go experiment. Twenty-three different Vanderbilt University undergraduate students (thirteen female; age 18–24 years) participated in the 2AFC experiment. All participants were compensated with course credit.

### Stimuli

The stimuli consisted of color images of naturalistic scenes from an online image database (Oliva & Torralba, 2001). Scene images were divided into categories of natural and man-made environments. The natural scene category included images of beaches, fields, mountains, and forests and the man-made scene category included images of skyscrapers, urban cities, and streets. Only scenes with salient objects were used in the behavioral experiments. Two independent observers tagged scenes that contained a salient object that was consistent or inconsistent with the scenes' natural or man-made category (reliability = 0.93). 192 natural scenes (64 with inconsistent objects) and 192 man-made scenes (64 with inconsistent objects) were randomly selected from the database for the experiments. This set of 384 scene images was used for every participant. Scene images were presented in color and subtended 10.2° × 10.2° of visual angle. Example stimuli are shown in Figure 1.

### Procedure

*Go/No-go Experiment*. In the go/no-go experiment, participants were randomly assigned to a "go" category of natural or man-made scenes (with the other category assigned "no-go"). On each trial, a fixation cross was presented for 500–800 ms followed by a brief presentation of the scene image for 26 ms. Participants were instructed to press the response key if the scene belonged to the target category and withhold any response otherwise. Responses could be made for 1000 ms after onset of the scene image and any responses made after this time window were considered no-go responses because the experiment moved on to the next trial. The trial concluded with a 500 ms blank period before the next trial began. The experiment consisted of two blocks of 192 trials with an even split of target and distractor trials. Scene images used as target trials for half of the participants served as distractors for the other half of participants. The entire experiment lasted approximately 25 minutes.

*2AFC Experiment*. The 2AFC experiment consisted of the same trial sequence as the go/no-go experiment. Participants were instructed to respond to the category of the scene (natural or man-made) by pressing one of two labeled response keys. Responses could be made for 1000 ms after onset of the scene image. The experiment consisted of two blocks of 192 trials with an even proportion of natural and man-made scenes. The entire experiment lasted approximately 25 minutes.

## Results

*Go/No-go Experiment*. Accuracy and reaction times for correct responses was analyzed separately by target category (natural or man-made scene) (Figure 2). For both target category groups, we observed a significant consistent-object advantage, with higher accuracy for scenes containing consistent objects compared to incon-

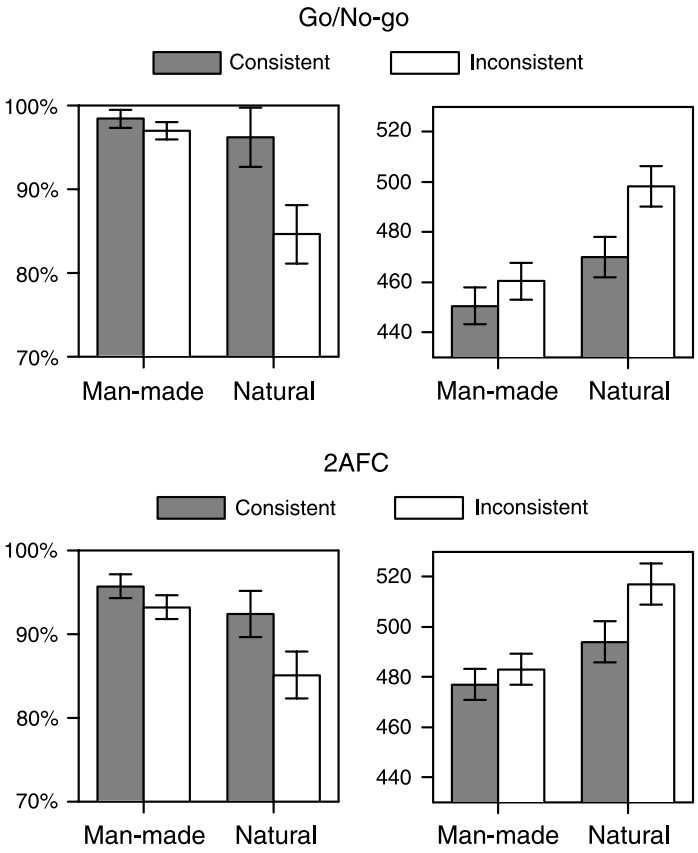


Figure 2. Behavioral results for the go/no-go (top) and 2AFC (bottom) decisions from Experiment 1. Average accuracy (left) and RT for correct responses (right) for consistent-object scenes (dark bars) and inconsistent-object scenes (white bars). Error bars represent 95% confidence intervals of consistent- versus inconsistent-object comparison.

sistent objects; this effect was larger for the natural scene group (11.6% difference; paired Wilcoxon test: $Z = 4.17$, $p < 0.001$) than the man-made scene group (1.4% difference; $Z = 2.648$, $p = 0.008$). Both groups also showed a consistent-object advantage in mean reaction times, with faster responses to scenes containing consistent objects; the effect was larger for the natural scene group (28 ms difference; $Z = 4.167$, $p < 0.001$) than the man-made scene group (10 ms difference; $Z = 2.435$, $p = 0.015$).

*2AFC Experiment*. Accuracy and reaction time for correct response were analyzed in the same manner (Figure 2). Results largely replicated those found in the go/no-go experiment. Both target category groups showed a significant consistent-object advantage, with higher accuracy for scenes containing consistent objects compared to inconsistent objects; this effect was larger for the natural scene group (7.4% difference; paired Wilcoxon test: $Z = 4.05$, $p < 0.001$) than the man-made scene target group (2.5% difference; $Z = 2.97$, $p = 0.003$). Both groups also showed a consistent-object effect in mean reaction times, with faster responses to scenes containing consistent objects; the effect was larger for the natural scene group (23 ms difference; $Z = 3.92$, $p < 0.001$) than the man-made scene group (7 ms difference; $Z = 1.97$, $p = 0.049$).

## Discussion

We replicated the consistent-object advantage in a go/no-go scene decision (Joubert et al., 2007) and found converging evidence for the consistent-object advantage in a 2AFC scene categorization decision. For both man-made and natural scene targets, scenes containing consistent objects were categorized faster and with fewer errors than scenes containing inconsistent objects. The consistent-object advantage was larger for natural scenes, but this may be explained by stimulus factors; we did not attempt to equate the natural and man-made scene images in terms of visual properties or similarity. A similar difference in magnitude of the consistent-object advantage for natural and man-made scenes was reported by Joubert et al. (2007).

# Diffusion model analysis

The diffusion model is a well-known model of perceptual decision making (Ratcliff, 1978). Decisions are made by a stochastic accumulation of noisy evidence over time toward a decision threshold (Figure 3). The rate of accumulation (called the drift rate, $v$) is determined by the quality of the perceptual evidence. Higher-quality evidence leads to faster accumulation and faster reaction times. Changing the decision threshold ($a$) affects the tradeoff between speed and accuracy. Overall reaction time is given by the time for the perceptual decision made by the diffusion plus time for non-decision factors ($T_{er}$) such as stimulus encoding and motor response times. In the full diffusion model, variability in drift rate, starting point, and nondecision time can be present and allow for the diffusion model to account for more detailed patterns

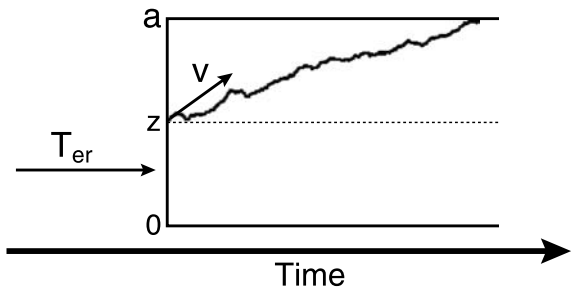


Figure 3. Illustration of the diffusion model. At starting point $z$, noisy evidence accumulates with a drift rate, $v$, towards decision bounds at 0 and $a$. Overall reaction time is given by the time of stochastic accumulation plus time for non-decision factors ($T_{er}$).

of reaction time distributions (Ratcliff & Rouder, 1998). However, since simple differences in mean reaction times and accuracy are typically predicted by changes in drift rate, decision threshold, and nondecision time (e.g., Wagenmakers, van der Maas, & Grasman, 2007), we focused our analysis on these three parameters.

The diffusion model is typically applied to 2AFC categorization. A recent paper extended the diffusion model to account for go/no-go categorization as well (Gomez et al., 2007). They tested two versions of the diffusion model, one where evidence accumulates towards a single decision boundary for the "go" response with the other boundary at negative infinity, and another where evidence accumulates to both "go" (explicit response) and "no-go" (no response) boundaries. The two-boundary model was found to provide the best account of behavior associated with several go/no-go categorization tasks (Gomez et al., 2007). Therefore, we modeled the data from the go/no-go scene categorization experiment using a two-boundary diffusion model, with one boundary for a go response and the other boundary for a no-go non-response.

Before combining the diffusion model with a scene-recognition front end, we used the pure diffusion model as a data analysis device in order to pinpoint the source of the consistent-object advantage in accuracy and reaction time. The consistent-object advantage could arise from a variety of difference sources in the diffusion model:

First, it could arise from a difference in the time to perceptually process and encode scenes containing consistent versus inconsistent objects. In other words, even though the advantage is observed behaviorally in terms of the overt task response, the advantage is not because of faster decisions per se. Instead, perceptual processing is more efficient when scenes contain consistent objects. This would be reflected by a difference in the $T_{er}$ parameter.

Second, recognizing consistent versus inconsistent objects might bias the decision process. Imagine that ultrarapid object recognition detects a man-made object or a natural object. Clearly, the task does not ask for an object recognition response. However, the results of object recognition could bias the scene categorization decision, leading to a difference in the decision threshold of the accumulation process (the $a$ parameter) in the direction consistent with the ultrarapid object recognition.

Finally, the advantage could arise from a difference in the quality of the perceptual evidence driving the accumulation process. This would be reflected in a difference in drift rate, $v$. As we will detail later, the drift hypothesis is most consistent with an hypothesized account where scene perception mechanisms alone lead to the consistent-object advantage because of differences in global scene statistics.

## Model fitting

The diffusion model was fitted to reaction time distributions from the two experiments using standard techniques (see Ratcliff & Tuerlinckx, 2002) with the Diffusion Model Analysis Toolbox (Vandekerckhove & Tuerlinckx, 2008). For each individual participant, RT data for scenes containing consistent versus inconsistent objects were grouped into 6 RT bins defined by the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles. To fit the go/no-go data, an additional bin was included to count the number of no-go responses. To fit the two-choice data, correct and error RTs were grouped separately into RT bins. Quantile RTs averaged across participants were then used to generate predicted cumulative distributions of response probabilities (Vandekerckhove & Tuerlinckx, 2007, 2008). Best-fitting model parameters were found using the SIMPLEX method that minimized the Pearson chi-square ($\chi^2$) for the observed versus predicted number of RTs within each RT bin; we also report Bayesian Information Criterion (BIC) statistics for model fits, which can be characterized as a maximum likelihood measure with a term that penalizes a model for its number of free parameters (Schwarz, 1978). The full diffusion model is defined by seven parameters: starting point of the accumulation process and its variability ($z$, $s_z$), decision threshold ($a$), drift rate and its variability ($v$, $\eta$), and the nondecision time and its variability ($T_{er}$, $s_t$). For our model fits, starting point ($z = a/2$) and its variability, variability of drift rate ($\eta$), and variability of nondecision time ($s_t$) were held constant across the consistent and inconsistent conditions; fitting such a highly parameterized model requires data from more conditions than we had. So following other recent work with the diffusion model (Wagenmakers et al., 2007), we fitted versions of the model where only the three key parameters, decision threshold ($a$), nondecision time ($T_{er}$), and drift rate ($v$), were free to vary or were held constant across the consistent and inconsistent conditions (see also, Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Grasman, Wagenmakers, & van der Maas, 2009; Matzke & Wagenmakers, 2009); we compared these versions of the diffusion model with a version where all three parameters were fixed across consistent and inconsistent conditions. In a sense, we are using the diffusion model as a data analysis tool in much the same way that psychophysicists routinely use signal detection theory.

## Results

Table 1 displays values for the chi-square statistic, BIC statistic, and the appropriate significance tests for each version of the diffusion model fitted to the data from the go/no-go and 2AFC experiments (see Appendix A for best-fitting parameter values). For the natural scene condition, the variant of the diffusion model with only drift rate as a free parameter provided a significantly better fit to the behavioral data than variants with only nondecision time or decision threshold as a free parameter. For the man-made scene condition, none of the diffusion model variants with a free parameter provided

| | Free parameters | BIC | $\chi^2$ | $p$ (vs. all fixed) |
|---|---|---|---|---|
| Go/No-go | fixed | 1621.1 | 1579.3 | – |
| Natural | $a$ | 1623.7 | 1576.1 | 0.071 |
| | $T_{er}$ | 1625.2 | 1577.6 | 0.184 |
| | **$v$** | **1606.6** | **1558.9** | **<0.001** |
| Go/No-go | fixed | 1345.6 | 1303.9 | – |
| Man-made | $a$ | 1350.7 | 1303.1 | 0.339 |
| | $T_{er}$ | 1350.5 | 1302.9 | 0.301 |
| | $v$ | 1350.8 | 1303.1 | 0.371 |
| 2AFC | fixed | 998.1 | 959.6 | – |
| Natural | $a$ | 1001.3 | 957.3 | 0.133 |
| | $T_{er}$ | 1003.1 | 959.0 | 0.462 |
| | **$v$** | **983.1** | **939.1** | **<0.001** |
| 2AFC | All fixed | 843.6 | 805.1 | – |
| Man-made | $a$ | 848.6 | 804.5 | 0.462 |
| | $T_{er}$ | 848.4 | 804.3 | 0.392 |
| | $v$ | 848.9 | 804.9 | 0.708 |

Table 1. Chi-square and BIC values for diffusion model fits to data from the natural and man-made scene target conditions of the go/no-go and 2AFC decisions from Experiment 1. The last column shows $p$ values for comparisons to the baseline model with all parameters held constant across the consistent- and inconsistent-object conditions (bold values indicate significantly better fits than the baseline model).

a better fit than the baseline model with all parameters held constant across the consistent and inconsistent object conditions; it is likely that the small magnitude of the consistent-object advantage for man-made scenes contributed to this nonidentifiability.

## Discussion

Diffusion model analyses revealed that a model with separate drift rates for the consistent and inconsistent object condition provided the better account of the behavioral data than models with separate residual times ($T_{er}$) or response boundaries ($a$), at least when natural scenes were the target category. The consistent-object advantage in scene categorization seems to arise from differences in the quality of perceptual evidence, not differences in the time to perceptually processes the scenes or from a biased decision process.

# Dynamic scene categorization model

To test this idea further, we extended a successful model of scene categorization (Oliva & Torralba, 2001). This model is the perceptual front-end that extracts evidence for a scene's category that then drives the diffusion model of decision making. Specifically, the scene categorization model establishes the drift rate of the diffusion process, rather than allowing the drift rate to be a free parameter. This is analogous to earlier work that used object categorization models to establish drift rate of a diffusion-like process in order to account for speeded object categorization decisions (e.g. Nosofsky & Palmeri, 1997; Palmeri, 1997).

## Model description

We started with the scene categorization model developed by Oliva and Torralba (2001). In this model, scenes are represented by a set of features that describe the global spatial structure of the scene (Oliva & Torralba, 2001). The feature space, known as the spatial envelope, is defined by measures of global shape properties that are extracted using a bank of Gabor filters of varying spatial scales and orientations for a particular spatial resolution. Linear combinations of these spatial frequency components correlate with subjective characteristics of scenes such as mean depth, openness, expansion, and naturalness (Oliva & Torralba, 2001). Since the scene categorization task asked participants to categorize scenes as natural or man-made, we concentrated on spatial frequency features diagnostic to the naturalness of a scene.

We followed the procedure outlined by Oliva and Torralba (2001). A bank of Gabor filters spanning four spatial scales and eight orientations were used to extract the scenes' global features. To reduce the dimensionality of the filter responses, each filter output was downsampled to a lower-resolution ($4 \times 4$) summary. This lower-resolution summary preserves spatial localization of global scene information at a scale of 2 cycles/image. Principal component analysis (PCA) was then used to further reduce the dimensionality creating a final scene representation consisting of a 50-element vector. Natural versus man-made scene categories were defined by a hyper-plane boundary extracted using linear discriminant analysis (see Figure 4).

We used the results of the linear discriminant function to establish the drift rate of the diffusion model for each scene image to be categorized. Specifically, for a given scene, the output of the linear classifier corresponds to the distance of that scene from the boundary separating natural versus man-made scenes. The sign of the distance signifies which category the scene is classified into and the magnitude of the distance represents the quality of the evidence for that classification. Distance is transformed into drift rate with a sigmoid function,

$$v = 1/(1 + e^{-nx}), \quad (1)$$

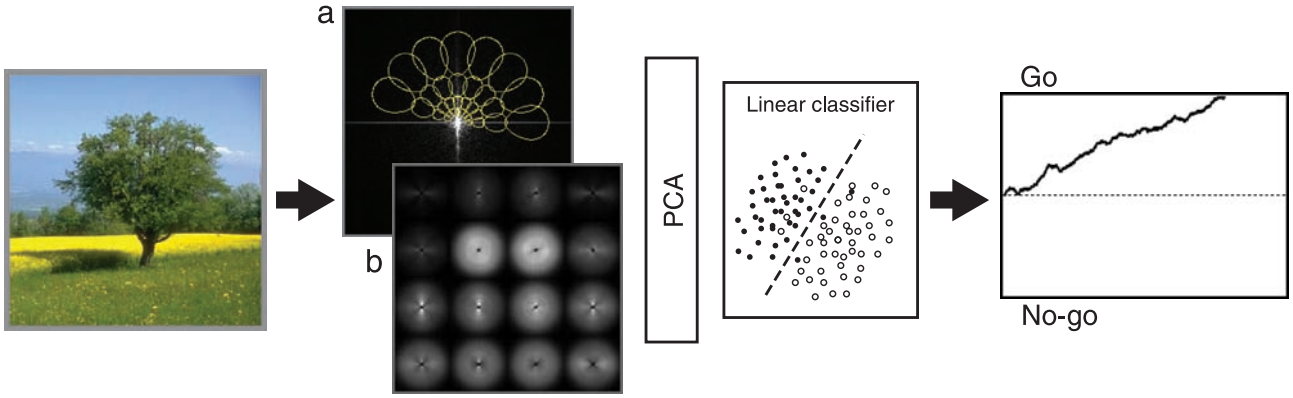


Figure 4. Illustration of the dynamic scene categorization model. Scenes are first classified by a scene categorization front-end. The scene's global spatial frequency is extracted with a bank of Gabor filters (*a*—polar plot of global spatial energy, spatial scale and orientation of filters shown by ellipses) and summarized into a low-resolution representation (*b*—subimages in the 4 × 4 grid show the global energy at that spatial location). Scene representations are projected onto a 50-dimensional principal component (PCA) space and classified by linear discriminant analysis (Linear Classifier). The resulting classification value drives a stochastic accumulation of evidence towards go or no-go response boundaries.

where $x$ is the quality of the evidence and $n$ is the drift rate scaling parameter. Using that drift rate, the decision process is carried out by the diffusion as a stochastic accumulation of evidence to a threshold. For the go/no-go model, the thresholds corresponded with a go response for the target category and a no-go nonresponse (Gomez et al., 2007). For the two-choice model, the thresholds corresponded with the natural and man-made scene categories (Ratcliff, 1978).

We must emphasize that this model assumes no parameters that vary across scenes containing consistent versus inconsistent objects. The scene categorization front-end uses the same discriminant function for scenes containing consistent and inconsistent objects. Distance from the discriminant function is transformed into drift rate using the same drift rate scaling function and the same scaling parameter for all scenes. The diffusion process determining the time-course of the decision is the same for all scenes, with the same boundaries. It should also be clear that the model contains no explicit object recognition process. Scenes are represented by global features that capture the scene's spatial frequency structure without extracting objects. The only difference between scenes containing consistent versus inconsistent objects is in the global content, not recognition of any individual objects in the scenes.

## Simulation method

First, a set of 200 natural and 200 man-made scene images were randomly selected from the scene database (same database as used in the behavioral experiments) for creating the PCA. A fifty-dimensional principal component space was extracted from these scenes' Gabor-filtered representations and saved. Next, a training set consisting of another 100 natural and 100 man-made scenes was randomly selected from the scene database. These scenes were passed through the Gabor filters, projected into the principal component space, and used to train the linear discriminant classifier.

The scene database we used had fewer inconsistent-object scenes compared to consistent-object scenes since, by definition, inconsistent objects are not typically found in those scenes. In order to ultimately test an equivalent number of scenes containing consistent and inconsistent objects, we randomly selected 500 consistent object scenes and inconsistent object scenes with replacement from the scene database. Scenes used for training were never included in the testing sets. Test trials consisted of first passing a scene through the scene categorization front-end. This stage generated a classification value from the discriminant function that was then transformed into a drift rate for the diffusion using Equation 1. The drift rate drove the stochastic accumulation of evidence until a decision threshold was reached or 1000 ms had elapsed. To model the go/no-go experiment, the decision thresholds corresponded with a go or no-go response as described in the diffusion model analysis above; the 2AFC experiment was modeled with decision thresholds corresponding to a natural or man-made scene. The three parameters of the model (drift rate scaling factor $n$, decision threshold $a$, nondecision processing time $T_{er}$) were optimized during training by fitting the predicted reaction time distributions to the observed data using the same procedure used in the earlier diffusion model analysis. We tested the model's performance with both natural and man-made scenes as targets. For generality, the entire simulation procedure was repeated with twenty-five separate training and testing sets for both the go/no-go and 2AFC models.

## Results

We analyzed the simulation predictions in the same way we analyzed the behavioral data. Simulated accuracy and reaction times for correct responses were analyzed separately by target category (natural and man-made). The go/no-go model predicted a significant consistent-object effect for natural scenes (Figure 5). Accuracy was higher ($Z = 4.24$, $p < 0.001$) and reaction times were faster ($Z = 4.37$, $p < 0.001$) for consistent-object scenes compared to inconsistent-object scenes. For man-made scenes as the target, mean differences in both accuracy and reaction time trended in the manner of a consistent-object advantage, but did not reach criterion for statistical significance ($Z = 0.977$, $p = 0.328$; $Z = 1.44$, $p = 0.15$); recall that the difference observed for human subjects was also quite small.

Similar results were found with the 2AFC simulations (Figure 5). For natural scenes as the target, accuracy was higher ($Z = 4.38$ $p < 0.001$) and reaction times were faster ($Z = 4.37$, $p < 0.001$) for consistent-object scenes compared to inconsistent-object scenes. For man-made scenes as the target, there was no significant difference in accuracy between consistent- and inconsistent-object scenes ($Z = 0.765$, $p = 0.509$) but reaction times were significantly faster ($Z = 4.36$, $p < 0.001$) for consistent-object scenes compared to inconsistent-object scenes.

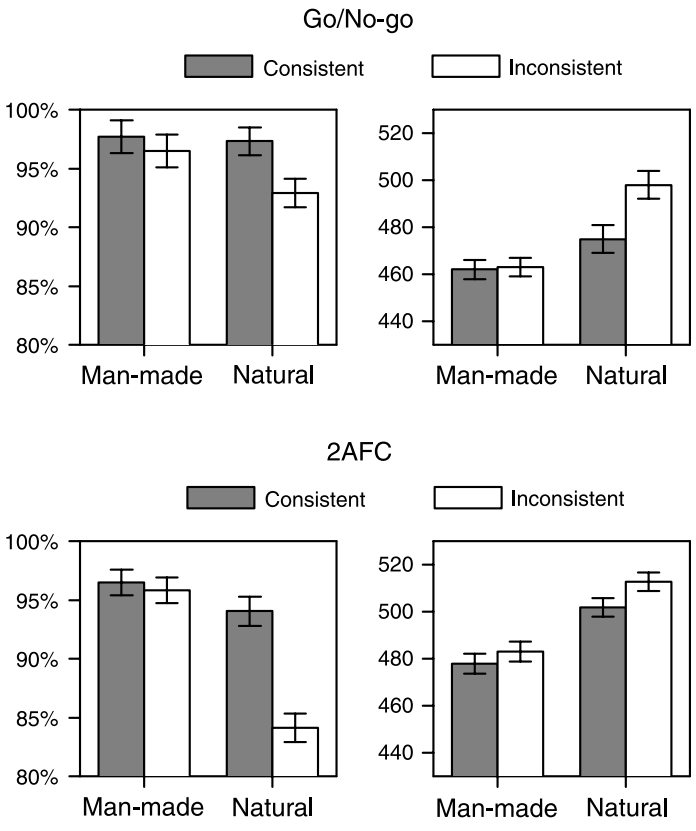


Figure 5. Simulation predictions for the go/no-go (top) and 2AFC (bottom) variants of the dynamic scene categorization model. Average accuracy (left) and correct response RTs (right) for consistent (dark) and inconsistent (light) object scenes. Error bars represented 95% confidence intervals of the comparison between consistent- and inconsistent-object conditions.

## Discussion

Simulations of ultrarapid natural scene categorization based on a go/no-go and 2AFC decision with the dynamic scene categorization model showed a significant consistent-object advantage for categorizing scenes. For the go/no-go model, an advantage was shown with natural scenes as targets and a small (but not significant) advantage for man-made scenes; for the two-choice model, an advantage was found with both natural and man-made scenes as targets. The larger consistent-object advantage with natural scenes compared to man-made scenes is qualitatively comparable to what was observed in human subjects in the behavioral experiments. These simulations suggest that the global features extracted by the perceptual front-end of the model were influenced by the presence of an inconsistent object. This subtle influence on global scene context may be sufficient to explain the lower accuracy and slower reactions times associated with scenes containing inconsistent objects.

# General discussion

The aim of our work was to test whether the consistent-object advantage in ultrarapid scene categorization observed by Joubert et al. (2007), and replicated and extended here, could be explained using global scene categorization mechanisms without the need to activate semantic information from object recognition. By this account, semantically inconsistent objects in scenes can influence the global perceptual evidence diagnostic for scene categorization without any explicit recognition of consistent versus inconsistent objects contained within the scene.

Consistent with this simple scene categorization account, we presented evidence from a diffusion model analysis that suggests a difference in the quality of the perceptual evidence available from scenes containing consistent versus inconsistent objects. Furthermore, we showed that a Dynamic Scene Categorization Model, which couples a scene categorization model based on global scene statistics (Oliva & Torralba, 2001) with the diffusion model of perceptual decision making (Ratcliff, 1978), accounts well for the consistent-object advantage in both go/no-go and 2AFC decisions. Instead of distinct scene and object recognition systems operating in parallel and competing or cooperating for categorization through activation of semantic information, the consistent-object advantage in

ultrarapid scene categorization of the sort reported by Joubert et al. (2007) can be explained by a single scene perception system that interprets the global visual properties found within scenes.

Our results are consistent with and extend the recent findings on rapid scene perception (Greene & Oliva, 2009a, 2009b). With just a glance at a scene, global properties, including their naturalness, are perceived rapidly even before a scene's basic-level category is determined or its objects are recognized (Greene & Oliva, 2009b). These global properties are captured by linear combinations of a scene's spatial frequency content (Greene & Oliva, 2009a; Oliva & Torralba, 2001). Here, we find that the rapid signal of a scene's superordinate category of natural or man-made as encoded by spatial frequency content is modulated by the presence of a consistent or inconsistent object; this modulation of visual information from objects can interfere with decisions about a scene's category. We also extend the earlier work by Oliva and colleagues by marrying their scene categorization model as a front-end to a diffusion model of perceptual decision making, resulting in a model that accounts well for both the accuracy and time to rapidly categorize scenes. With the growing interest in the temporal dynamics of object and scene categorization, it is necessary to have computational models that can explain the time-course of those decisions.

It is important to place our findings in their appropriate context. We are not arguing that explicit recognition of objects never matters for scene categorization. It goes without saying that fully understanding the environments we encounter during our everyday visual experience requires successful object recognition. There is evidence for semantic influences on scene categorization from object recognition (e.g., Davenport & Potter, 2004) that cannot be explained with the Dynamic Scene Categorization model. For these effects, some kind of two-system account of parallel object and scene perception interacting through activation of relevant semantic information remains a viable explanation. However, in the case of ultrarapid "ultrasuperordinate" (natural vs. man-made) scene categorization, we have shown that explicit representation and recognition of objects in those scenes is not necessary to account for an influence of consistent or inconsistent objects. It is striking—and not all that intuitive—that global scene statistics alone can reflect the presence of (semantically) consistent or inconsistent objects within those scenes and that quantitative differences in those global scene statistics can explain the observed quantitative differences in the consistency effect for natural vs. man-made scenes.

Clearly, our demonstration is one of sufficiency and not necessity. Further converging evidence is needed to understand whether the mechanisms described in our model underlie ultrarapid scene categorization in humans more generally. One possibility is that ultrarapid ultrasuperordinate scene categorization is governed by the processes instantiated in our model, but that more deliberate finer-grained scene categorization is modulated by interactions with explicit object recognition processes.

Alternative explanations for the original Joubert et al. (2007) results are also possible. Since we explicitly followed their experimental design, our article inherits those alternative explanations as well. Following Joubert et al., we used real images of scenes containing consistent objects or inconsistent objects. While there are clear benefits of using real images that preserve the range of low-level image properties found in real world visual experience, real images also engender a lack of control that introduce potential confounds. In one sense, our work demonstrates that these seemingly subtle visual confounds are (perhaps surprisingly) sufficient to produce significant differences in predicted behavior in ultrarapid scene categorization. In that sense, our work provides a cautionary tale, oft repeated, that relatively low-level visual differences can lead to what might erroneously be ascribed to relatively high-level semantic explanations.

One potential explanation for the difference between consistent and inconsistent object in the real scenes used by Joubert et al. (2007) and us is the factor of object salience. Maybe consistent and inconsistent objects in scenes simply differ in salience. Now, the term "salience" is used often in visual science but it can mean different things in different contexts by different investigators. Let's consider a couple possibilities. In Figure 1, the church in the bottom left scene may simply be more salient than the tree in the top left scene based solely on the low-level visual properties of the objects. Of course, in order to predict any kind of consistent-object advantage overall, there would need to be a similar salience difference for consistent and inconsistent objects across both natural and man-made scenes. For example, all consistent objects in man-made scenes would need to be more salient, on average, than inconsistent objects in man-made scenes, and all consistent objects in natural scenes would need to be more salient, on average, than inconsistent objects in natural scenes. Then the consistent-object advantage would come about because consistent objects are more salient, perhaps they are categorized quickly, and their categorization supports and speeds up the scene categorization. But maybe the salience difference goes the other way. Perhaps it is inconsistent objects that are always more salient in their respective scenes. Because the inconsistent objects are more salient, their categorizations are made quickly, and those categorizations interfere with the categorization of the rest of the scene. An explicit model that calculates image-based salience of objects in scenes, uses that information to select objects for categorization, and allows their categorization to interact positively or negatively with a scene categorization would need to be formalized in order to fully test this intriguing alternative explanation.

Another relevant definition of salience is of an object relative to its context. An inconsistent object appearing in

natural or man-made scenes is simply more salient than consistent objects in those same scenes (e.g., a car in a natural scene is simply more salient than a tree in a natural scene). A consistent-object advantage arises not because of the presence of an inconsistent object, *per se*, but because of the difference in salience that interferes with the recognition process. In some ways, this explanation can be recast as a restatement of our claim. Our model is based on the assumption that a scene's naturalness is represented by a linear combination of global image statistics. These global-based representations are influenced to some extent by the objects found in scenes suggesting that the stored representation for a natural scene will be biased towards the visual properties of objects typically found in natural scenes and likewise for man-made scenes with man-made objects. The presence of an inconsistent object in a scene shifts the scene's global representation away from the expected visual regularities captured in the target representations. In this sense, an inconsistent object in a scene ends up being more salient in that it shifts the overall scene representation away from how it might be represented if the scene did not contain that object.

In closing, evidence for interactions of object and scene processing has been well known for many years (Biederman et al., 1982; Friedman, 1979; Palmer, 1975), yet few formal computational accounts of these effects have been proposed. The Dynamic Scene Categorization model presents a preliminary step in understanding the mechanisms behind object and scene processing. This computational model extends a current class of successful scene categorization models to predict both response probabilities and reaction times. This model offers a richer description of scene categorization by accounting for the time course of the perceptual decision, much in the tradition of some object categorization models (e.g., Lamberts, 2000; Nosofsky & Palmeri, 1997; Palmeri, 1997). Further behavioral research and application of this model is necessary to better understand the underlying mechanisms of scene categorization and to characterize the relationship between scene and object perception.

# Appendix A

Table A1.

| | Model Variant | Diffusion parameter values | | |
|---|---|---|---|---|
| | | *a* | $T_{er}$ | *v* |
| Go/No-Go | fixed | .131 | .369 | .649 |
| Natural | *a* | .119, .132 | .371 | .620 |
| | $T_{er}$ | .133 | .364, .373 | .665 |
| | *v* | .152 | .360 | .911, .725 |
| Go/No-go | fixed | .340 | .257 | .839 |
| Man-made | *a* | .352, .359 | .271 | .901 |
| | $T_{er}$ | .411 | .290, .230 | .779 |
| | *v* | .431 | .198 | .873, .854 |
| 2AFC | fixed | .151 | .314 | .728 |
| Natural | *a* | .141, .156 | .317 | .742 |
| | $T_{er}$ | .176 | .320, .331 | .965 |
| | *v* | .179 | .314 | .991, .863 |
| 2AFC | fixed | .301 | .208 | .759 |
| Man-made | *a* | .297, .302 | .260 | .602 |
| | $T_{er}$ | .332 | .362, .370 | .589 |
| | *v* | .386 | .166 | .885, .860 |

Table A1. Best-fitting parameter values from the diffusion model analyses for the four model variants fit to the behavioral data from the two experiments (Go/No-Go and 2AFC) with natural and man-made scenes as targets. Model variants are identified in the third column by the parameter that was free to vary across the consistent versus inconsistent condition. The two values for free parameters associated with the consistent and inconsistent conditions are shown within the same cell separated by a comma.

# Acknowledgments

This work was supported by a grant from the James S. McDonnell Foundation, NSF grant HSD-DHBS05, and by the Temporal Dynamics of Learning Center (NSF Science of Learning Centers grant SBE-0542013).

Commercial relationships: none.
Corresponding author: Michael L. Mack.
Email: michael.mack@vanderbilt.edu.
Address: Vanderbilt University, Wilson Hall, 111 21st Avenue South, Nashville, TN 37240, USA.

# Footnote

[1]We adopted the definition of consistency between objects and scenes used in Joubert et al. (2007), which may appear at odds with earlier studies of object/scene consistency (e.g., Davenport & Potter, 2004; Palmer, 1975). Is a tree in a street scene actually inconsistent? Here, consistency between an object and scene is defined in reference to the relevant categories of the task being performed. The Joubert et al. (2007) experiments (and our replications and extension) asked participants to categorize scenes as natural versus man-made. In this sense, a tree (a natural object) is semantically inconsistent with a street scene's category (man-made).

## References


Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177. [PubMed]

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition, 35,* 393–401. [PubMed] [Article]

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15,* 559–564. [PubMed]

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review, 16,* 1026–1036. [PubMed]

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108,* 316–355. [PubMed]

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General, 136,* 389–413. [PubMed] [Article]

Grasman, R. P. P. P., Wagenmakers, E.-J., & van der Maas, H. L. J. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology, 53,* 55–68.

Greene, M., & Oliva, A. (2009a). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58,* 137–176. [PubMed] [Article]

Greene, M., & Oliva, A. (2009b). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20,* 464–472. [PubMed] [Article]

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243–271. [PubMed]

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47,* 3286–3297. [PubMed]

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review, 107,* 227–260. [PubMed]

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognition natural scene categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. II* (pp. 2169–2178). New York.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review, 16,* 798–817. [PubMed]

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300. [PubMed]

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34,* 72–107. [PubMed]

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175.

Palmer, S. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3,* 519–526.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 324–354. [PubMed]

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481. [PubMed] [Article]

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44,* 2301–2311. [PubMed]

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5,* 195–200.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14,* 1011–1026. [PubMed] [Article]

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavioral Research Methods, 40,* 61–72. [PubMed]

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14,* 3–22. [PubMed] [Article]