Behav Res
DOI 10.3758/s13428-016-0820-3

SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes

Sabine Öhlschläger[1] · Melissa Le-Hoa Võ[1]

Abstract Our visual environment is not random, but follows compositional rules according to *what* objects are usually found *where*. Despite the growing interest in how such *semantic* and *syntactic* rules – a scene grammar – enable effective attentional guidance and object perception, no common image database containing highly-controlled object-scene modifications has been publically available. Such a database is essential in minimizing the risk that low-level features drive high-level effects of interest, which is being discussed as possible source of controversial study results. To generate the first database of this kind – SCEGRAM – we took photographs of 62 real-world indoor scenes in six consistency conditions that contain semantic and syntactic (both mild and extreme) violations as well as their combinations. Importantly, always two scenes were paired, so that an object was semantically consistent in one scene (e.g., ketchup in kitchen) and inconsistent in the other (e.g., ketchup in bathroom). Low-level salience did not differ between object-scene conditions and was generally moderate. Additionally, SCEGRAM contains consistency ratings for every object-scene condition, as well as object-absent scenes and object-only images. Finally, a cross-validation using eye-movements replicated previous results of longer dwell times for both semantic and syntactic inconsistencies compared to consistent controls. In sum, the SCEGRAM image database is the first to contain well-controlled semantic and syntactic object-scene inconsistencies that can be used in a broad range of cognitive paradigms (e.g., verbal and pictorial priming, change detection, object identification, etc.) including paradigms addressing developmental aspects of scene grammar. SCEGRAM can be retrieved for research purposes from http://www.scenegrammarlab.com/research/scegram-database/.

Keywords Scene grammar · Scene perception · Image database · Semantic and syntactic inconsistency · Eye-movements

Introduction

From the basic shapes used in the classic perception experiments of the Gestalt school, vision science has evolved to use more and more complex stimuli including depictions of real-world scenes. While more simple stimuli (e.g., oriented lines or gabor patches) remain essential to gain insights into the fine-grained mechanisms of human visual perception, more complex cognitive processes – like scene understanding – require the use of more realistic stimuli. This increase in complexity for the sake of realism goes hand in hand with a decrease in controllability of low-level image features and other confounding factors (e.g., object familiarity). In order to study real-world scene perception and understanding, a highly-controlled image database is essential.

Scenes can be considered a composition of objects that follow organizational principles. Biederman, Mezzanotte, and Rabinowitz (1982) provided a first terminology that included a semantic-syntactic distinction which was based on whether the properties of the relations require retrieval of the

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0820-3) contains supplementary material, which is available to authorized users.

✉ Sabine Öhlschläger
oehlschlaeger@psych.uni-frankfurt.de;
http://www.SceneGrammarLab.com

[1] Scene Grammar Lab, Department of Cognitive Psychology, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, 60323 Frankfurt am Main, Germany

meaning of the object. According to this definition object-scene relations including probability, size, and position concern *semantics*. That is, in order to know that a fire hydrant does not belong in the kitchen or on a mailbox, you need to know what it is. *Syntax* – in Biederman's terminology – comprises violations of support and interposition, as both relations violate physical constraints of gravity independent of the meaning of the single object, for instance a hydrant hovering above the street.

Võ and Wolfe (2013) introduced a refined, but slightly different taxonomy of semantic and syntactic object-scene violations: The authors refer to violations of *semantics* for an object that does not fit into the *global meaning* of the scene (e.g., milk in the bathroom). In contrast to Biederman and colleagues (1982), they refer to violations of *syntax* for objects that are not at their *typical location* within the scene. The term syntax is further divided into mild syntax violations (e.g., toothbrush in the bathroom sink) versus physically impossible syntax violations (e.g., toothbrush hovering above the bathroom sink). Indeed, using event-related potentials (ERPs), Võ and Wolfe (2013) provided evidence for independent neural representations of the types of scene semantics and syntax in line with their taxonomy, as well as differential processing of extreme and mild violations of syntax. Exposing observers to such object-scene violations helps studying the underlying knowledge system – our scene grammar – and elucidating how it interacts with ongoing behavior.

In the last decades, the effects of scene grammar on eye-movement control have been studied using different types of stimulus material ranging from line drawings (e.g., De Graef, Christiaens, & D'Ydewalle, 1990; Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1978) to 3D-rendered images (Võ & Henderson, 2009, 2011) and color photographs of real-world scenes (e.g., Becker, Pashler, & Lubin, 2007; Bonitz & Gordon, 2008; Underwood & Foulsham, 2006; Underwood, Templeman, Lamming, & Foulsham, 2008). For instance, Henderson et al. (1999) used line drawings extracted from scenes (adapted from De Graef et al., 1990) containing an object that was either semantically consistent (e.g., cocktail glass in the bar) or inconsistent (e.g., microscope in the bar). They reported longer dwell times (i.e., the sum of all fixation durations) on the inconsistent object, once it was fixated.

Meanwhile further research accumulated to form consensus that semantic (Bonitz & Gordon, 2008; Rayner, Castelhano, & Yang, 2009), and also syntactic inconsistencies (Spotorno, Malcolm, & Tatler, 2015; Võ & Henderson, 2009) have an influence on processing stages upon the first fixation of the object as shown in longer dwell times. However, whether these high-level inconsistencies also exhibit eye-movement control at earlier stages of processing prior to fixation of the object is still under debate. While some studies have found evidence for attentional attraction towards inconsistent objects prior to their fixation (Becker et al., 2007; Bonitz & Gordon, 2008; Loftus & Mackworth, 1978; Spotorno et al., 2015; Underwood & Foulsham, 2006; Underwood, Humphreys, & Cross, 2007), other studies did not find any evidence for such an attraction of eye movements (De Graef et al., 1990; Henderson et al., 1999; Võ & Henderson, 2009, 2011). A possible reason for these controversial results might lie in the fact that the stimuli used so far have varied widely across studies (for a discussion see Võ & Henderson, 2009). The level of clutter in the scene has been proposed as one possible factor that could explain some of the variance between study results. For instance, the scenes in the original "octopus in farmyard" study by Loftus and Mackworth (1978), which reported inconsistency effects on initial eye movements, only contained a few objects between big unoccupied areas of white background that might have favored a "pop-out" of the semantic inconsistencies.

Along the same lines, Võ and Henderson (2009) discuss that the average low-level saliency rank of the critical objects in the scene – according to Itti and Koch's (2000) saliency model – varied across studies between the third (Underwood, Humphreys, & Cross, 2007) and the ninth (Võ & Henderson, 2009) most conspicuous area of the image (see also Spotorno et al., 2015, but only medians indicated). Manipulating both salience and semantic consistency of critical objects revealed that both components might not function independently (Spotorno, Tatler, & Faure, 2013; Underwood & Foulsham, 2006; but see Underwood, Templeman, Lamming, & Foulsham, 2008). For instance, in a scene viewing task similar to the one used by Henderson et al. (1999), inconsistent objects were only fixated earlier and with fewer fixations compared to consistent objects when they were inconspicuous, whereas when they were conspicuous both inconsistent and consistent objects were fixated early (Underwood & Foulsham, 2006). These results indicate that when studying high-level scene-object inconsistencies high-saliency ranks might be less appropriate.

Furthermore, in some studies the critical objects were inserted post-hoc into color photographs which might have resulted in artificial shadows, edges, or depth cues that might have affected or even attracted the gaze of observers (Becker et al., 2007; Spotorno et al., 2015; Underwood et al., 2007; for discussions see Spotorno et al., 2015; Võ & Henderson, 2009). As another consequence, post-hoc editing might affect the ambiguity of the intended inconsistency manipulation: For instance, when objects do not appear true to scale with their context, an intended violation of only position might be confounded with the violation of size. Such a double-inconsistency might have contributed to some of the previously reported eye-movement effects rendering their interpretation more difficult (Spotorno et al., 2015).

This debate is supposed to illustrate the importance of well-controlled stimulus material to ensure the reliability and replicability of eye-movement findings in general (e.g., studies on the relations between dwell times and age, personality traits, and cultural background) and emphasizes which image parameters are the most relevant to control for. Hence, a database of images, well-controlled for visual saliency and image editing artifacts, which depicts a range of well-defined and counterbalanced inconsistency manipulations that can be used in a variety of paradigms might shed new light on this or other debates and could provide the grounds for more converging results.

There are numerous databases containing hundreds or even thousands of images – such as the Caltech-256 object category dataset (Griffin, Holub, & Perona, 2007), the ImageNet Database (Deng et al., 2009), and for scenes specifically the SUN Database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010) and the MIT-CSAIL Database of Objects and Scenes (Torralba, Murphy, & Freeman, 2004), relying on photographs retrieved from various internet sources. On the one hand this provides a nearly inexhaustible number of images, as it is needed for certain research purposes. On the other hand, many important image features and the objects themselves cannot be as carefully controlled for or manipulated as in photographs that are purposefully generated. There are also databases of photos purposefully generated such as the Columbia Object Image Library (COIL-100; Nene, Nayar, & Murase, 1996), the McGill Calibrated Color Image Database (Olmos & Kingdom, 2004), and the Berlin Object in Scene Database (BOiS; Mohr et al., 2016), but only the latter contains one type of object-scene manipulation, i.e. the mild-syntactic one.

Some other studies (additionally) also included semantic manipulations and took great care to control for effects related to the salience and familiarity of the critical objects when designing their own stimulus set for their studies (e.g., Henderson et al., 1999; Võ & Henderson, 2009). The line drawings of 24 real-world scenes used by Henderson et al. (1999) were paired such that a critical object occurred as consistent in one scene, but inconsistent in the other scene. Similarly, Võ and Henderson (2009) paired 20 3D-rendered images of real-world scenes to counterbalance the assignment of objects to conditions. In addition, they created for each scene a condition of physically impossible syntax violations (i.e. the critical object hovering in midair), in addition to semantic inconsistencies making both directly comparable. However, the latter study only investigated extreme syntax violations, not mild ones. Furthermore, while 3D-rendering is becoming more and more sophisticated and allows for high control of image features, the 3D-rendered scenes used in experiments usually do not reach the same level of realism that can be achieved by digital photography. More importantly, due to the fact that 3D-rendered scenes are generated from scratch by the experimenter, these rarely contain the amount of clutter that is natural in our environment and realistically captured in photographs.

Võ and Wolfe (2013) generated a stimulus set of photographs of real-word scenes that represented all possible conditions of scene-object inconsistency. However, a given scene was not photographed in all syntax conditions, i.e. scenes either had a mild or an extreme syntax violation, but not both. Furthermore, the scenes were not paired to avoid effects of mere object familiarity. This renders such stimuli less appropriate for developmental studies with children, where varying degrees of familiarity with different objects per se might influence children's looking times as has been suggested by studies using the habituation paradigm (for a review see Turk-Browne, Scholl, & Chun, 2008).

In order to address the shortcomings presented above, we generated an image database with SCEne GRAMmar manipulations (SCEGRAM) that consists of highly-controlled photographs of objects in real-world scenes. Each of the scenes was photographed in all six possible semantic and syntactic conditions: (1) a consistent control, (2) an inconsistent-semantic condition, (3) a mild inconsistent-syntax condition, (4) a mild double-inconsistency condition, (5) an extreme inconsistent-syntax condition, and (6) an extreme double-inconsistency condition. Two scenes were always paired so that each object occurred once in a consistent and once in an inconsistent scene context. Furthermore, each scene was also photographed without the critical object and finally each object was photographed individually against a white background. In addition, we obtained consistency ratings for each scene-object version, included a visual salience analysis of the entire database, and performed a cross-validation of two condition subsets.

Methods

The SCEGRAM database can be retrieved for research purposes from our website at: http://www.scenegrammarlab.com/research/scegram-database/. In the following, we describe the creation and composition of the image database in more detail.

Apparatus

Digital photographs were taken with a Nikon D5100 single-lens reflex (SLR) digital camera with an 18–55 mm zoom lens. The original image aspect ratio was 3:2 with a resolution of 4,928 × 3,264 pixels and images were stored in JPG format (JPG fine).

Procedure

The photos were taken in apartments of the private living spaces of colleagues, friends, and family in Frankfurt, Germany, and rural surroundings. Because of the varying and uncontrollable lighting conditions in different rooms and apartments, we took the photos according to the following scheme: First, we determined the recommended ISO and shutter speed using the auto (flash off) mode, then we switched to manual mode and used those parameters. All photographs were taken without using the flash under natural lighting conditions, but the room light was switched on if needed. Focus and zoom were adapted manually for each scene, but were kept constant across all conditions within the same scene. A tripod was used to ensure identical positioning of the camera and therefore identical placement of all objects within a photograph for all conditions of a given scene.

We did not insert features post hoc, that is, hovering objects were photographed actually hovering in mid air using a stick and transparent cord. Thus post hoc editing via, for example, Adobe Photoshop (see following paragraph) of the images was kept to a minimum to avoid artificially introduced artifacts. Where possible, floating objects were photographed to remain upright in the air to underline their impossible physical state rather than, for example, giving the impression that objects were photographed while being thrown.

Image post-processing

In the first post-processing step, scenes were modified in Adobe Photoshop CS (Adobe, USA) only in the following cases:

(a) If the transparent cord was visible in the extreme syntax-violation scenes, the corresponding pixels were adapted to the surrounding pixels using the Photoshop retouching and healing tools.

(b) If an identically looking object occurred in another scene as a distractor object (only in scenes 28, 42, and 51), the color of the distractor object was changed into a realistic color for that specific object using the Photoshop replace color adjustment.

Then, the photographs were cropped in width so that they matched the standard aspect ratio 4:3 and resized to a resolution of 1,024 × 768 pixels in MATLAB (The MathWorks Inc., USA) using the Image Processing Toolbox. The cropped and resized images were then exported to PNG format and saved in addition to the original image files. All following steps in the methods section refer to the cropped images.

Conditions

Object present images (see Fig. 1): A given scene was photographed with

(a) a semantically consistent object in a consistent location (consistent control condition; CON),

(b) a semantically inconsistent object in a syntactically consistent location (inconsistent-semantics condition; SEM),

(c) a semantically consistent object in a syntactically inconsistent, but physically possible location (mild inconsistent-syntax condition; SYN),

(d) a semantically inconsistent object in a syntactically inconsistent, but physically possible location (mild double-inconsistency condition; SEMSYN),

(e) a semantically consistent object in a syntactically inconsistent, but physically impossible location, that is hovering in midair (extreme inconsistent-syntax condition; EXSYN),

(f) a semantically inconsistent object in a syntactically inconsistent, but physically impossible location, that is hovering in midair (extreme double-inconsistency condition; EXSEMSYN).

In order to create these conditions, two objects were always paired according to their visual attributes (i.e., shape, color, and size) as illustrated in Fig. 2 and both were photographed once in a consistent and once in an inconsistent scene at consistent and inconsistent locations. As can be seen in Fig. 1, the toilet paper in the bathroom fits into the context and is at its most probable location, the toilet paper holder (CON). When instead of the toilet paper, a cup appears at this location, it is inconsistent with the global semantics of the scene (SEM). When occurring on the toilet seat cover, the toilet paper is at a physically possible yet inconsistent location, but still fits into the meaning of the scene (SYN), whereas for the cup both the context and the location are inconsistent (SEMSYN). When hovering in midair above the toilet, both toilet paper and cup are at physically impossible inconsistent locations, but only for the cup the meaning of the scene is inconsistent (EXSYN vs. EXSEMSYN). It should be noted that an object that does not fit into the meaning of a scene obviously has no appropriate position in that scene. Therefore, the object locations of the CON, SYN, and EXSYN conditions were maintained to control for location in the double-inconsistency conditions.

Object absent images:

We additionally photographed each scene without the critical object (absent condition; ABS) immediately after creating the present scene. This was done to minimize changes – e.g. due to changes of lighting conditions –

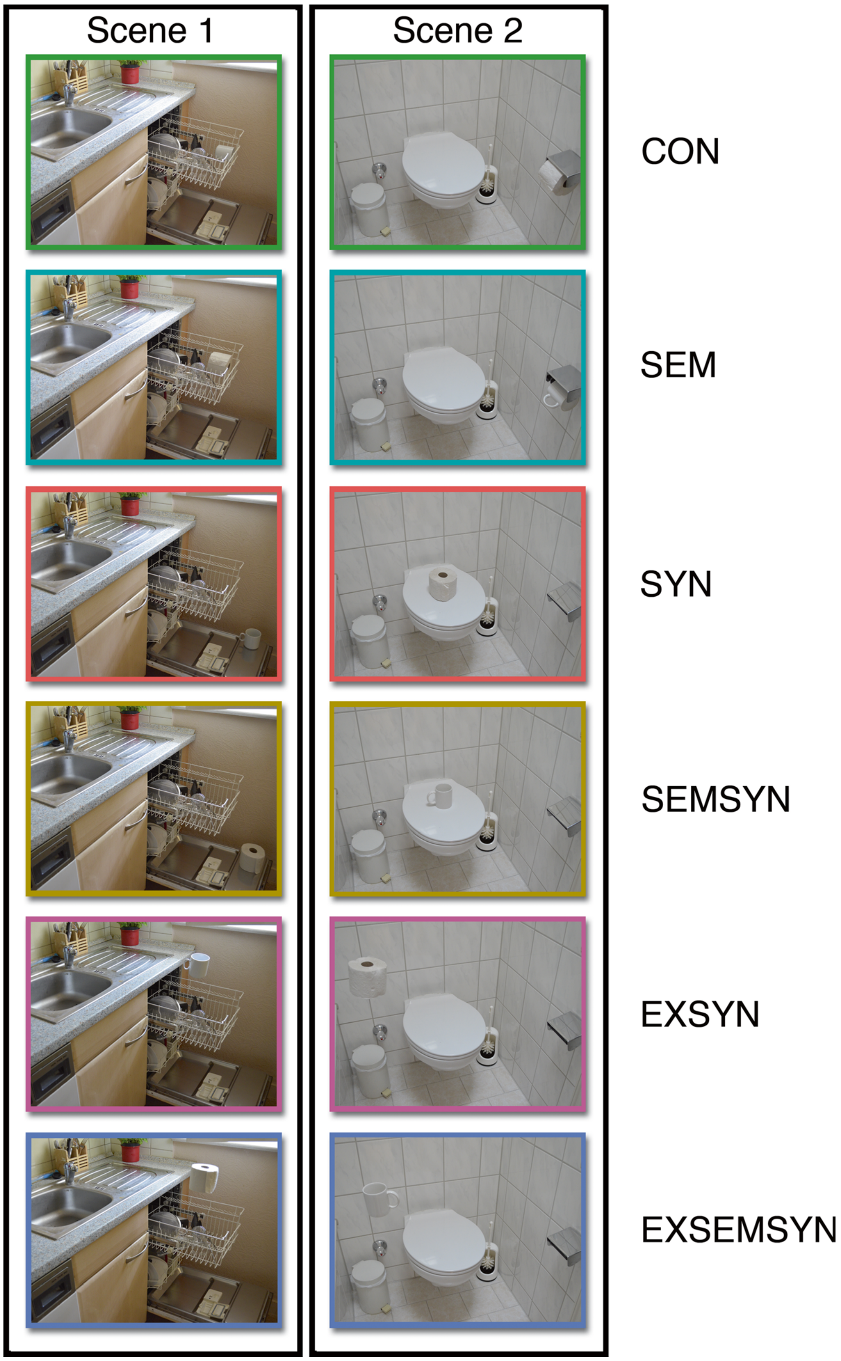

Fig. 1 Consistency conditions and counterbalancing. The consistency conditions are illustrated for the two example scenes 1 and 2. The occurrence of objects in scenes was counterbalanced in a way that each object was photographed in a consistent and inconsistent context, respectively (e.g., cup in kitchen vs. cup in bathroom)

when transitioning between object present and absent versions of each consistency condition, allowing for smooth sequential presentation of both photos with only the critical object changing.

Object-only images (see Fig. 2): Furthermore, each object was photographed in isolation in front of the identical white background. This gives the opportunity to use the object image to prime the target or the scene or to test recognition performance for the objects (object only condition; OBJ). The object-only portrait images were post-processed by cropping the original white background as well as centering, and resizing the object to fit into a central box of 2,144 × 2,144 pixel on a uniform white background using Adobe Photoshop CS (Adobe, USA). In a second step, object-only images were cropped and resized in the same manner as the scenes. The resulting object-only photos are depicted in Fig. 2.

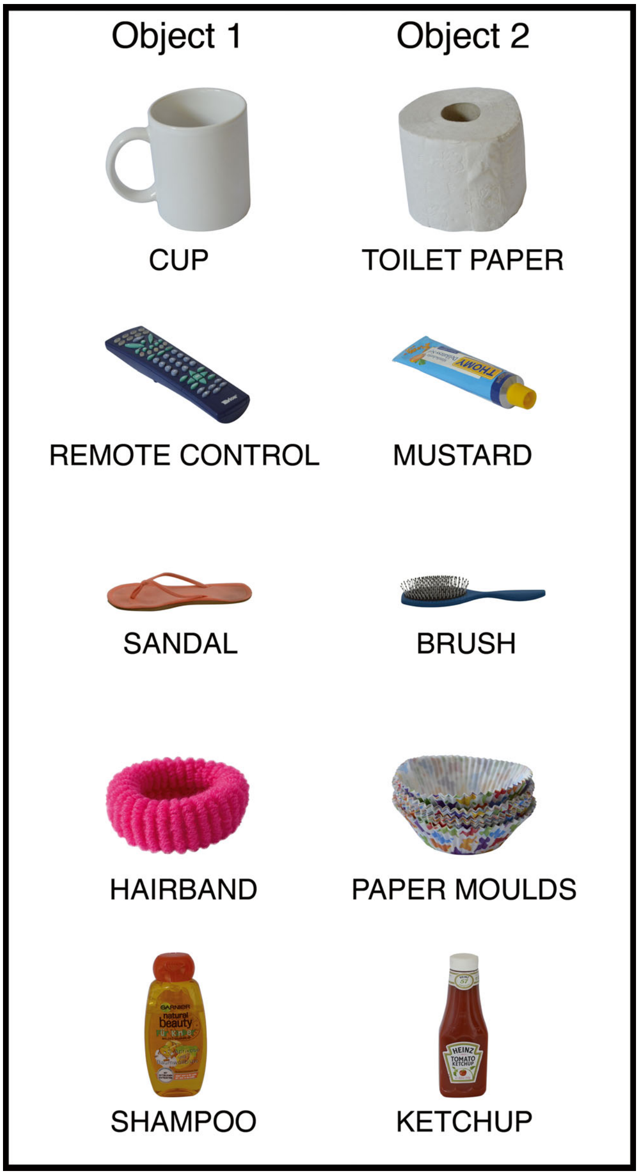

Fig. 2 Examples of the object-only portrait images included in the SCEGRAM database. Two objects (e.g., cup and toilet paper) were always paired according to their visual attributes (i.e., shape, color, and size) and served as consistent and inconsistent object, respectively

Counterbalancing

We took great care to eliminate as many confounding factors as possible (e.g., object familiarity, low-level saliency of image features) that might co-vary with the inconsistency conditions. To do so, we tried to match the semantically inconsistent and consistent object of each scene in their appearance (see Fig. 2). As a major feature of this database, the same object occurs in two paired scenes (with even and odd numbers): once in a scene where it was semantically consistent and once in a different scene where it was not. For instance, in Scene 1 (see Fig. 1), the cup appears in the kitchen as consistent condition (CON), whereas in Scene 2 the same cup appears in the bathroom, where it is semantically inconsistent with the scene context (SEM). As objects were photographed at both consistent and inconsistent locations in both the consistent and inconsistent scenes, each object occurs in each of the six consistency conditions. For instance, in Scene 1 (see Fig. 1), the cup appears on the door of the dishwasher (SYN), whereas in Scene 2 the same cup appears on the toilet seat (SEMSYN). This principle also applies for the extreme-syntax condition.

Areas of interest (AOIs)

Areas of interest (AOIs) were drawn manually around the critical object in each scene so that they form the smallest possible rectangle that encloses the object. In case of occlusion of object parts, only the visible object contours were considered for defining the AOIs, however, this does not mean that the remaining part of the object is not entirely included as AOIs are always rectangular. The AOIs were obtained in image coordinates and described in the format according to their horizontal and vertical center, as well as their width and the height. The AOI information can be found in columns 13 to 16 of the database excel sheet (see following section). Note that for certain purposes (e.g., eye tracking, saliency checks) the size of the AOIs across conditions should be equalized within the same scene by any user of this database. This is to prevent differences in dependent measures only due to differently sized AOIs. This equalization procedure is described in detail in the sections for the saliency check and the eye-tracking cross-validation.

Low-level image features: Saliency check

The counterbalancing of critical objects between conditions reduces the influence of confounding factors. To further control for saliency differences that might co-vary with the consistency manipulation, we calculated the saliency rank within 15 simulated fixations using the Saliency Toolbox (Walther & Koch, 2006). The saliency rank reflects the serial rank of a model fixation calculated based on brightness, color, contrast, and edge orientation. It has a value of 1 for the most salient area of an image. For the saliency analysis we used the AOIs defined in the previous step and equalized their size in a given scene in all six scene-object conditions. For instance in Scene 1 (see Fig. 1) the AOIs of cup and toilet paper were adjusted to be equally large in all six images, even if the size of the objects themselves differed. The AOIs covered the identical area of the scene for objects that were supposed to be at the same location (i.e., CON and SEM, SYN and SEMSYN, EXSYN and EXSEMSYN) by defining the smallest rectangle that encloses the critical objects. When the first model fixation fell within the critical object AOI, this AOI received the respective saliency rank. In case none of the simulated fixations fell into a given AOI, its fixation rank was set to 15. The mean saliency rank was compared across all conditions of a given scene using a one-way repeated-measures ANOVA with consistency condition as factor with six levels. The mean saliency rank for the critical objects did not differ significantly between consistency conditions, $F < 1$. As this analysis is conducted in favor of the null hypothesis, we furthermore quantified the evidence for the null compared to the alternative hypothesis by calculating the Bayes Factor (BF_{01}) using the BayesFactor package (Morey, Rouder, & Jamil, 2013) implemented in R. The resulting BF_{01} of 82.86 indicates a "very strong evidence for the null hypothesis" according to Wagenmakers, Wetzels, Borsboom, and van der Maas (2011; adapted from Jeffreys, 1961), thus minimizing the influence of the low-level saliency of image features on response differences between the consistency conditions. As can be seen in Fig. 3, the mean saliency rank was larger than 4.5 in all conditions. In other words, the AOI was on average not fixated before the 5th simulated fixation, which is lower than the high-saliency definition of Underwood and colleagues (2008) and Underwood and Foulsham (2006). Note that the saliency calculations resulting from this procedure are only valid when considering the full database in all conditions and when adjusting the AOIs in the way previously described. Therefore, this step has to be adapted by every user, who uses the database in a customized manner.

SCEGRAM database

The database consists of 62 scenes in all six consistency conditions (372 individual images) and the ABS condition (372 individual images), resulting in 744 scene photos in total. In addition, it includes 62 object photos (OBJ). The excel file containing the database information can be downloaded from http://www.scenegrammarlab.com/research/scegram-database/.

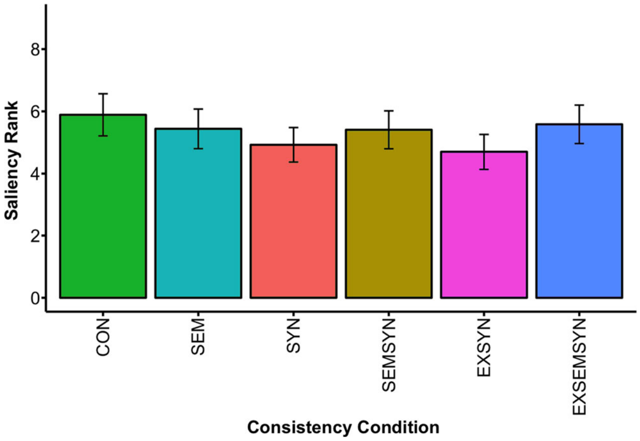

Fig. 3 Mean saliency rank per consistency condition. Error bars indicate ±1 standard error

Within this excel file, the spreadsheet "scenes" contains all relevant information for using the scenes. It is described in detail as follows:

Column #1: name of the scene image file indicating the number of the scene and the consistency and absent condition.
Column #2: scene ID (62 scenes in total)
Column #3: image ID, irrespective of scene and condition (744 images in total)
Column #4: presence of critical object (*1 = present, 0 = absent*).
Column #5-6: name of the critical object in English and German respectively and 'XXX' in case the object was absent
Column #7-8: scene category, in English and German, respectively
Column #9-10: scene category, in which the critical object was consistent, in English and German, respectively
Column #11-12: consistency condition, abbreviated and number coded, respectively
Column #13-16: location of critical object as coordinates of the AOI center on the horizontal and vertical axes, and the width and height of the AOI, containing '-99', in case the critical object was absent
Column #17-18: scene image width and height in pixels
Column #19: average consistency rating *(1=consistent, 6 = inconsistent)* for the scene-object condition rated by 12 participants (see following section)

The spreadsheet "objects" contains all relevant information for using the individual object images and is described in detail as follows:

Column #1: name of the image file
Column #2-3: object names in English and German, respectively
Column #4: object ID (62 objects in total)
Column #5-8: scene category, in which the object is consistent and the one in which it is inconsistent in English and German, respectively
Column #9-10: scene ID, in which the object was consistent and inconsistent

Database validation

Consistency rating

As a manipulation check we obtained consistency ratings of the critical object in a given scene across all six consistency conditions (see Column #19). Since people tend to differ in the way they interpret how consistent or inconsistent an object is within a scene (similar to valence ratings on emotional images), we encourage every user to always collect their own ratings of the stimulus set they use. In the following, we provide one of many possible examples of how to do this.

Apparatus

Participants were seated at an approximate viewing distance of 65 cm. The scenes were presented on a 24-in. monitor (resolution: 1,920 × 1,080 pixels, refresh rate: 60 Hz) and subtended a visual angle of 23.54° horizontally and 18.06° vertically. Stimulus presentation and response recording were controlled with Matlab (The MathWorks Inc., USA) using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

Participants

Twelve undergraduate students (mean age = 20.83, SD = 2.25, range: 18–25 years, six female) participated in the consistency rating for course credit. All participants had normal or corrected-to-normal vision and color vision.

Procedure

The participants were presented with all 372 object present scenes (62*6 conditions) successively and were asked to rate on a discrete scale from 1 to 6 how well the critical object fits into a given context or the location of the context (Supplementary Materials #1 for the original instructions). For each scene, the critical object was indicated by a red frame that appeared on the scene 500 ms after scene onset and remained visible for 2 s. Then the original scene was presented without the frame for another 500 ms before the grey response display was shown. The participants rated each scene and gave confidence ratings in each condition in a within-subject design. The presentation of the 372 scenes was divided into six blocks of 62 trials with each block containing exactly one condition-scene version of the scene. The occurrence of the condition-scene versions in the image sequence was counterbalanced across groups of six participants according to a latin-square design.

Results

All types of scene-object violations were rated more inconsistent compared to the consistent control condition, $F(5, 55) = 87.56, p < .001, \eta^2G = 0.80$, all *t*s > = 5.79, indicating that our object-scene manipulations had their intended effects (see Fig. 4). Semantic violations were judged higher in inconsistency compared to both mild, $t(11) = 10.68$, p < .001, and extreme-syntactic violations, $t(11) = 4.16, p < .05$, which did not differ significantly from each other after controlling for multiple comparisons using Bonferroni adjustments, $t(11) = 2.49, p = .451$. These observations diverge from those reported by Võ and Wolfe (2013) and are partly attributed to the inter-subject variability in the extreme-syntax condition as will be discussed in the following section. The rating data of individual participants is displayed in the supplementary materials (Fig. 7; Supplementary Materials #2).

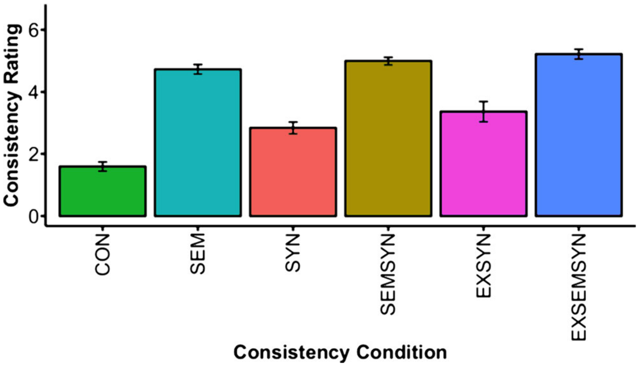

Fig. 4 Mean consistency ratings judged from 12 participants on a scale from 1 to 6 (*1=consistent, 6=inconsistent*) as a function of consistency condition. Error bars indicate ±1 standard error

The participants' rating confidence also differed between consistency conditions as shown in Fig. 5, $F(5, 55) = 4.59, p < .05, \eta^2G = 0.21$. Post-hoc comparisons controlling for multiple comparisons using Bonferroni adjustments revealed that this effect was driven by the participants being less confident in their ratings of mild-syntax violations compared to the consistent object-scene pairings, $t(11) = 5.36, p < .01$. Again subjects especially differed in their ratings for extreme-syntax violations. The confidence ratings of individual participants are displayed in the supplementary materials (Fig. 8; Supplementary Materials #2).

Cross-validation using eye-movements

The aim of the cross-validation was to replicate the well documented eye-movement findings of longer dwell times to semantic and syntactic violations (Henderson et al., 1999; Võ &

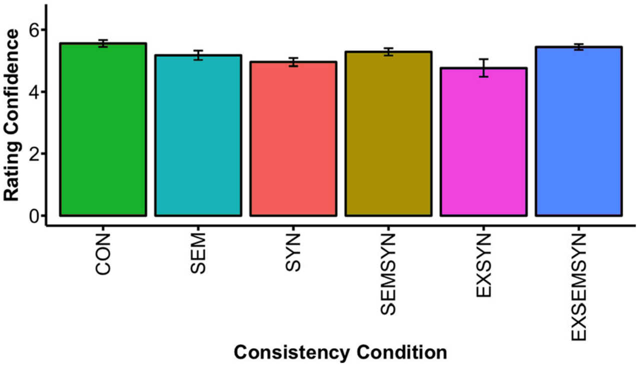

Fig. 5 Mean rating confidence judged from 12 participants on a scale from 1 to 6 (*1=consistent, 6=inconsistent*) as a function of consistency condition. Error bars indicate ±1 standard error

Henderson, 2009). We wanted to accumulate further evidence that the SCEGRAM stimuli have their intended effects on eye-movement control, while we leave it to the users of this database to address theoretical questions and debates about inconsistent findings in the field. For this eye-tracking paradigm we focused on the semantic and both the mild- and extreme-syntactic inconsistencies. Even with the lower than expected ratings for the extreme-syntax violations we anticipated to observe a strong consistency effect to these violations in the more sensitive eye-movement measure and wanted to extend this finding to mild-syntax violations for the first time.

Apparatus

Participants were seated at an approximate viewing distance of 80 cm. The scenes were presented on a 24-in. monitor (resolution: 1,920 × 1,080 pixels, refresh rate: 60 Hz) and subtended a visual angle of 19.49° horizontally and 14.84° vertically. Stimulus presentation and response recording were controlled with Matlab (The MathWorks Inc., USA) using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). For the cross-validation experiment, eye movements were recorded monocularly with an Eyelink 1000 Plus desktop mount eye tracker (SR Research, Canada) at a sampling rate of 500 Hz in remote mode.

Participants

Three different groups of participants non-overlapping with the sample from the rating experiment, were either tested on the inconsistent-semantics condition (n = 15, mean age = 20.98, SD = 2.71, range: 18–27 years, 12 female) or the mild inconsistent-syntax condition (n = 7, mean age = 21.13, SD = 1.99, range: 19–25 years, seven female) and extreme inconsistent-syntax condition (n = 7, mean age = 25.66, SD = 5.50, range: 19–33 years, five female) in comparison to the consistent control condition. All participants had normal or corrected-to-normal vision and color vision.

Stimuli

For the cross-validation experiment, a subset of 40 scenes was used in the inconsistent-semantics and the inconsistent-syntax experiments, which where overlapping between semantics and syntax conditions apart from six scenes. To apply the strength of SCEGRAM to compare the identical scenes between conditions, we did a sub-analysis of the eye-movement data for the 34 scenes that were identical in the three conditions with a consequence that our counterbalancing was not maintained on analysis level. However, results were comparable to those obtained with the full design, so that in the following only the results of the sub-analysis are reported. The original AOIs were equalized to have the same size for the inconsistent-syntax and control condition, and the same size and location for the inconsistent-semantics and control condition, respectively. Online, a buffer of 75 pixels was added to each side of the AOI to counteract possible tracking imprecision. The mean saliency rank of the 34 scenes did not differ between the consistent and the semantic (CON: M = 5.74, SD = 5.14, SEM: M = 6.47, SD = 5.58, *t*(33) = 0.60, p = 0.55), the mild-syntactic condition (CON: M = 6.06, SD = 5.36, SYN: M = 7.32, SD = 5.71, *t*(33) = 1.3, p =.20), or the extreme-syntactic condition (CON: M = 5.18, SD = 4.93, EXSYN: M = 4.65, SD = 3.98, *t*(33) = 0.77, p = .44).

Procedure

Before the experiment, a 5-point-calibration and validation were performed. Drift checks were inserted every ten trials, and a recalibration was administered if necessary during the experiment. Following two practice trials, the 40 experimental trials started. The trial sequence was as follows: A blank screen was presented for 500 ms, followed by an animated fixation spiral randomly presented on the left or right side in half of the trials. As soon as the participants were looking at the fixation spiral for 500 ms, the scene was presented for 7 s upon the first gaze sample detected on the scene. After half of the scenes, a 10 s reward video was presented. The participant's task was to simply view the scene presented. No manual response was required. Fifteen participants viewed 20 scenes with inconsistent semantics and 20 consistent control scenes, whereas two different groups of 20 participants viewed 20 scenes with mild or extreme inconsistent syntax and 20 consistent control scenes. After terminating the experiment, the participants rated the scene-object consistency for the subset of scenes presented during the experiment – a procedure recommended for each user of this database (see Discussion section).

Results

In order to directly compare our results to previous findings on the effects of semantic and syntactic inconsistencies on eye movement control (Henderson et al., 1999; Võ & Henderson, 2009), we calculated the dwell time. The dwell time is defined as the sum of all durations of fixations that fell on the critical object starting with the first gaze sample on the scene until the offset of the scene. Fixation durations shorter 100 ms, as well as fixation durations that deviated for more than 2.5 SD from the mean were considered as artifacts and discarded from the analysis (SEM: 5 %, SYN: 7 %, EXSYN: 7 % of all fixations excluded).

Three paired student t-tests were conducted to calculate within-subject comparisons between the consistent control and the inconsistent-semantics experiment and the inconsistent-syntax experiments, respectively. Participants showed longer mean dwell times when the object was either semantically inconsistent, $t(14) = 6.37$, $p < .001$, when it was at a syntactically mild, $t(6) = 2.96$, $p < .05$, or syntactically extreme (i.e. floating) location, $t(6) = 4.37$, $p < .01$, compared to the consistent control condition. It should be noted that the results rely on different sample sizes (n = 15 vs. n = 7 vs. n = 7). However, in all experiments corresponding effect size measures (Cohen's d = 1.64 vs. d = 1.12 vs. d = 1.65) can be considered as large (Cohen, 1992). As shown in Fig. 6, the consistency effects in dwell times were comparable for semantic and syntax violations experiments.

Discussion

To our knowledge SCEGRAM is the first ready-to-use image database which includes real-world photographs depicting a range of well-controlled object-scene inconsistencies. So far, SCEGRAM includes 62 indoor scenes in six consistency conditions. Average low-level salience was calculated and controlled between consistency conditions for the whole database. Furthermore, consistency ratings were obtained for each object-scene condition in each scene. To our surprise, the extreme-syntax condition was not judged more extreme than the mild-syntax condition as could have been expected based on the ratings reported by Võ and Wolfe (2013). However, in their study both violations were not present in the identical scene and they also included objects balancing on edges, while our extreme syntax condition only included floating objects. As another possible explanation, one could argue that when semantic violations are presented, the syntactic violations seem less inconsistent. We observed higher degrees of inter-individual variance for the extreme-syntax than the other violations as reflected in higher standard errors and participant's verbal reports (e.g., one participant reported imagining a hand holding the object in midair): some participants considered the extreme-syntactic violations as inconsistent, while for others the hovering objects did not appear inconsistent as long as they fit the global meaning of the scene. Such inter-individual differences have not been systematically addressed so far, but suggest that depending on the specific sample, the ratings might vary. In general, we encourage any user of the database to replicate consistency and salience checks for the purpose of their own experiments. We provide detailed usage instructions in the Methods section.

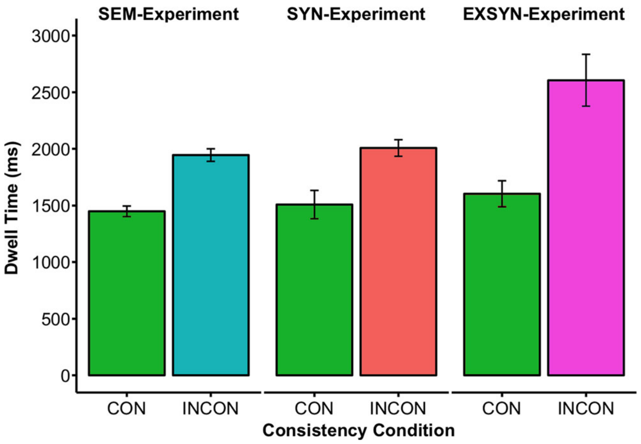

Fig. 6 Mean dwell times as a function of consistency condition (CON vs. INCON) in the inconsistent-semantics (SEM-Experiment, n = 15), the mild (SYN-Experiment, n = 7), and extreme (EXSYN-Experiment, n = 7) inconsistent-syntax experiment. Error bars indicate ±1 standard error

Despite the general ambiguity in any subjective consistency ratings, the cross-validation of the database using a simple eye-tracking paradigm provided a replication of previous findings in that we found longer dwell times for both semantic and syntactic inconsistencies compared to the control condition (e.g., Henderson et al., 1999; Võ & Henderson, 2009; see also Bonitz & Gordon, 2008; Rayner et al., 2009; Spotorno et al., 2015). We can now even expand and directly compare these findings to mild-syntax violations suggesting that our manipulations have the expected effects on eye-movement control. Future research might want to cross-validate the database in populations with different cultural backgrounds to make it accessible for studies on intercultural differences with regard to the perception of semantic and syntactic inconsistencies. The SCEGRAM database further includes a scene condition, in which the critical object is absent. When these scenes are presented successively with the corresponding object present scenes, the critical object is the only changing stimulus feature as it might be needed for some tasks (e.g., change detection) or gaze-contingent paradigms. To make the database also useful for pictorial priming paradigms, as well as tests of recognition memory for critical objects, we additionally provide object-only images, where the isolated object is displayed on a uniform white background.

In contrast to BOiS (Mohr et al., 2016), the only other publically accessible database on scene-object inconsistencies, SCEGRAM not only refers to the syntactic placement of the objects, but also contains object pairs of similar appearance, which occur in a semantically consistent and inconsistent scene. While BOiS was intentionally designed for visual search experiments, the SCEGRAM database can also be used in other tasks such as passive viewing, priming, change detection, or in flash-preview moving-window paradigms etc., as well as for developmental or EEG recording purposes, for which the objects are supposed to be easily identifiable. Hence, in terms of research purposes both databases might be considered as a complementation of one another.

In sum, with SCEGRAM we have created a well-controlled multifunctional database that we hope helps interested researchers to better address a broad range of research questions even beyond those aimed at investigating scene grammar.

Author notes Many thanks to Daniela Gresch, Anna Semenkova, Elisabeth Zey, Vanessa Tratsch, and Renaud Clerel for their help in the creation of the SCEGRAM database. This work was funded by DFG grant VO 1683/2-1 to MLV.

References

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology. Human Perception and Performance, 33*(1), 20–30. doi:10.1037/0096-1523.33.1.20

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177. doi:10.1016/0010-0285(82)90007-X

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica, 129*(2), 255–263. doi:10.1016/j.actpsy.2008.08.006

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. doi:10.1163/156856897X00357

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

De Graef, P., Christiaens, D., & D'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52*(4), 317–329. doi:10.1007/BF00868064

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. doi: 10.1109/CVPR.2009.5206848

Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset. Caltech, Technical Report*. Retrieved from http://authors.library.caltech.edu/7694

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*(1), 210–228. doi:10.1037/0096-1523.25.1.210

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert hifts of visual attention. *Vision Research, 40*(10-12), 1489–1506. doi:10.1016/S0042-6989(99)00163-7

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology. Human Perception and Performance, 4*(4), 565–572. doi:10.1037/0096-1523.4.4.565

Mohr, J., Seyfarth, J., Lueschow, A., Weber, J. E., Wichmann, F. A., & Obermayer, K. (2016). BOiS - Berlin object in scene database: Controlled photographic images for visual search experiments with quantified contextual priors. *Frontiers in Psychology, 7*, 749. doi:10.3389/fpsyg.2016.00749

Morey, R. D., Rouder, J. N., & Jamil, T. (2013). Package "BayesFactor": Computation of Bayes Factors for Common Designs. R package version 0.9.12-2. Retrieved September 5, 2016, from https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Nene, S. A., Nayar, S. K., & Murase, H. (1996). *Columbia Object Image Library (COIL-100). Technical Report* (Vol. 95). Retrieved from http://www.cs.columbia.edu/CAVE/databases/papers/nene/nene-nayar-murase-coil-100.ps

Olmos, A., & Kingdom, F. A. A. (2004). A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception, 33*(12), 1463–1473. doi:10.1068/p5321

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. doi:10.1163/156856897X00366

Rayner, K., Castelhano, M. S., & Yang, J. (2009). Eye movements when looking at unusual/weird scenes : Are there cultural differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 254–259. doi:10.1037/a0013508

Spotorno, S., Malcolm, G. L., & Tatler, B. W. (2015). Disentangling the effects of spatial inconsistency of targets and distractors when searching in realistic scenes. *Journal of Vision*, *15*(2, Art. 12), 1–21. doi:10.1167/15.2.12

Spotorno, S., Tatler, B. W., & Faure, S. (2013). Semantic consistency versus perceptual salience in visual scenes: Findings from change detection. *Acta Psychologica, 142*(2), 168–176. doi:10.1016/j.actpsy.2012.12.009

Torralba, A., Murphy, K. P., & Freeman, W. T. (2004). Sharing features: Efficient boosting procedures for multiclass object detection. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 I.E. Computer Society Conference on* (pp. 762–769). doi:10.1109/CVPR.2004.1315241

Turk-Browne, N. B., Scholl, B. J., & Chun, M. M. (2008). Babies and brains: Habituation in infant cognition and functional neuroimaging. *Frontiers in Human Neuroscience, 2,* 16. doi:10.3389/neuro.09.016.2008

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology, 59*(11), 1931–1949. doi:10.1080/17470210500416342

Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency and gist in the inspection of objects in natural scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye Movements: A window on mind and brain* (pp. 563–579). Elsevier. doi:10.1016/B978-008044980-7/50028-8

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. Consciousness and Cognition, *17*(1), 159–170. doi:10.1016/j.concog.2006.11.008

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3, Art. 24), 1-15. doi:10.1167/9.3.24

Võ, M. L.-H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception & Psychophysics, 73,* 1742–1753. doi:10.3758/s13414-011-0150-6

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science, 24*(9), 1816–1823. doi:10.1177/0956797613476955

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432. doi:10.1037/a0022790

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks: The Official Journal of the*

International Neural Network Society, 19(9), 1395–1407. doi:10.1016/j.neunet.2006.10.001

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492). doi:10.1109/CVPR.2010.5539970