

Perceptual and Decisional Factors Influencing the Discrimination of Inversion in the Thatcher Illusion

Katherine Cornes
Nick Donnelly
Hayward Godwin

Michael J. Wenger

The Pennsylvania State University

The University of Southampton

Draft April 22, 2010: In press, *JEP:HPP*. Please do not cite or distribute without permission.

The Thatcher illusion (Thompson, 1980) is considered to be a prototypical illustration of the notion that face perception is dependent on configural processes and representations. We explore this idea by examining the relative contributions of perceptual and decisional processes to the ability of observers to identify the orientation of two classes of forms—faces and churches—and a set of their component features. Observers were presented with upright and inverted images of faces and churches, in which the components (eyes, mouth, windows, doors) were present either upright or inverted. Observers first rated the subjective grotesqueness of all of the images, then performed a complete identification task, in which they had to identify the orientation of the overall form, and the orientation of each of the interior features. Grotesqueness ratings for both classes of image showed the standard modulation of rated grotesqueness as a function of orientation. The complete identification results revealed violations of both perceptual and decisional separability, but failed to reveal any violations of within-stimulus (perceptual) independence. In addition, exploration of a simple bivariate gaussian signal detection model of the relationship between identification performance and judged grotesqueness suggests that within-stimulus violations of perceptual independence on their own are insufficient for producing the illusion. This lack of evidence for within-stimulus configural processing suggests the need for a critical re-evaluation of the role of configural processing in the Thatcher illusion.

The vernacular conception of the perceptual and cognitive processing of faces is that it is in some (usually undefined) way holistic, or driven by the processing of configurations. However, what has generally been lacking is a detailed understanding of what is involved in the processing of facial configurations. One problem that has made understanding what configural processing entails is that a concrete definition of configural processing has not been made both explicit and precise.¹ In the current paper we investigate whether dependencies exist between features in upright and inverted faces to help elucidate the nature of configural processing within the context of a specific empirical effect: the Thatcher illusion.

This illusion involves the presentation of a facial image (originally that of Margaret Thatcher, former

Prime Minister of Great Britain, Thompson, 1980) in which some of the internal anatomical features (e.g., the eyes and mouth) are inverted from their biologically-appropriate orientation. When this altered stimulus is presented upright, the inversion of the internal features is detected quite readily; however, this is not the case when the stimulus is presented inverted. In many experimental investigations of the illusion, the dependent variable is a subjective rating, such as judged grotesqueness. Thus, any characterization of the illusion and its putative implications for configural percepts needs to address both objective performance (orientation detection) and subjective phenomenology.

For the most part, previous research that has examined whether there are dependencies between features has been based upon experimental paradigms using speed and/or accuracy judgments (e.g., Bartlett & Searcy, 1993; Tanaka & Farah, 1993; Young, Hellawell & Hay, 1987), or matching across different conditions (e.g. Davidoff & Donnelly, 1990), guided almost completely by attempts to define configural processing

This work was supported in part by an award (to KC) from the Economic and Social Research Council of Great Britain. Sincere thanks are due to Noah Silbert, Jim Townsend, Tammy Menneer, Leslie Blaha, Jennifer Bittner, Rebecca Von Der Heide, and Danny Fitousi for comments on an earlier version of this paper. Correspondence: Michael J. Wenger, Department of Psychology, The Pennsylvania State University, mjw19@psu.edu.

¹ This, of course, is a long-standing criticism of multiple approaches to perceptual organization and object perception and identification (e.g. Uttal, 1988).

(holism, gestalts, etc.) operationally rather than theoretically. One problem with adopting such an approach is that it is often difficult to determine whether the results that are generated are underpinned by true differences in processing, or whether they are due to other potentially-unrelated factors, such as variations in performance that potentially have little or nothing to do with the central theoretical distinction. In addition, the set of criteria for inferring configural processing that exists in the literature has not, until recently, considered “varieties” of configural processing and representation. For example, O’Toole, Wenger, and Townsend (2001) noted that one possible type of configural processing may have a decisional, as well as a perceptual (or representational) source.

In order to empirically consider such hypotheses, and to distinguish variations in performance due to variations in sensitivity from shifts in response criterion, one needs a theoretical “language” capable of clearly characterizing the logical distinctions, and associated methods and measures capable of connecting theory and data. Traditionally, the distinction between perceptual and decisional components has been done by way of signal detection theory (SDT, e.g., Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). For example, one can approach this theoretical distinction empirically through the use of the signal detection parameters for sensitivity (d') and response criterion ($z(FA)$,² β , c). However, classic signal detection theory is framed with respect to single signal dimensions. If the question of interest pertains to multiple stimulus dimensions and their relations, it is necessary to consider generalizations of the approach to multiple dimensions (e.g., Ashby & Townsend, 1986; Macmillan & Creelman, 2005, chapters 6-10).

A Theory-based Characterization of Configurality

The particular generalization that has been used to address these questions is known as general recognition theory (GRT, Ashby & Townsend, 1986). In a classic, unidimensional SDT framework, participants are typically asked to respond to the presence of a signal embedded in a background of noise. SDT assumes that observers collect information from a stimulus, and then use the value of that information to make a decision about the nature of that stimulus (i.e. whether or not the signal is judged to be present). In order to make that decision, a *criterion*, is set, so that encoded values at or above that cut-point result in “signal-present” responses, and values below that cut-point result in “noise-alone” responses. The distributions of encoded values resulting from presentations of signal-present and noise-alone stimuli are typically idealized as gaussian, and the degree to which these two distributions are separated determines how sensitive observers are to the true state of the stimulus. This sensitivity is typically quantified using an estimate of the distance between the two distributions, such as d' . The location of the criterion, which can be

set independent of the distance between the distributions, determines the extent to which an observer will give a particular response; this response bias is typically quantified using an estimate of the location of the cut-point, such as c .

GRT is a multidimensional generalization of SDT. The most powerful aspect of GRT is that it can explicitly (theoretically) characterize and empirically assess the *perceptual* and *decisional* dependencies that exist among the dimensions. Thus, the tools of GRT can be used to predict and then measure how sensitivity and criterion values for a given dimension are affected by changes in the other dimensions in a stimulus. If the sensitivity for one dimension changes depending on the level of another dimension then we can say that we have evidence suggesting a perceptual dependency. Additionally, if the criterion for one dimension changes depending on the level of another dimension then we can say that we have evidence suggesting a decisional dependency. These two types of dependencies can exist for all possible elements of a stimulus. Both of these dependencies can be interpreted as revealing global configural processing, of two very different types (see discussions in O’Toole et al., 2001; Wenger & Ingvalson, 2002).

The best way to illustrate the theoretical distinction between perceptual and decisional dependencies and how each can be used to represent hypotheses regarding configurality is to consider the simplest possible case: that of two stimulus dimensions. Assume that we wish to present observers with stimuli composed by using two levels of each of these two dimensions, resulting in a total of four possible stimuli. We begin by plotting theoretical distributions of perceptual evidence resulting from the presentation of signal and noise for each stimulus dimension in a two-dimensional joint probability space (see panel (a) of Figure 1). To simplify the representation it is possible to insert a horizontal plane through the space at some arbitrary level. When viewed from above, the representation then becomes a set of equal likelihood contours (see panel (b) of Figure 1).

Perceptual dependencies can occur in two ways. First, a perceptual dependency may exist as a violation of perceptual independence (PI). Stimulus dimensions are perceptually independent when the perceptual effect of one dimension is statistically independent of the perceptual effect of another dimension. This is a strong form of dependency, and exists within a *single* stimulus. This distinguishes violations of PI, as all of the other types of dependencies exist as relations *across* stimuli. If we assume that the theoretical distributions are (for simplicity) equal-variance bivariate Gaussians, then perceptual independence is represented by circular equal likelihood contours, indicating a lack of correlation.

Second, a perceptual dependency may exist as a vi-

² The inverse z -transform of the false alarm (FA) rate, labeled as λ in some applications (e.g., Wickens, 2002) and C in others (e.g., Kadlec, 1999). We define the quantities relevant to the present study below.

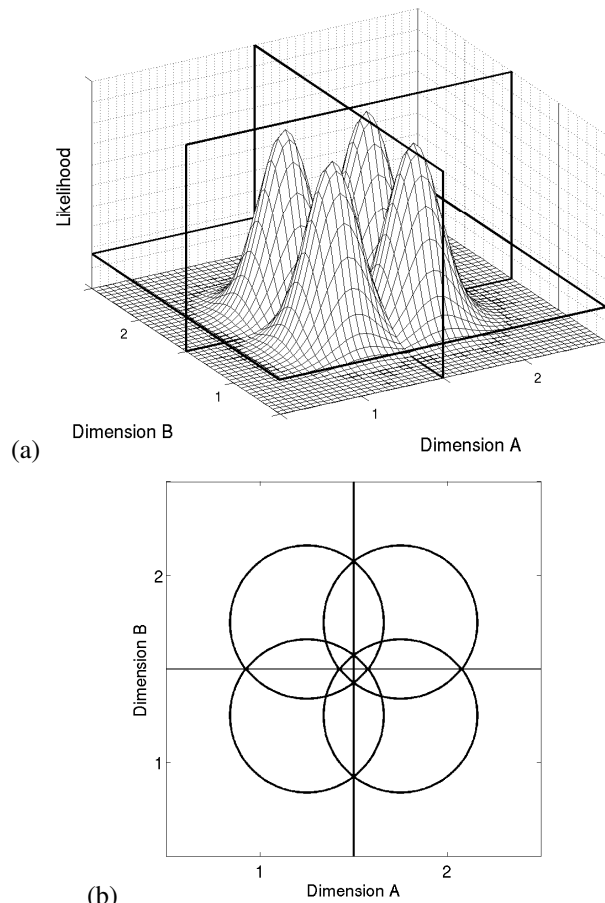


Figure 1. Theoretical representation of distribution of perceptual information from the perspective of GRT. (a) Joint probability distributions for stimuli composed of the combination of two dimensions each at two levels. The planes in the interior of the space represent the two decision bounds (b) Equal-likelihood contours for this same configuration, along with the decision bounds.

olation of perceptual separability (PS). For two dimensions to be perceptually separable the perceptual effects of one dimension must not vary across the levels of the other dimension. If PS holds for both dimensions then the centers of the equal likelihood contours can be connected to form a rectangle. Decisional dependencies between dimensions of a stimulus can exist as a violation of decisional separability (DS). Stimulus dimensions are decisionally separable when the location of the decision boundary (criterion) for one dimension is unaffected by the level of the other dimension. If decisional separability holds and if it can be assumed that the bounds are linear (as in Ashby & Townsend, 1986; Kadlec & Townsend, 1992a,b), then the decision boundaries are parallel to the dimensional axes.

Applications of GRT to Configural Processing of Faces

The GRT approach has been adopted in a number of recent studies where classic demonstrations of configural processing have been re-examined. For example, Richler and colleagues (2008) re-examined the composite face effect. In traditional versions of the composite face effect participants are asked to learn a set of faces. The top half of one of these faces is then presented with the bottom half of another face. When the face halves are aligned participants are slower and less accurate at identifying either face half compared to when the face halves are misaligned. Traditionally this result has been used to argue that dependencies exist in the perception of the two halves of the face, which would be consistent with configural processing. In Richler et al.'s (2008) version of the composite face task, participants were presented with a face that had the top and bottom halves of the face, either aligned, misaligned or very misaligned. Participants were subsequently presented with another face that was also aligned, misaligned or very misaligned and were asked to determine whether the top and bottom halves of the face were the same, whether the top was the same and the bottom was different, whether the top was different and the bottom was the same or whether both the top and the bottom were different. In this task, a violation of perceptual separability meant that one face half is perceived as being *more similar to* or *more different from* the study face depending on the perceived same or differentness of the other face half. A violation of perceptual independence meant that the variability in the perceived same or different status of one face half is correlated with the perceived same or differentness of the other face half. Finally, a violation of decisional separability means that the location of the decision criterion for generating the response depends on the same or differentness of the other face half.

The results showed violations of perceptual separability and decisional separability, but not violations of perceptual independence. In addition, simulation results suggested that violations of PI alone would not be sufficient to produce the standard behavioral results indicative of configural processing. In other words, the results showed that the variability in the perceived same or different status of one face half was not correlated with the perceived same or different status of the other face half, and that the location of the decision criterion depended on the same or different status of the other face half. Importantly, the results showed that the magnitude of the violations of decisional separability decreased with the magnitude of the relative misalignment. In other words, as the face halves became more aligned the magnitude and number of violations of decisional separability increased. In contrast, the number and magnitude of the violations of perceptual separability did not change with relative misalignment. These results demonstrate that the difference between the aligned and non-aligned conditions, in the

composite face effect, is underpinned by differences in decisional rather than perceptual factors (but see Richler et al., 2007).³

In earlier studies, Wenger and Ingvalson (2002, 2003) re-examined the task developed by Farah, Wilson, Drain and Tanaka (1998), and used in earlier work by Tanaka and colleagues (e.g., Tanaka & Farah, 1993; Tanaka & Sengco, 1997). In Farah et al.'s experiment, participants were presented with a target face. They were then presented with a test face and were asked to make same/different decisions about an individual pre-designated feature. In compatible trials, if the target features were the same, then the irrelevant features were also the same (alternatively, if the target features were different, then the irrelevant features were also different). In incompatible trials, if the target features were the same, then the irrelevant features were different (alternatively, if the target features were different, then the irrelevant features were the same). Farah et al. (1998) showed that reaction times (RTs) to make same decisions were faster when irrelevant features were compatible, compared with when irrelevant features were incompatible, suggesting that RTs were influenced by task irrelevant features. The authors used these results to argue that responses to specific features were dependent on the level of other features. Although the authors suggested that the effect was perceptual in nature, their data revealed a shift in response bias as a possible contributing factor.⁴ Using the GRT approach, Wenger and Ingvalson (2002, 2003) showed that differences in decisional dependencies distinguished between compatible and incompatible conditions in the overwhelming majority of cases.

Applying GRT to the Thatcher Illusion

The aim of the work presented in this paper is to further investigate whether there is evidence of dependencies between features in faces. To do this we examine observers' ability to identify the orientation of internal features and external forms in the types of stimuli used in work on the Thatcher illusion (Thompson, 1980). In the present experiment, normal faces are discriminated from faces with both eyes and mouths inverted (what we refer to as the full illusion condition) and faces where only eyes or mouths have been inverted (what we refer to as the partial illusion condition). In order to use the empirical tools associated with GRT, we take an approach that is unusual for explorations of the Thatcher illusion but that is a standard in work applying GRT. The stimuli will be composed of three experimentally-manipulated features: the external form (the facial surround), and two internal features (the eyes and the mouth). Each of these features will exist at two levels: upright and inverted. Forming all possible combinations across these factors results in a total of eight possible stimuli, two of which are typically used in an experiment investigating the Thatcher illusion: one in which the external form is upright and the two internal features are inverted (Fig-

ure 2(a)), and the other in which all three dimensions are inverted (Figure 2(b)).

At this point, we can take advantage of that fact that GRT allows one to explicitly model perceptual and decisional interactions and obtain explicit predictions for the results of those interactions. To do this in the context of the Thatcher illusion, we begin with the assumption that repeated presentation of stimuli constructed as described in the preceding paragraph results in distributions of (encoded) evidence (Ashby & Lee, 1993) for each of the encoded dimensions of the stimulus. For the purposes of illustration, assume that these are the external form and one of the internal anatomical features (e.g., the eyes), each of which can exist in two states (upright, inverted). Assume next that the form of the internal distributions for the encoded evidence is bivariate gaussian. Finally, assume that responses are generated using either continuous- or piecewise-linear decision bounds.

Thus, a fully-specified GRT model for the example situation of two dimensions each with two levels requires at least 22 parameters. Each of the bivariate gaussian distributions (on the level of encoded evidence) requires two means (expressed as a mean vector)

$$\mu = \begin{bmatrix} \mu_E \\ \mu_I \end{bmatrix},$$

two variances, and a correlation parameter (expressed as a covariance matrix)

$$\Sigma = \begin{bmatrix} \sigma_E^2 & \rho\sigma_E\sigma_I \\ \rho\sigma_E\sigma_I & \sigma_I^2 \end{bmatrix},$$

where the subscripts E and I index the parameters for the marginal distributions on perceptual evidence for the external and internal features, respectively. In addition, each of the decision bounds requires an intercept, one for each dimension in the case of continuous linear bounds (γ_E and γ_I) and two for each dimension in the case of piece-wise linear bounds (e.g., γ_{I1} and γ_{I2}).

The specific choices of values for these parameters allows us to explicitly predict the effects of violations

³ Richler et al. (2007) have suggested that there may be models absent decisional components that can account for these effects. The particular model of concern is one originally published by Dailey and Cottrell (1998), and the extent to which this model (or other similar models) can produce these types of effects is potentially of great significance. However, at the time of this writing, the potential correspondences between the mechanisms of that model and the definitions of PI, PS, and DS have yet to be completely articulated, making it difficult to determine whether there should be any critical concerns above and beyond those already noted in the literature (beginning, e.g., with Ashby & Townsend, 1986). Still, we acknowledge the concern and address it by adopting a method of converging evidence to support our inferences.

⁴ It is important to note, though, the Farah and colleagues were not using signal detection theory to frame their hypotheses or make any distinctions between perceptual and decisional determinants of the outcomes.



Figure 2. An illustration of the Thatcher illusion. A face that is perceived as grotesque when upright is perceived as much-less grotesque when inverted.

of perceptual independence, perceptual separability, and decisional separability on variations in response frequencies and thus, by extension, measures of accuracy, sensitivity, and response bias. In order to model the effects of these violations on the relevant phenomenology of the Thatcher illusion, we simply assume that judged grotesqueness is proportional to the ability to correctly identify the orientation of the internal feature.

We begin with a baseline model, representing the null hypothesis of no perceptual or decisional interactions. A schematic representation of this model (model 0) is presented in Figure 3(a), the predictions for measures of response accuracy (hit rates) and sensitivity (d') are presented in Figure 4(a), and the predictions for two measures of response bias are presented in Figure 4(b). The parameter values used for this and all of the examples in this section are presented, along with a description of the methods used for the numeric simulations, in Appendix A. As expected, none of the measures show any differences as a function of orientation. The lack of any difference in the accuracy and sensitivity measures suggests no difference in judged grotesqueness. This most naturally represents the *absence* of the Thatcher illusion.

The first of the models that incorporate dimensional interactions is parameterized as a violation of perceptual separability for the internal feature across the two levels of the external form. A schematic representation of this model (model 1) is presented in panel (b) of Figure 3, and the predictions for the measures of performance are presented in panels (a) and (b) of Figure 4. For this model, both accuracy and sensitivity to the orientation of the internal feature was greater when the external form was upright, relative to when it was inverted. Under our simple assumption regarding judged grotesqueness, this model would produce the expected orderings on grotesqueness ratings indicative of the Thatcher illusion. With respect to the measures of response bias, note that one of the measures— c —was invariant across the two orientations, the other— $z(FA)$, the inverse normal transformation of the false alarm rate—was not. This can be understood by considering the information used in each of these mea-

asures. Where c is calculated using both hit and false alarm rates, $z(FA)$ uses only the false alarm rates. Thus, although the false alarm rate changes across orientation, the hit rate also increases, and does so proportionally. Thus, c correctly indexes a lack of change in the relative location of the criterion, while $z(FA)$ does not. That being said, however, $z(FA)$ is correctly detecting a change in the false alarm rates.

The second of the models to incorporate dimensional interactions (model 2) is a second violation of PS. The schematic representation of this model is presented in panel (c) of Figure 3, and the predictions for the measures of performance are presented in panels (a) and (b) of Figure 4. As was true for model 1, both accuracy and sensitivity for the internal feature was greater when the form was upright rather than inverted, indicating that (under our assumptions) this model is capable of producing the Thatcher illusion. However, in contrast to model 1, this manner of violating PS produces opposite outcomes with respect to the two measures of response bias. For model 2, $z(FA)$ is invariant across orientation while c is not. Note however, that this is correct relative to the geometry of the model that generated the data, which produces no increase in false alarm rates, and an increase in hit rates.

Here it is important to note that the two measures of response bias performed differently for these two types of violations, which are distinguished by the concurrent changes predicted for the hit and false alarm rates. This suggests that, should empirical differences be observed in the response bias measures, the hit and the false alarm rates should be examined to determine whether there is a possibility that an inference of a violation of DS may be suspect.

The third model to consider is one that represents the dimensional interactions in terms of a violation of decisional separability for the internal features across the two levels of the exterior form.⁵ The schematic repre-

⁵ A standard and perhaps modal interpretation of a violation of DS (a shift in criterion) is in terms of a conscious, volitional, strategic choice, and thus a “late” component of processing.

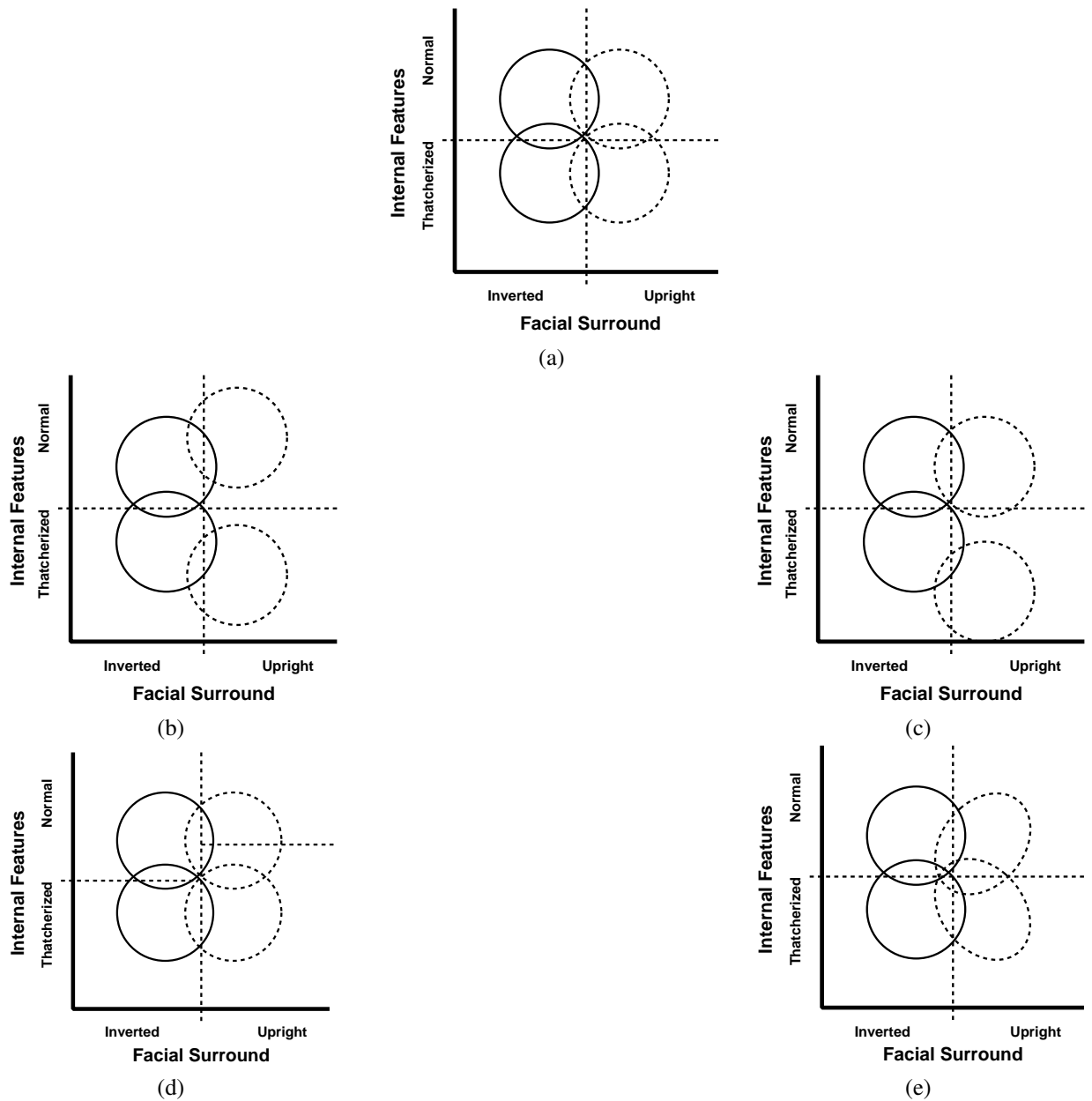


Figure 3. Schematic representation of the five models used to illustrate predictions for the Thatcher illusion: (a) no violations, (b) violation of PS, (c) alternative violation of PS, (d) violation of DS, (e) violations of PI.

sensation of this model (model 3) is presented in panel (d) of Figure 3, and the predictions for the measures of performance are presented in panels (a) and (b) of Figure 4. This model predicts that the measure of response accuracy is higher when the exterior is upright rather than inverted, but that sensitivity is actually invariant across orientation. This is because the measure of response accuracy uses only the hit rates, while the measure of sensitivity uses both the hit and false alarm rates. This model also (correctly) produces estimates of response bias that do vary as a function of orientation. Consequently, this

model can be interpreted as suggesting that a violation of DS is capable of producing the Thatcher illusion, but only on the basis of changes in the hit rate (correctly identifying the orientation of the internal feature).

The fourth and final model to consider is one that represents the dimensional interaction in terms of a violation of perceptual independence. The schematic repre-

Although this is possible, it is not the only means by which a shift in criterion might occur, there being a number of reasonable possibilities for “early” or unconscious mechanisms. We return to this issue briefly in the general discussion.

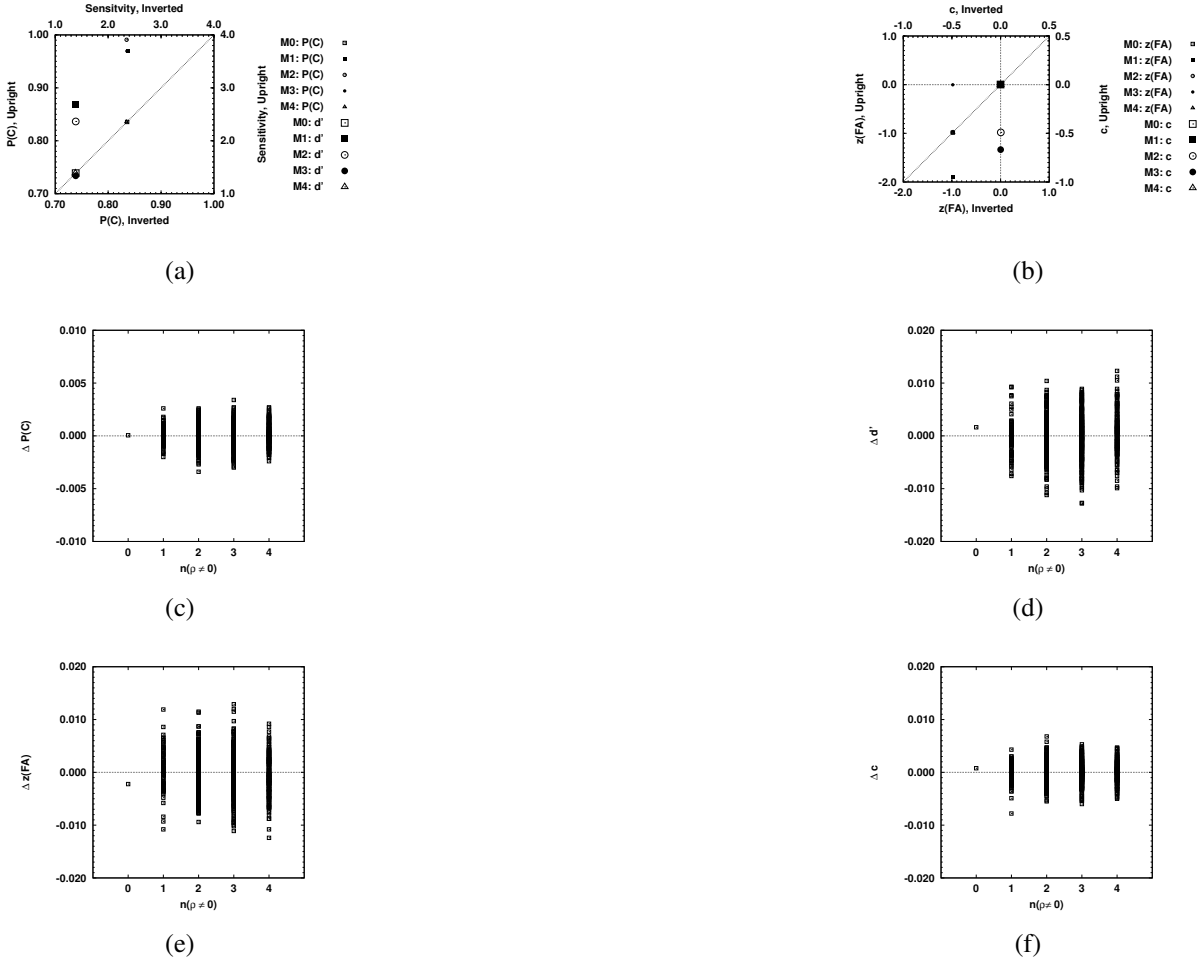


Figure 4. Results of the simulations of the five models used to illustrate predictions for the Thatcher illusion. (a) accuracy and response sensitivity for each of the five models, (b) two measures of response bias for each of the five models. Predicted differences in (a) response accuracy (hit rates), (b) response sensitivity (d'), (c) one measure of response bias ($z(FA)$), (d) a second measure of response bias (c) for internal features across orientation of the image, for 0-4 violations of PI at all values of sign and magnitude (see text for details).

sentation of this model (model 3) is presented in panel (e) of Figure 3, and the predictions for the measures of performance are presented in panels (a) and (b) of Figure 4. This model—the one that best captures the notion of a within-stimulus gestalt or configularity—suggested that the measures of accuracy, sensitivity, and response bias were *all* invariant across orientation. This somewhat paradoxical *lack* of ability of a violation of PI to produce what has been interpreted as an empirical “signature” of perceptual holism has been noted in other contexts as well (Richler et al., 2008).

To get a sense as to whether this outcome was at all dependent on the specific parameters we used, we conducted an additional series of numeric simulations, in which we considered all possible combinations of number, sign, and magnitude of non-zero correlation parameters. The method for these numeric simulations is presented in Appendix A, and the results are presented in the

panels (c)-(f) of Figure 4. Here it can be seen that across all possible parameter values and their combinations, the measures of accuracy (panel (c)), sensitivity (panel (d)), and both measures of response bias (panels (e) and (f)) varied very little (less than one-half of one percent for accuracy and little more than one one-hundredth of a standard deviation unit for the other measures) across orientation. Thus, it appears that a violation of PI *by itself* is unable to produce the changes in performance indicative of the Thatcher illusion.⁶

The models presented thus far show how performance differences across image orientation indicative of the Thatcher illusion can result from violations of both per-

⁶ We thank Noah Silbert for pointing this potential out to us during a review of an early version of this paper. Of course, it is possible that a violation of PI could exist along with violations of either PS, DS, or both.

ceptual and decisional separability, but not a violations of perceptual independence (alone). In addition, it is logically possible to consider how combinations of violations could produce the Thatcher illusion. A critical implication of this range of possibilities is that a similar range of possibilities could exist across observers: that is, variations across the models suggest the strong potential for differences across individuals. All of this, of course, is to be understood given the assumption that judged grotesqueness is proportional to the accuracy of identifying the orientation of the dimensions of the stimulus. Thus, along with the hypotheses for the various perceptual and decisional components of orientation detection performance, we have the hypothesis that variations in judged grotesqueness should be systematically related to variations in identification accuracy, sensitivity, and (potentially) response bias.

In order to empirically test for the possible violations that could produce the Thatcher illusion, we need to be able to collect the data required to support inferences regarding the predicted effects of the violations of PI, PS, and DS. The standard experimental approach used for this purpose is referred to as the complete identification paradigm (Ashby & Townsend, 1986; Kadlec & Townsend, 1992a,b; Kadlec & Hicks, 1998; Townsend, Hu & Ashby, 1981). In the complete identification paradigm participants are presented with stimuli that have multiple features present at multiple levels and every possible combination of levels is present. In the complete identification paradigm implemented here, faces (or churches) could have either eyes (windows), mouths (doors), or face (church) outlines upright or inverted. We included churches as comparison stimuli due to the similarity of the configuration of internal features to the configuration of internal features in faces, thus maintaining a standard geometry across two classes of images. Every possible combination of manipulations were presented to participants. Participants were required to make decisions about the orientation of each feature on every trial. Participants were asked to determine whether the face (church) outline was upright or inverted, whether the eyes (windows) were upright or inverted and whether the mouth (door) was upright or inverted.

Method

Participants

Four undergraduate students from The Pennsylvania State University were recruited for this experiment and were remunerated at a rate of \$8 per hour. Participants had a mean age of 23.1 years ($SD = 0.2$ years). Three were female and one was male. All were right handed and all reported normal or corrected to normal vision.

Ten photographs were taken of five males and five females who were enrolled in the Psychology program at The Pennsylvania State University. Models were photographed in front view against a white background wearing surgical scrubs with a surgical hat covering

their hair. These photos were then manipulated using Adobe Photoshop. Faces were extracted from their backgrounds, cropped below the chin and resized so that the distance between pupils measured 0.7 cm (see Figure 5 for examples of the stimuli). In addition to unaltered faces, three sets of altered faces were created, one with the eyes ‘Thatcherised’, one with the mouth ‘Thatcherised’ and one with both the eyes and mouth ‘Thatcherised’. ‘Thatcherised’ faces were created by cutting out eyes and/or mouths, inverting them and then pasting them back into the context of the face. The blur tool was used to remove high contrast edges that could act as local featural cues.

Ten photographs of churches were obtained from sources on the internet. Churches were photographed in front view and were selected so that they had a central door and two windows either side of the door to approximately equate the position of the eyes and mouth within a face. Churches were extracted from their backgrounds using Adobe Photoshop and were pasted onto a white background. Churches were resized so the distance between the centres of the windows measured 0.7 cm (see Figure 5 for examples of the stimuli). In addition to unaltered churches, three sets of altered churches were created, one with the windows ‘Thatcherised’, one with the door ‘Thatcherised’ and one with both the windows and the doors ‘Thatcherised’. ‘Thatcherised’ churches were created by cutting out the windows and/or door, inverting them and then pasting them back into the context of the church. The blur tool was used to remove high contrast edges that could act as local featural cues.

Design and Procedure

Participants completed 1-2 sessions per day for 9 consecutive days, excluding weekends. In session 1, participants completed a face/church rating task. Each face/church was presented twice and participants were asked to rate the face/church for grotesqueness, for a total of 160 rating trials for each stimulus class. Participants responded by clicking a bar that was displayed below each face/church and was labeled from 0-100. In addition, each end of the scale was accompanied by three descriptors. The positively-valenced descriptors were “typical,” “normal,” and “pleasant.” The negatively-valenced descriptors were “unusual,” “strange,” and “grotesque.” For two of the observers, the negatively-valenced descriptors were placed on the left end of the scale (low values) and the positively-valenced descriptors were placed on the right end of the scale (high values). This placement was reversed for the other two observers.

In sessions 2-9, participants were presented with each face/church and were asked to decide whether the outline was upright or inverted, the eyes (windows) were upright or inverted and whether the mouth (door) was upright or inverted. Each session lasted approximately 1 hr. Each session was comprised of 5 identical blocks. In each block, each of the 10 faces and 10 churches was presented 8 times, four times upright and four times in-

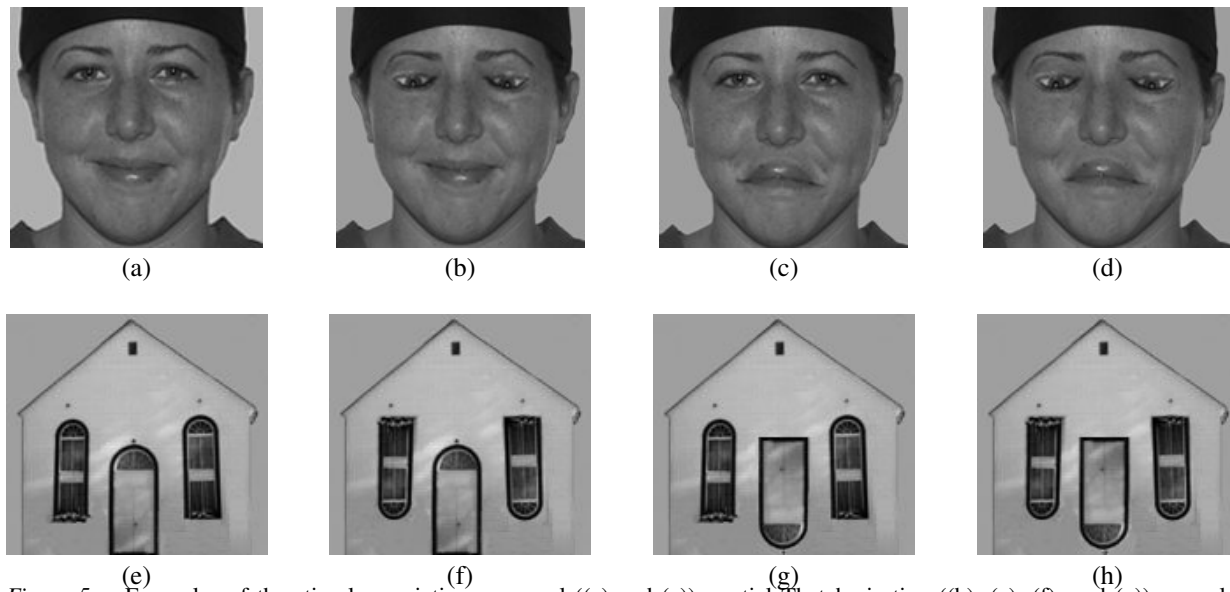


Figure 5. Examples of the stimulus variations: normal ((a) and (e)), partial Thatcherisation ((b), (c), (f), and (g)), complete Thatcherisation ((d) and (h)).

verted. Faces were presented once with eyes and mouths upright (i.e. a normal face), once with the eyes inverted, once with the mouth inverted and once with the eyes and mouth inverted. Churches were presented once with the windows and doors upright, once with the windows inverted, once with the door inverted, and once with the windows and doors inverted. The order of faces and churches was counterbalanced, as was orientation and manipulation type.

Each trial started with a fixation cross that was present for between 400 and 800 ms; the specific value used on each trial was drawn from an exponential distribution with a mean of 600 ms, using only values between 400 and 800 ms. Once the fixation cross disappeared, a pattern mask was presented for 100 ms. Following the mask, the face was presented for 100 ms. The mask was then presented again to prevent afterimages. Participants were asked three successive questions:⁷ Is the face outline upright or inverted? Are the eyes upright or inverted? Is the mouth upright or inverted? Participants responded using the right and left mouse keys. Response keys were counterbalanced between participants.

Results and Discussion

The data were examined with respect to three general questions. First, did the stimuli used in this task produce the standard Thatcher illusion in the grotesqueness ratings? Second, did those same stimuli produce any systematic violations of PI, PS, and/or DS? Third, were there systematic relationships between the patterns in the phenomenological data (the grotesqueness ratings) and the complete identification data, in particular shifts in the sensitivity and bias measures? All data were analysed at the level of the individual observers. Unless otherwise noted, a criterion (α) of 0.05 was used to infer statistical

reliability.

Phenomenology: Grotesqueness Ratings

Prior to analysis, observers ratings' of each of the images' grotesqueness were first transformed so that all ratings ranged from low (normal) to high (grotesque).⁸ Each rating for each observer was then transformed to a Z-score, by subtracting the mean for each stimulus type (face or church) for that observer, then dividing by the corresponding standard deviation. These Z-scores were then analysed, separately for each observer, using a 2 (type: face, church) \times 2 (orientation of the test image: upright, inverted) \times 3 (level of Thatcherization: none, partial, complete) analysis of variance (ANOVA). Note that the "partial" Thatcherization collapses across the two types of single-element modifications (mouth/door only, eyes/windows only) as preliminary analyses indicated that these two changes resulted in statistically indistinguishable effects in the subjective ratings ($t < 1.0$, ns).

The means for all four observers, for all image types and modifications, are presented in the two panels of Figure 6, and the results of the ANOVA are presented in Table 1. In the figure, the smallest symbols represent the unmodified images, the medium symbols repre-

⁷ This type of sequential responding was used in some of the earliest applications of GRT to perceptual organization (e.g., Kadlec & Hicks, 1998; Townsend & Thomas, 1994; Townsend, Hu & Kadlec, 1988), and was more recently used in Richler et al.'s (2008) examination of the composite face effect.

⁸ Overall, the entire range of the rating scale (transformed from screen coordinates to values running from 0 to 100) was used for both classes of stimuli, with a mean (on this transformed scale) of 44 for all faces and 49 for all churches.

sent the partially-Thatcherized (one feature change) images, and the large symbols plot the means for the fully-Thatcherized images. Means for upright versions of each image are given by the y -coordinate, and means for the inverted versions are given by the x -coordinate. The diagonal line in each pattern represents equality for the upright and inverted ratings; points above the line are means for which rated grotesqueness when upright exceeds rated grotesqueness when inverted.

Two results are critical. The first is the one that establishes the standard Thatcher illusion for these data: a reliable interaction between stimulus orientation and level of Thatcherization. This effect (reliable for three of the four observers) can be seen in Figure 6 as the increase in distance above the diagonal for the medium and large symbols: relative to the unmodified stimuli, increasing levels of Thatcherization produced grotesqueness ratings which are increasingly larger when upright rather than inverted. The second result is the one that substantiates a lack of qualitative difference between faces and churches as stimuli: the lack of a reliable three-way interaction between image type, level of Thatcherization, and stimulus orientation. Although it is the case that, overall, the magnitude of the variations in rated grotesqueness are far lower for churches than for faces, the basic qualitative pattern is the same for both stimulus types.

Multidimensional Signal Detection Analyses

The basis for our inferences regarding potential violations of PI, PS, and/or DS were two sets of analyses applied to the identification/confusion matrixes summarizing each observer's performance with each of the two classes of stimuli. The first set of analyses involves tests of equalities that have been shown (analytically) to hold under assumptions regarding PI, PS, and DS (Ashby & Townsend, 1986; Kadlec, 1999; Kadlec & Townsend, 1992a,b). The measures used are presented in Table 2 and the logic of inference is presented in Table 3. All of these analyses were done on all possible pairings of each of the two internal features and the external form. The foundations of this first approach were initially developed by Ashby and Townsend (Ashby & Townsend, 1986), and Kadlec and Townsend (Kadlec & Townsend, 1992a,b), and has been used in a variety of studies of face perception and memory (e.g., Richler et al., 2008; Thomas, 2001a,b; Wenger & Ingvalson, 2002, 2003), and a broader range of work in perceptual organization in general (e.g., Amazeen, 1999; Copeland & Wenger, 2006; Kadlec & Hicks, 1998). The present study differed from these earlier works in that we restricted consideration to analyses of marginal measures of performance, along with a non-parametric test of sampling independence (see Table 2) that speaks to inferences regarding PI.⁹

As can be seen in Table 3, there are a number of cases in which inferences regarding possible violations are uncertain. This potential inferential challenge has been ac-

knowledged in a number of ways since the initial work on GRT (Ashby & Townsend, 1986) was published. Specifically, many of the logical relations between theory and data were predicated on DS holding, with inferential indeterminacy existing if this could not be assumed (Ashby & Townsend, 1986; Kadlec, 1999; Kadlec & Townsend, 1992a,b; Thomas, 2001a,b, 2006). In addition, Kadlec's (1999, see p. 385) computational implementation of the inferential logic explicitly acknowledged potential difficulties in making inferences regarding violations of DS. However, as can be seen in panels (b) and (f) of Figure 4, the estimator for response bias used in this and our previous studies (c, Copeland & Wenger, 2006; Richler et al., 2008; Wenger & Ingvalson, 2002, 2003) show little if any problems of mis-estimation under varying types and magnitudes of violations in PI and PS, and ongoing work examining alternative statistical methods (Menneer, Wenger & Blaha, 2009b; Menneer, Silbert, Cornes, Wenger, Townsend & Donnelly, 2009a) show limited Type I error rates for the present set of estimators.

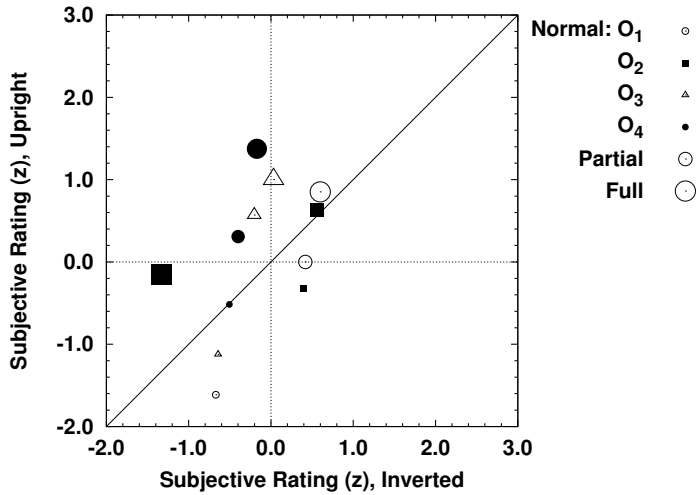
This being said, there are always reasonable concerns regarding inferential errors. In addition, we wanted to augment the non-parametric measures used to guide inferences regarding PI with an estimate of the correlation parameter for each of the bivariate gaussian distributions. To do this, we followed the approach advanced by Thomas (2001a), and used a hierarchical model-fitting procedure to fit a range of parameterized GRT models to the identification/confusion matrixes of each of the observers.¹⁰ Details of the model-fitting procedure are provided in Appendix B. Thus, with the combination of the marginal analyses and the fits of the gaussian models, we have two independent and converging methods for guiding our inferences.

Results of the comparisons on the marginal measures and the test of sampling independence are presented in Table 4, and the inferences drawn from both the marginal analyses and the model fits are presented in Table 5.¹¹ A few comments are in order regarding the correspondences between the inferences drawn from the marginal analyses and the model fits. The first is the generally high level of correspondence: the two methods supported identical inferences in 80% (77 of 96) of the total comparisons. The model fits resolved ambiguities in the in-

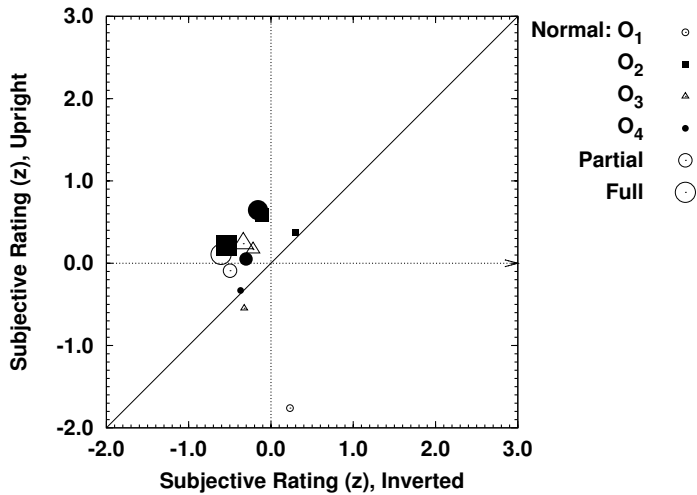
⁹ The relationship is established in Theorem 1, p. 160, of Ashby and Townsend (1986). In brief, (a) sampling independence within any one stimulus holds in any one stimulus if PI and DS both hold, (b) PI holds in any one stimulus if sampling independence holds across differences in decisional criteria as long as those criteria are parallel to the coordinate axes (piecewise linear bounds), and (c) if the decision bounds are not parallel to the coordinate axes then sampling independence is not logically related to PI.

¹⁰ Note that fitting parameterized models to behavioral data has been a standard and often-used approach in the literature on GRT (e.g., Copeland & Wenger, 2006; Maddox & Bogdanov, 2000; Maddox, 2001; Thomas, 2001a)

¹¹ The numerical values for all of these comparisons and all of the model fits are available on request from the authors.



(a)



(b)

Figure 6. Mean normalized (Z) subjective grotesqueness ratings for (a) faces and (b) churches, for each of the four observers, as a function of the orientation of the external form (inverted, upright).

ferences based on the marginal analyses in 17% (16 of 96) of the total comparisons. And the remaining 3% (3 of 96) of the cases, the model fits suggested an inference that was nominally different than the corresponding inference suggested by the marginal analyses, specifically suggesting a violation where one was not suggested by the marginal analyses (e.g., a difference in distribution means suggesting a violation of PS). However, in all of these cases, examination of the confidence intervals for the parameters in question overlapped, meaning that the inference of a violation was not supported, thus making the inferences drawn from the two approaches consistent. This high level of correspondence is one that has been observed in other work with GRT (e.g., Copeland & Wenger, 2006).

With respect to violations of PI, there was no evidence for any violations of PI, for any of the stimulus types, for any of the four observers. Note that in the marginal analyses we used an adjustment (Bonferonni) for multiple comparisons in these analyses that is known to be quite conservative (see, e.g., Zar, 1999). However, even when this correction was not applied, the evidence for any violations of PI was extremely limited. This result is consistent with the predictions of the example models in the introduction, and is one that has been documented frequently in studies applying the GRT approach to face perception and memory (Richler et al., 2008; Wenger & Ingvalson, 2002, 2003) This suggests that the “types” of holism or configularity (O’Toole et al., 2001) implied by many operational definitions of holism or configularity

Table 1

ANOVA results (*F*-ratios) for the grotesqueness ratings, by observer. Note: error *df* for all comparisons was 148; * = $p < .05$, † = $p < .01$, ‡ = $p < .001$.

Factor	<i>df</i>	<i>F</i> -ratio, by Observer			
		1	2	3	4
<i>MSE</i>		0.49	0.17	0.79	0.82
Type (T)	1	0.01	0.01	0.06	0.01
Orientation (O)	1	35.69‡	0.45	1.59	10.42†
Version (V)	2	52.34‡	3.37*	15.84‡	10.71‡
T × O	1	7.88†	1.35	2.47	2.98
T × V	2	1.83	1.20	1.74	0.41
O × V	2	10.07‡	2.63	3.14*	3.11*
T × O × V	2	1.46	1.32	1.44	0.94

Table 2

Marginal measures of performance used to guide inferences regarding PI, PS, and DS, considering the relationship between the two internal (top and bottom) elements of the stimuli. These same quantities were also estimated for each combination of one of the internal elements with the external form. Note: lower case letters refer to responses, upper-case letters refer to stimuli, A refers to top elements (eyes, windows), B refers to bottom elements (mouths, doors), T as a subscript indicates Thatcherized, N as a subscript indicates normal.

Quantity	Definition and description
$P(a_T b_i A_T B_i)$	marginal probability of correctly identifying the top element as being Thatcherized (“hit”) when the bottom element is in state <i>i</i> (Thatcherized, Normal)
$P(a_T b_i A_N B_i)$	marginal probability of incorrectly identifying the top element as being Thatcherized (“false alarm”) when the bottom element is in state <i>i</i> (Thatcherized, Normal)
$d'(A, B_i)$	marginal sensitivity to the correct state of the top element when the bottom element is in state <i>i</i> (Thatcherized, Normal) $= \frac{1}{\sqrt{2}} \{z[P(a_T b_i A_T B_i)] - z[P(a_T b_i A_N B_i)]\}$
$c(A, B_i)$	marginal criterion for reporting the top element as Thatcherized when the bottom element is in state <i>i</i> (Thatcherized, Normal) $= -\frac{1}{2\sqrt{2}} \{z[P(a_T b_i A_T B_i)] + z[P(a_T b_i A_N B_i)]\}$
MRI	non-parametric equality testing for the presence of marginal response invariance, pertinent to inferences regarding PS and DS $P(a_T b_T A_T B_T) + P(a_T b_N A_T B_T) = P(a_T b_T A_T B_N) + P(a_T b_N A_T B_N)$
SI	non-parameteric equality testing for the presence of sampling independence when the top element is Thatcherized and the bottom element is in state <i>i</i> (Normal, Thatcherized), pertinent to inferences regarding PI $P(a_T b_i A_T B_i) = [P(a_T b_T A_T B_i) + P(a_T b_N A_T B_i)] \times [P(a_T b_T A_T B_i) + P(a_N b_T A_T B_i)]$

are not within-stimulus effects.

Figure 7 plots the total number of violations of PS and DS, for upright and inverted stimuli, across all pair-wise combinations of exterior form and internal elements, summed across all four observers. The diagonal line in the figure represents equality of total number of violations for upright and inverted presentation. Overall, there were more violations of PS than of DS, and the total of violations of both PS and DS were greater with inverted rather than upright stimuli. The exception was for the interior elements across levels of the exterior form. Here there were far more violations of DS when the stimuli were presented upright than when they were presented inverted. Most critically, the violations of both PS and DS were greatest for the combination of exterior form and one of the internal elements (either the eyes/windows or mouth/door). This means that the perception and judg-

ments of one of the internal elements are affected by the level of the external form more than they are by the level of the other internal element.

Documenting that there are violations of PS and DS does not, however, indicate the ways in which these two types of separability are violated. To understand this, the clearest source of evidence comes first from the marginal measures of sensitivity and response bias, and second from the changes in the marginal hit and false alarm rates. Marginal sensitivity measures (d') for the face stimuli are plotted for each of the four observers in the four panels of Figure 8; the corresponding measures for the church stimuli are plotted in the four panels of Figure 9. The diagonal line in each plot indicates equality for the sensitivity measure across the two orientations of the second feature (noted in the figure key); points above the line indicate means for which the sensitivity

Table 3

Logic relating marginal measures of performance to inferences regarding PS and DS. Note: T for the evidence indicates the equality in question held, T for the inference indicates no violation; F for the evidence indicates the equality in question did not hold, F for the inference indicates a violation; ? indicates an uncertain inference.

MRI?	Evidence		Inference	
	Equal marginal d'	Equal marginal c	PS	DS
T	T	T	T	T
T	T	F	T	F
T	F	T	F	T
T	F	F	F	F
F	T	T	T	?
F	T	F	T	F
F	F	T	F	?
F	F	F	F	?

Table 4

Summary of the marginal analyses and the non-parametric test of sampling independence (SI), for each of the four observers (Obs.). Note: MRI = test of marginal response invariance, d' = test of marginal sensitivity, c = test of marginal response bias, T = equality holds, F = equality does not hold; the number in the SI column indicates the total number of failures of SI.

Stimulus	Comparison	Obs. 1				Obs. 2				Obs. 3				Obs. 4				
		MRI	d'	c	SI	MRI	d'	c	SI	MRI	d'	c	SI	MRI	d'	c	SI	
Faces	Img × eyes, mouth invd	T	T	T	0	T	T	T	0	T	T	T	1	T	T	T	0	
	Eyes × img, mouth invd	F	F	F	1	T	F	T	0	T	F	T	0	T	F	T	0	
	Img × eyes, mouth uprt	T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
	Eyes × img, mouth uprt	F	F	F	2	T	T	T	0	T	T	T	0	F	F	F	1	
	Img × mouth, eyes invd	T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
	Mouth × img, eyes invd	F	F	T	0	F	F	T	1	T	F	T	0	T	F	T	0	
	Img × mouth, eyes uprt	T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
	Mouth × img, eyes uprt	F	T	T	0	F	F	F	2	F	F	F	1	T	F	F	2	
	Eyes × mouth, img invd	T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
	Mouth × eyes, img invd	T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
	Eyes × mouth, img uprt	T	T	T	0	T	F	T	0	F	T	F	1	T	T	T	0	
	Mouth × eyes, img uprt	T	T	T	0	F	T	T	0	T	T	T	0	T	T	T	0	
	Churches	Img × windows, door invd	T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0
		Windows × img, door invd	F	F	T	1	T	F	T	0	F	T	T	0	F	T	F	1
Img × windows, door uprt		T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
Windows × img, door uprt		T	T	T	0	T	F	F	1	T	T	T	0	F	T	T	0	
Img × door, windows invd		T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
Door × img, windows invd		F	F	F	1	T	F	T	0	T	T	T	0	T	F	F	1	
Img × door, windows uprt		T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
Door × img, windows uprt		F	T	T	0	F	F	F	1	F	F	T	1	F	F	T	2	
Windows × door, img invd		F	T	F	1	T	T	T	0	T	T	T	0	T	T	T	0	
Door × windows, img invd		T	T	T	0	T	T	T	0	T	T	T	0	T	T	T	0	
Windows × door, img uprt		T	F	F	1	F	F	F	2	F	T	T	0	T	T	T	0	
Door × windows, img uprt		T	F	T	0	T	T	T	0	F	T	T	0	T	T	T	0	

to the Thatcher manipulation (i.e., the ability to detect the inversion) was higher when the second of the two dimensions was upright rather than inverted. Points with increasing distance above the diagonal indicate an increasingly larger Thatcher illusion. For the face stimuli, across the four observers, it appears that this difference was highest for the ability to identify the orientation of the eyes across the two orientations of the external form, holding the orientation of the mouth constant (the dark circles in each panel). The highest levels of sensitivity were obtained for the orientation of the external form, and the lowest levels of sensitivity were obtained for the orientation of the mouth (the square symbols in each panel). For the church stimuli, for three of the four observers, the distance above the diagonal (indexing the magnitude of the Thatcher illusion) was greatest for the door across the two levels of the overall form, holding the orientation of the windows constant (the dark triangles in each of the panels). The highest level of sensitivity was generally obtained with the overall form, while the lowest level was obtained with the doors.

An additional note of interest is that only a minority

of the points have x -axis confidence intervals that include 0, indicating that the ability to identify the inversion (the “Thatcherizing” manipulation) when the dimension held constant was inverted was not reliably different from 0. Instead, in the majority of cases, for both the faces and churches, the ability to identify the inversion was reliably greater than 0, even when the dimension held constant was inverted rather than upright. This is consistent with the notion that inversion may be exerting quantitative (i.e., reductions in performance) rather than qualitative (e.g., a shift in the type of processing or form of representation) influences (see also Sekuler, Gaspar, Gold & Bennett, 2004; Wenger & Ingvalson, 2002, 2003; Wenger & Townsend, 2001).

The marginal response bias measures (c) for the face stimuli, for each of the four observers, are plotted in the four panels of Figure 10, with the corresponding values for the church stimuli plotted in the four panels of Figure 11. The diagonal lines in each of the panels of these figures indicate equality of response bias as a function of the orientation of the dimension held constant (indicated in parentheses in the figure key). Points above the

Table 5
Inferences regarding PI, PS, and DS for both classes of stimuli and all four observers (Obs.). Note: OT = outcome type (correspondence of the marginal analyses and the model fits), img = image, invd = inverted, uprt = upright.

Stimulus	Comparison	Obs. 1				Obs. 2				Obs. 3				Obs. 4			
		PI	PS	DS	OT	PI	PS	DS	OT	PI	PS	DS	OT	PI	PS	DS	OT
Faces	Img × eyes, mouth invd	T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
	Eyes × img, mouth invd	T	F	F	2	T	F	T	1	T	F	T	1	T	F	T	1
	Img × eyes, mouth uprt	T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
	Eyes × img, mouth uprt	T	F	T	2	T	F	F	1	T	F	F	1	T	F	F	2
	Img × mouth, eyes invd	T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
	Mouth × img, eyes invd	T	F	F	2	T	F	T	2	T	F	T	1	T	F	T	1
	Img × mouth, eyes uprt	T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
	Mouth × img, eyes uprt	T	T	T	3	T	F	F	1	T	F	T	2	T	F	F	1
	Eyes × mouth, img invd	T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
	Mouth × eyes, img invd	T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
	Eyes × mouth, img uprt	T	T	T	1	T	F	T	1	T	T	F	1	T	F	T	1
	Mouth × eyes, img uprt	T	T	T	1	T	F	T	2	T	T	T	1	T	T	T	1
	Number of violations	0	3	6		0	6	2		0	4	2		0	5	2	
	Churches	Img × windows, door invd	T	T	T	1	T	T	T	1	T	T	T	1	T	T	T
Windows × img, door invd		T	F	T	2	T	F	T	1	T	T	T	2	T	T	F	1
Img × windows, door uprt		T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
Windows × img, door uprt		T	T	T	1	T	F	F	1	T	T	T	1	T	T	T	2
Img × door, windows invd		T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
Door × img, windows invd		T	F	F	2	T	F	T	1	T	T	T	1	T	F	F	1
Img × door, windows uprt		T	T	T	1	T	T	T	1	T	T	T	1	T	T	T	1
Door × img, windows uprt		T	T	T	1	T	F	F	2	T	F	F	2	T	F	F	2
Windows × door, img invd		T	T	F	1	T	T	T	1	T	T	T	1	T	T	T	1
Door × windows, img invd		T	T	T	3	T	T	T	1	T	T	T	1	T	T	T	1
Windows × door, img uprt		T	F	F	1	T	F	T	2	T	T	T	2	T	T	T	1
Door × windows, img uprt		T	F	T	1	T	T	T	1	T	T	T	1	T	T	T	3
Number of violations		0	4	3		0	5	2		0	1	1		0	2	3	

Outcome type (OT):
 1. identical inferences supported by marginal analyses and model fit
 2. model fit resolves ambiguity from marginal analyses
 3. inferences from models and marginals differ nominally, but agree by overlapping confidence intervals on parameter estimates

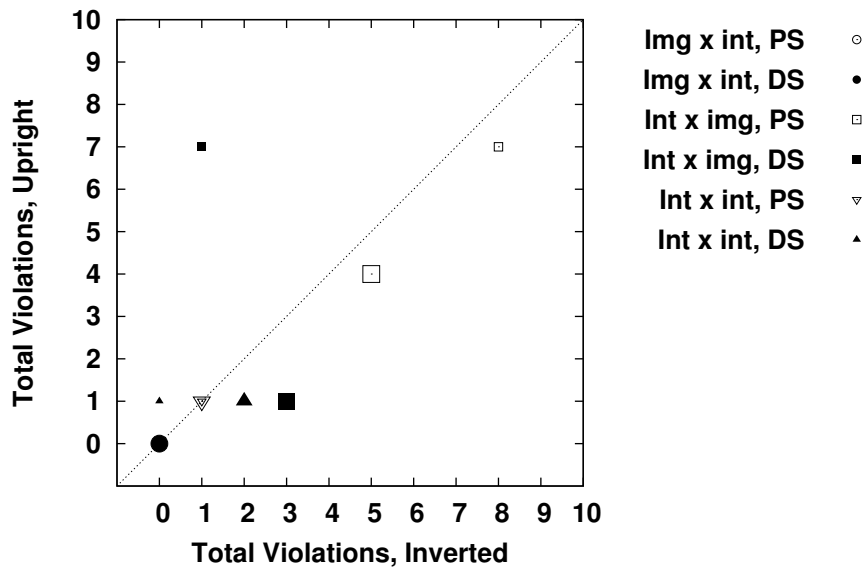


Figure 7. Total number of violations of PS and DS, for all four observers, for the faces (small symbols) and churches (large symbols), as a function of stimulus orientation. Note: Img = image (exterior form), int = interior elements (eyes/windows, mouth/door).

diagonal indicate a response bias that is relatively more liberal (i.e., more willing to judge a dimension as inverted than not) when the second of the two dimensions is upright rather than inverted; points below the diagonal indicate a response bias that is relatively more conservative when the second of the two dimensions is inverted rather than upright. For faces, the majority of the shifts in marginal response bias involved observers being relatively more conservative when the other dimensions were inverted rather than upright, meaning that observers

were much less likely to respond that a given dimension was inverted when either or both of the other dimensions were inverted, relative to when they were upright. In addition, the largest shifts in marginal criterion involved judgments on the eyes relative to the outer form, consistent with the largest number of violations of DS being associated with an internal element and the outer form. The same qualitative pattern was observed with the church stimuli, though two of the observers (1 and 2) showed a more conservative response criterion for a subset of the

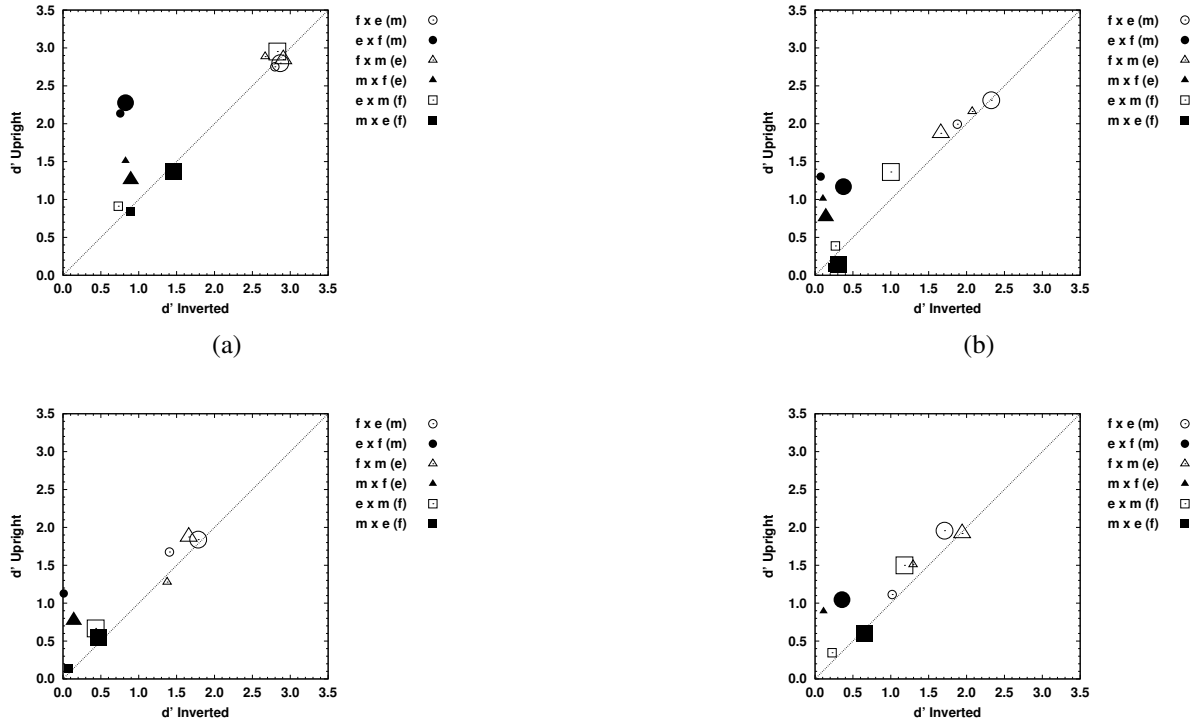


Figure 8. Marginal sensitivity (d') for faces, for all four observers. Note: small symbols plot mean sensitivity when dimension held constant (noted in parentheses) is inverted, large symbols plot mean sensitivity when the dimension held constant is upright; f = exterior form, e = eyes, m = mouth.

stimuli when the other dimension was upright rather than inverted.

The data to this point suggest that the ability to discriminate orientation (the basis of the Thatcher illusion) has both perceptual and decisional sources, for both faces and churches, based on the differences in the marginal sensitivity and bias estimates. Observers were both less sensitive to feature inversions, and less likely to give inversion judgments, when other elements of the stimulus were inverted rather than upright. These patterns were most pronounced for the relationship between the outer form and one of the internal elements, generally the eyes. A finer level of analysis is possible by examining the changes in the hit and false alarm rates that were used to compute the sensitivity and bias measures; these data are plotted in Figure 12 (faces) and Figure 13 (churches). The values of each point in the plots represent the differences in the marginal hit and false alarm rates for the first of the two dimensions (listed in the key) as a function of the orientation of the second of the two dimensions (upright minus inverted), for the third dimension at a fixed level. The small symbols plot the data for the trials on which the third dimension was upright and the large symbols plot the data for the trials on which the third dimension was inverted. The dashed line along the negative diagonal of each panel represents equality of the magnitude of change in the marginal hit and false alarm rates.

For both types of stimuli, the marginal changes for

one dimension due to inversion of a second involved both reductions in the marginal hit rate and increases in the marginal false alarm rate (a pattern sometimes referred to as a *mirror effect*, e.g., Dobbins & Kroll, 2005; Murdock, 1998; Sikstrom, 2001). For faces, the largest changes were reductions in marginal hit rates, with corresponding (though smaller) increases in marginal false alarm rates. This can be seen in the four panels of Figure 12 as the vertical clustering of the data. In addition, changing the third dimension in almost all cases caused an additional reduction in marginal hit rates for any given pair of dimensions. This can be seen by comparing symbols that are the same in form but change in size. Consequently, it appears that the effects of inversion were at the level of the ability to correctly detect the inversion (the Thatcherizing manipulation), with a smaller increase in the rate of incorrectly reporting the inversion.

For the churches, two of the observers (1 and 3) showed patterns that were very similar to those obtained for the faces. However, the other two observers (2 and 4) showed increases in marginal false alarm rates that were larger, with the data from observer 4 showing increases in marginal false alarm rates, accompanying reductions in marginal hit rates, that were in many cases equal in magnitude to the changes in marginal hit rates. In addition, two of the observers (3 and 4) actually showed improvements in the rate of correctly identifying the inversion as a function of inversion, with these improvements being specific to the two internal elements (doors

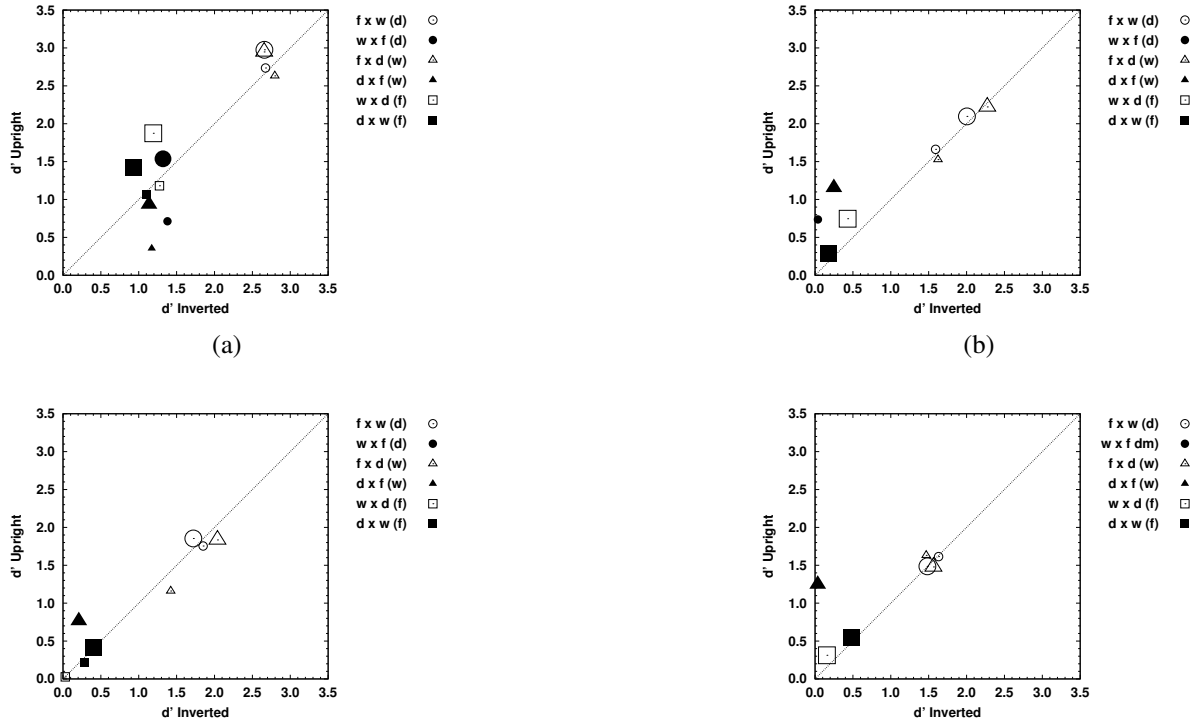


Figure 9. Marginal sensitivity (d') for churches, for all four observers. Note: small symbols plot mean sensitivity when dimension held constant (noted in parentheses) is inverted, large symbols plot mean sensitivity when the dimension held constant is upright; f = exterior form, w = windows, d = door.

and windows).

The importance of observing changes in both the hit and false alarm rates can be appreciated by examining the predictions of the simulation models in the introduction. In particular, consider the predictions for shifts in response bias made by model 1, plotted in Figure 4(b). This model implemented a violation of PS that involved increases in hit rates and decreases in false alarm rates, as was observed in our experimental data. That model predicts that this violation of PS will not produce evidence suggesting a violation of DS, as the response bias measure was (accurately) invariant. Thus, the finding of a shift in c (accompanied by changes to both the hit and false alarm rates suggests that the inferences of violations of DS are justified.

Relating Marginal Sensitivity and Bias to the Phenomenology

The results to this point support the following conclusions: (a) the same stimuli that produce the reliable relationship between rated grotesqueness and orientation that is the empirical signature of the Thatcher illusion also produce reliable violations of PS and DS; (b) analysis of those violations indicate that the perceptual and decisional components of the ability to correctly discriminate orientation result from decreases in the rate with which stimuli that have been modified are correctly identified as such (hit rates), as a function of orientation, and

from changes in the rate with which unmodified stimuli are incorrectly identified as being inverted (false alarm rates); (c) these patterns hold for both faces and churches as stimuli; and (d) there are large individual differences at all levels of the data. The one remaining question is whether the patterns of rated grotesqueness show any reliable relationships with measures of sensitivity and bias. To address this question, albeit somewhat coarsely, we estimated an overall d' and c for each stimulus of each type (face, church), for each observer. We then estimated the correlation (r) between these two overall measures and the mean (normalized) measure of grotesqueness for each stimulus, of each type, for each observer. The results of this analysis are presented in Table 6.

The correlations between grotesqueness and sensitivity were uniformly positive, for both stimulus types, and for all observers. Increases in the level of sensitivity were associated with increases in rated grotesqueness. This relationship, however, was reliable for only two of the observers (1 and 3, for both faces and churches). The correlations between grotesqueness and bias were uniformly negative, again for both stimulus types and all observers. Stimuli that were rated as highly grotesque were associated with relatively low (liberal) values of c while stimuli rated as not at all grotesque were associated with relatively high (conservative) values of c . The strength of this relationship reached the criterion for statistical reliability for only two observers (2 and 3, for both faces and churches). Finally, it is worth noting that

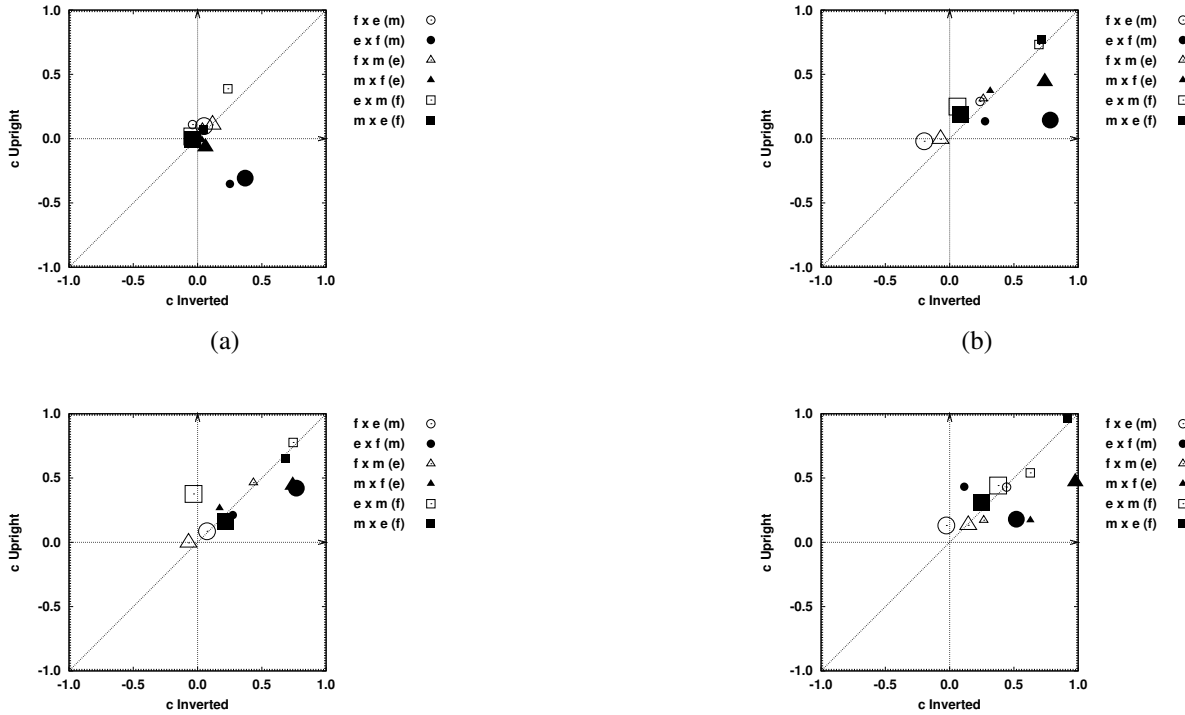


Figure 10. Marginal bias (c) for faces, for all four observers. Note: small symbols plot mean sensitivity when dimension held constant (noted in parentheses) is inverted, large symbols plot mean sensitivity when the dimension held constant is upright; f = exterior form, e = eyes, m = mouth.

Table 6

Correlations between rated grotesqueness and summary measures of sensitivity and bias. Note: $*$ = $p < .05$, \dagger = $p < .01$.

Observer	Faces		Churches	
	d'	c	d'	c
1	.71*	-.41	.75*	-.26
2	.44	-.66*	.39	-.68*
3	.62*	-.75*	.41	-.89†
4	.38	-.27	.47	-.71*

at least one of the two correlations was reliable for each observer. Consequently, the correlations reinforce the inference that the illusion has both a perceptual and decisional component, though with substantial individual differences, as suggested by the example models presented in the introduction.

General Discussion

The effect of orientation in the Thatcher illusion results from changes in perceptual sensitivity and decision criterion such that perceptual encoding is more sensitive to the presence of the image distortions and decision criterion more liberal (i.e., observers are generally more willing to indicate that they perceive the inversions) with upright than inverted faces. Importantly, the Thatcher illusion is not underpinned by different feature interactions across orientation such that there is more evidence for configularity (larger numbers of violations of inde-

pendence and separability) when faces are upright rather than inverted. Furthermore, the changes in perceptual sensitivity and decisional criteria that come when comparing upright with inverted faces are also found, although of a lesser magnitude, with churches. We interpret this as showing that the effect of orientation on perceptual and decisional processes that underpin the discrimination of orientation of internal elements and external forms in the Thatcher illusion are not face-specific. Finally, the underlying effect of orientation on perceptual sensitivity and decisional criteria that underpin the Thatcher illusion are subject to substantial individual differences, as has been true in similar studies of perceptual organization (e.g., Copeland & Wenger, 2006).

In addition, the simulations presented in the introduction demonstrate that there are multiple ways to produce the Thatcher illusion. This implies that different individuals may have different sources for the illusion: one

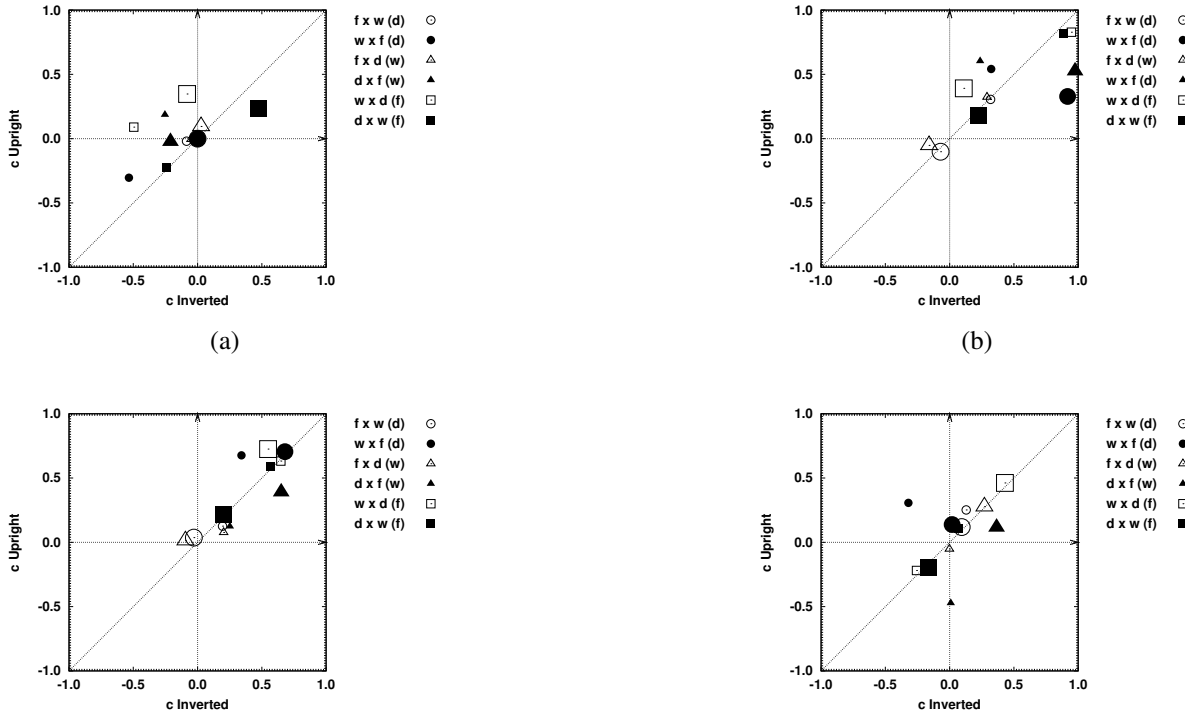


Figure 11. Marginal bias (c) for churches, for all four observers. Note: small symbols plot mean sensitivity when dimension held constant (noted in parentheses) is inverted, large symbols plot mean sensitivity when the dimension held constant is upright; f = exterior form, w = windows, d = door.

may show violations of perceptual separability while another might show violations of decisional separability. It could also be possible that a single individual observer may have different sources for the violation in different stimulus conditions (e.g., Townsend & Fific, 2004). All of this was observed in our data. However, no observer evidenced violations of PI, either alone or in combination with any other violations. Had this been the case, the ability of the GRT analysis to account for the effects would have been seriously challenged, as the strong prediction of the simulations was that violations of PI *alone* were not sufficient to produce the effect.¹²

All told, the outcomes strongly suggest that the modal explanations and interpretations of the Thatcher illusion be reassessed. In the remainder of the General Discussion we consider two questions that follow. First, in the light of the results of the GRT analysis, how should we reinterpret the Thatcher illusion? Second, what does the consistency of the results across this experiment and other related experiments mean for our view of the configural processing of faces?

Beyond Configural Processing as an Explanation of the Thatcher illusion

The experimental findings require that we reconsider what the Thatcher illusion reveals about face processing. The fact that none of the models incorporating violations of PI (alone) were capable of producing the be-

havioral pattern indicative of the Thatcher illusion, and the failure to find violations of PI in the data of the observers means that we must look beyond the formation of a strong configural state within a single percept as an explanation of the Thatcher illusion (cf., Bartlett & Searcy, 1993). Any explanation must be in terms of an underlying model and not by reference to some other phenomenology (e.g. biting versus grimacing, Parks, Coss

¹² An obvious concern here is the extent to which failure to find any violations of PI might be at all due to the small sample size associated with our using a within-observer approach. The concern is especially important as we report variation across observers. It is right to ask whether adding more participants might allow for violations of PI to be present in some participants. In order to address this question, at least in part, we used the largest value of the within-stimulus correlation parameter estimated in the model fits ($|r| = 0.0349, \hat{s}_r = .0192$) to try and estimate the sample size required to reject the null hypothesis of $\rho_0 = 0$. Assuming $\alpha = 0.05$, with desired power $1 - \beta = 0.90$, estimated sample size can be estimated as

$$n = \left[\frac{Z_{\beta(1)} + Z_{\alpha}}{Z_r} \right]^2 + 3$$

where the Z s are the Fisher transforms of the critical correlations (Zar, 1999, p. 386). This results in an estimated minimum sample size, under these assumptions, of more than 10,000 observers, suggesting some additional confidence in our conclusions. Nevertheless, it remains an empirical question to be tested.

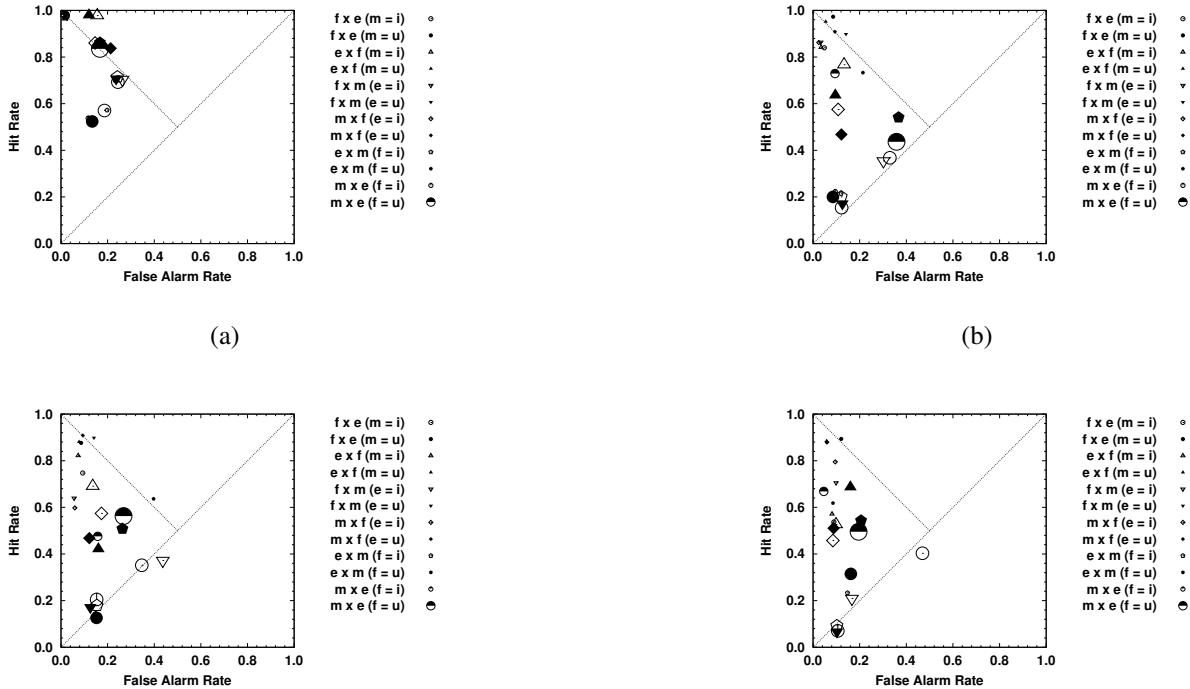


Figure 12. Changes in marginal hit and false alarm rates for faces, for the first of the two dimensions as the second of the two dimensions changed from upright to inverted. Note: small symbols plot these changes when the dimension held constant (in parentheses) was inverted, and the large symbols plot these changes when the dimension held constant was upright; f = exterior form, e = eyes, m = mouth.

& Coss, 1985). Furthermore, the explanation must allow for variation in both intensity of the phenomena and causation at the level of the individual, as the GRT analysis predicts and the empirical results show evidence of substantial differences across participants in both measures of grotesqueness and in the patterns of violations. This latter issue is one that has been critical to understanding theoretical distinctions in a range of cognitive tasks (e.g., Ashby, Maddox & Lee, 1994), beginning with Estes' (Estes, 1956) insights more than 50 years ago.

In our view, the changes in perceptual sensitivity and decisional criteria due to manipulating stimulus orientation are manifestations of visual expertise with upright but not inverted faces (note the same issues regarding expertise with upright objects are pertinent to churches as much as they are to faces). Visual expertise, in this regard, reflects only that we develop expertise for classes of images presented in specific conditions (e.g. upright and not inverted, with appropriate contrast as opposed to reversed contrast, with critical variations in judged sources of illumination, as in Talati, Rhodes & Jeffery, 2010) with which we have very significant levels of experience (Gauthier & Tarr, 1998; Mondloch, Maurer & Ahola, 2006). We thus take it that inverted *images* form a different class relative to upright images, or, more generally, images presented in their canonical orientation.

Studies of facial inversion typically show that inversion impacts more on judgments of the spatial location of

features than of feature identity itself (Rhodes, Brake & Atkinson, 1993); studies published seemingly illustrating variants on this basic point are interpreted as demonstrating a qualitative effect of orientation on face processing. However, there is a line of experimentation that provides an alternative explanation (Gaspar, Sekuler & Bennett, 2008a; Gaspar, Bennett & Sekuler, 2008b; Sekuler, Gaspar, Gold & Bennett, 2004). Using the reverse correlation method, these authors demonstrated that the classification images of upright and inverted stimuli were surprisingly similar. Both upright and inverted faces tended to be identified from information around the eyes and eyebrow region. The subtle differences that did exist in the classification images of upright and inverted faces were strongly correlated with size of the inversion effect, when computed across participants. Sekuler et al. (2004) concluded that sampling strategies accounted for the inversion effect and that there was no need to assume differences in the differential contributions of feature and relational processes across orientation and, as such, proposed that inversion had a quantitative effect on face processing. Similar conclusions have been drawn in examinations of the use of featural and configural information (Ingvalson & Wenger, 2005), short-duration memory for faces (Wenger & Ingvalson, 2002, 2003), and feature search in faces (Wenger & Townsend, 2006). We take it that these findings are consistent with the data from the present experiment and that the current findings regard-

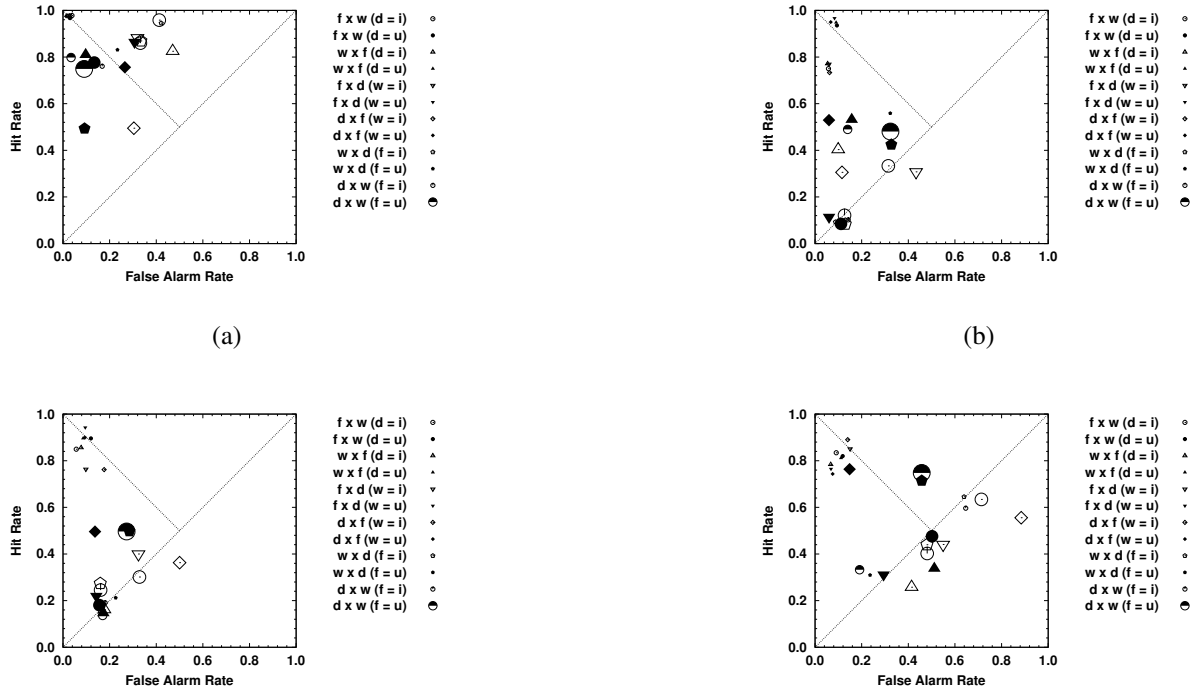


Figure 13. Changes in the marginal hit and false alarm rates for churches, for the first of the two dimensions as the second of the two dimensions changed from upright to inverted. Note: small symbols plot these changes when the dimension held constant (in parentheses) was inverted, and the large symbols plot these changes when the dimension held constant was upright; f = exterior form, w = windows, d = door.

ing perceptual sensitivity are accounted for by modest changes in sampling across upright and inverted faces.

In the absence of evidence for violations of PI (even in the perception of upright non-Thatcherized faces), and the adequacy of accounts of PS and DS to account for the inversion effect found with faces, there is an issue regarding how the phenomenology of the Thatcher illusion is generated given that it is not associated with configurality in *individual* upright faces. There are two possible answers to this question, with the adequacy of any answer depending on whether one considers the phenomenology of Thatcher faces as resulting from a purely aesthetic judgment of an atypical face relative to a face norm (cf., Langlois & Roggman, 1990), or an emotional response to an atypical faces. The studies to distinguish these two accounts have not been done, although preliminary analysis of functional imaging data indicates that Thatcher faces generate activation in a widespread network of brain areas typically activated in emotion processing (Donnelly, Cornes, Hadwin & Hadjikani, 2009). If Thatcher faces engage the mechanisms that code emotions in faces, then a range of novel questions emerges, as the mapping of faces to emotions has been considered only with respect to upright and whole faces whereby the standard patterns of facial expression are well understood (i.e. upturned mouths indicate smiling.) Of course, the Thatcher illusion is not a standard stimulus pattern and, moreover, it is likely that isolated features are capable

of being categorised with respect to their emotional valence. In the absence of an explicit model explaining the integration of the emotional states generated from different features, it is possible that the grotesqueness might result from the product of the negative valences of independent features rather than the summation of valence independent features (per a formal model of information integration, e.g., Ellison & Massaro, 1997; Massaro & Friedman, 1990) or simply the valence of the first feature to be processed. It is possible that such integration is only possible (or most efficient) with upright faces and not inverted faces.

The Effect of Orientation on Face Processing

Orientation is manifest in the present experiment in terms of two orientations only (upright and inverted) and this sets a limit on what can be concluded about how variations in orientation away from the canonical orientation for a class of images influences sensitivity and bias. When performance has been measured with intermediate orientations in other tasks, there is usually a continuous function relating performance to orientation. For example, the effect of orientation on figure-ground segmentation of merged and transparent face-pairs follows a quadratic/cubic trend (e.g., Donnelly, Hadwin, Cave & Stevenage, 2003), and 2AFC discrimination judgments of the Thatcher illusion itself follows a cubic trend (e.g.,

Lewis, 2001; Sturzel & Spillmann, 2000). The emerging consensus is that the function has quadratic and/or cubic components with a sharp decline between 90 and 115 deg. The extent to which this function is modulated by changes in sensitivity, bias or some combination is an open question. In relation to the Thatcher illusion it is, however, possible that the function measured across intermediate orientations is driven not by changes in sensitivity or bias but by one of the other factors we have discussed. For example, the function may relate to the capability to match the stimulus against stored facial prototypes to determine grotesqueness or the integration of the values for emotional valence is possible across orientations. Again, such studies have yet to be done, and the results of the present study underscore the importance of such work.

Relating Measures of Brain and Behavior

Despite the likely involvement of many brain areas in discriminating Thatcher from normal faces (Donnelly et al., 2009), with one exception (Rothstein et al., 2001), previous exploration of brain responses to Thatcher faces has been restricted to a set of event related potential (ERP) studies. These studies have largely focused on occipital-temporal brain regions and have explored the effect of Thatcher versus normal faces on the amplitude and latency of the N170 and other proposed markers of early face processing (e.g. P100, P250). For example, Boutsen, Humphreys, Praamstra and Warbrick (2006) showed evidence of delayed N170 with reduced amplitudes for Thatcher faces versus normal faces when participants performed an oddball task: these results holding for both upright and inverted stimulus presentations (although the latencies were longer for inverted than upright faces). However, other studies, requiring different judgments such as an identity or gender decision (Milivojevic, Clapp, Johnson & Corballis, 2003 and Carbon, Schweinberger, Kaufmann & Leder, 2005), have reported increased amplitude of N170 to Thatcher relative to normal faces. In addition, Milivojevic et al. also report effects of Thatcher versus normal faces on the P250 component and the P1 component. In this case, Thatcher faces led to reduced amplitude relative to normal faces on the P250 but enhanced amplitude on the P1. What is apparent is that Thatcher faces elicit different ERP responses from normal faces (in terms of both amplitude and latency), although the nature of this difference is dependent on task as well as the epoch in which the comparison is made. The question of interest here is if, and how, these findings map onto the patterns of violations described in the present study.

In part, this comparison must be informed by a recent study suggesting that face specific components to ERP are only reliably evident from the N170 onwards (Rousselet, Hush, Bennett & Sekuler, 2008: see also Rossion & Jacques, 2008). To be sure, the behavioral data from the present study do show the standard Thatcher

face inversion effect in terms of grotesqueness ratings. To that extent that this is contributed to by orientation specific perceptual changes in processing, then violations in PS presumably will be reflected in moderation of the N170. An intriguing possibility is that the violations of DS may correspond either to effects in later processing epochs (i.e. P250 measured over parieto-occipital areas: Milivojevic et al., 2005) or to responses in other epochs and over other sites yet to be considered because of the focus on strictly perceptual processing loci. However, it may also be the case that the interpretation of violations DS (or more generally shifts in response criterion) *solely* in terms of late, top-down, or conscious influences may be in error. If, alternatively, empirical shifts in decisional criteria could come about by experience-dependent changes in early visual processing (as suggested by, e.g., Wenger, Copeland, Bittner & Thomas, 2008; Wild & Busey, 2004), meaning that there could be possible empirical relationships between features of ERPs (e.g., N170) and violations of DS. These are empirical questions that have yet to be asked.

One further comment on the mapping between PI, PS, DS and ERP studies is appropriate. ERP data tend to be examined with respect to grand averages computed across participants. The approach taken by Rousselet et al. (2008) is interesting in that the intra-subject variability is considered explicitly in pursuit of effects that are deemed reliable across all subjects. In the present study, we have reported different patterns of violations across subjects. It might be that the potential to map between ERP and the patterns of violations noted in GRT studies requires a rather different approach to analysing ERP data than that typically adopted. Utilising the data cleaning methods of Rousselet et al. (2008), and mapping these data onto the patterns of violations reported in GRT analysis, provides an exciting avenue to pursue in the future

Development of Configural Processing and Individual Differences

The emerging consensus from GRT-based studies of face perception and memory is that there is little if any evidence of the “strongest” form of configurality—violations of PI (O’Toole et al., 2001)—in the behavioral regularities that have been offered as signatures of configural perception. The evidence, that the operational tasks used to demonstrate configural processing in upright faces should not be interpreted as showing within-stimulus effects, is particularly powerful as it now encompasses a set of primary tasks that have been used to argue for configural processing: same-different judgments on individual faces, whole-part differences, the interference from facial composites, and detection of the Thatcher illusion (Richler et al., 2008; Thomas, 2001a,b; Wenger & Ingvalson, 2002, 2003). To this extent the message is clear: the configural processing that is revealed in these tasks requires computations and comparisons *across* stimuli, rather than strong dependencies

within stimuli. In addition, the consistent finding of violations of DS suggests that the locus of the obtained configurality must lie as much if not more in how the perceptual evidence is used as in the intrinsic relations among the dimensions of that evidence.

An immediate implication is with respect to the understanding of the development/acquisition of the behavioral patterns that have generally been interpreted in terms of strong within-stimulus configurality (i.e., violations of PI). Many authors have been concerned with the developmental trajectory of configural processing; with accounts of processing shifts (Carey & Diamond, 1977) or more graduated development over time (e.g., Donnelly & Hadwin, 2003). The present data require that this debate be reconstructed in a more tractable form. Specifically, the present data suggest that studies should focus on exploring two primary issues: identifying the determinants and mechanisms of increases in sensitivity across development, and identifying the determinants and mechanisms of shifts in response criteria across development. Beyond these two issues there may be other questions regarding the development of facial prototypes or the capacity to integrate information regarding other sources of information (e.g., indicators of emotions), but the important points are that the questions that need to be asked are quite different from those that have typically been assumed, and that those questions can be couched in a manner that is both quite general (i.e., not tied to any one particular theory) and quite clearly defined (i.e., given formal representation using GRT).

One issue that is evident in the developmental literature, but usually subsumed under group variance in studies with adults, is the issue of individual differences. A notable exception to this statement is the study of Schwarzer (2002). Using a categorization task where categorization of individual features or sets of features allowed the attribution of processing style to individual participants on the basis of individual features or overall similarity, Schwarzer (2002) demonstrated that analytic and holistic processing styles were present in child and adult groups of participants. However, the results showed that overall similarity increased in importance through development but that analytic strategies were still present in adulthood. In the present study, violations of PS and DS were present in all participants but the pattern was not consistent across participants or conditions. We take the view that violations of PS and DS will vary across individuals, and perhaps across time with the same individuals (see in particular Townsend & Fific, 2004). Moreover, such variation is probably present in all studies but tends to be hidden by the statistics used to describe performance differences across groups. Much as in the Schwarzer study, we suggest individual differences are very likely to be found across many face processing tasks, and that differences in dominant processing mode may be a better characterization of processing across development, psychopathology, and tasks.

In conclusion, by using GRT to analyse orientation

categorisation of facial features we have shown that discrimination in upright faces is differentiated from that in inverted faces by changes in PS and DS and not PI: the effects are quantitative, not qualitative. Given that number of violations correlate with grotesqueness ratings for the Thatcher illusion, we suggest that the orientation specificity of the illusion is driven by violations of PS and DS. The phenomenology of the Thatcher illusion requires consideration of either atypicality of Thatcher faces relative to face norms or the emotion coding of facial features. The results support other studies that have reanalysed the composite face effect and whole-part differences in perception of facial features in attempts to clarify what is meant by configural processing of faces. Together these studies suggest a critical evaluation of the centrality of the concept of configural processing in the perception of and memory for faces and objects.

References

- Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute*, 50, 277–290.
- Amazeen, E. L. (1999). Perceptual independence of size and weight by dynamic touch. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 102–119.
- Ashby, F. G. & Lee, W. W. (1993). Perceptual variability as a fundamental axiom of perceptual science. In S. C. Masin (Ed.), *Foundations of perceptual theory* (pp. 369–399). Amsterdam: Elsevier.
- Ashby, F. G., Maddox, W. T. & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity choice model. *Psychological Science*, 5, 144–151.
- Ashby, F. G. & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154–179.
- Bartlett, J. C. & Searcy, J. (1993). Inversion and configuration of faces. *Cognitive Psychology*, 25, 281–316.
- Carey, S. & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science*, 195, 312–314.
- Copeland, A. M. & Wenger, M. J. (2006). An investigation of perceptual and decisional influences on the perception of hierarchical forms. *Perception*, 35, 511–529.
- Dailey, M. N. & Cottrell, G. W. (1998). Task and spatial frequency effects on facial specialization. In *Advances in Neural Information Processing Systems*, Volume 10. Cambridge, MA: MIT.
- Davidoff, J. & Donnelly, N. (1990). Object superiority - a comparison of complete and part probes. *Acta Psychologica*, 73(3), 225–243.
- Dobbins, I. G. & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6), 1186–1198.
- Donnelly, N., Cornes, K., Hadwin, J. A. & Hadjikani, N. (2009). Neural responses to a 2AFC version of the Thatcher illusion. in preparation.
- Donnelly, N. & Hadwin, J. (2003). Children's perception of the Thatcher illusion: Evidence for development in configural face processing. *Visual Cognition*, 10(8), 1001–1017.
- Donnelly, N., Hadwin, J., Cave, K. & Stevenage, S. (2003). Perceptual dominance of oriented faces mirrors the distribution of orientation tunings in inferotemporal neurons. *Cognitive Brain Research*, 17(3), 771–780.
- Ellison, J. W. & Massaro, D. W. (1997). Featural evaluation, integration, and judgment of facial affect. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 213–226.
- Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Review*, 53, 134–140.
- Farah, M. J., Wilson, K. D., Drain, M. & Tanaka, J. N. (1998). What is “special” about face perception? *Psychological Review*, 105, 482–498.
- Gaspar, C., Sekuler, A. B. & Bennett, P. J. (2008a). Spatial frequency tuning of upright and inverted face identification. *Vision Research*, 48(28), 2817–2826.
- Gaspar, C. M., Bennett, P. J. & Sekuler, A. B. (2008b). The effects of face inversion and contrast-reversal on efficiency and internal noise. *Vision Research*, 48(8), 1084–1095.
- Gauthier, I. & Tarr, M. J. (1998). Becoming a “greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, 37, 1673–1682.
- Gourevitch, V. & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, 32, 25–33.
- Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Ingvalson, E. M. & Wenger, M. J. (2005). A strong test of the dual mode hypothesis. *Perception & Psychophysics*, 67, 14–35.
- Kadlec, H. (1999). MSDA2: Updated version of software for multidimensional signal detection analyses. *Behavior Research Methods, Instruments, and Computers*, 31, 384–385.
- Kadlec, H. & Hicks, C. L. (1998). Invariance of perceptual spaces and perceptual separability of stimulus dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 80–104.
- Kadlec, H. & Townsend, J. T. (1992a). Implications of marginal and conditional detection parameters for the separabilities and independence of perceptual dimensions. *Journal of Mathematical Psychology*, 36, 325–374.
- Kadlec, H. & Townsend, J. T. (1992b). Signal detection analysis of dimensional interactions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 181–228). Hillsdale, NJ: Erlbaum.
- Langlois, J. & Roggman, L. (1990). Attractive faces are only average. *Psychological Science*, 1(2), 115–121.
- Lewis, M. (2001). The lady's not for turning: Rotation of the thatcher illusion. *Perception*, 30(6), 769–774.
- Macho, S. (2007). Feature sampling in detection: Implications for the measurement of perceptual independence. *Journal of Experimental Psychology: General*, 136, 133–153.
- Macmillan, N. A. & Creelman, C. D. (2005). *Detection theory: A user's guide (second edition)*. Mahwah, NJ: Erlbaum.
- Maddox, W. T. (1992). Perceptual and decisional separability. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 147–180). Hillsdale, NJ: Erlbaum.
- Maddox, W. T. (2001). Separating perceptual processes from decisional processes in identification and categorization. *Perception & Psychophysics*, 63, 1183–1200.
- Maddox, W. T. & Bogdanov, S. V. (2000). On the relation between decision rules and perceptual representation in multidimensional perceptual categorization. *Perception & Psychophysics*, 62, 984–997.
- Massaro, D. W. & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, 97, 225–252.
- Menner, T., Silbert, N., Cornes, K., Wenger, M. J., Townsend, J. T. & Donnelly, N. (2009a). Contrasting methods of model estimation for configural and holistic perception. Poster presented at the 2009 Vision Sciences Society Meeting, Naples FL.
- Menner, T., Wenger, M. J. & Blaha, L. M. (2009b). Multiple methods of modeling and detecting perceptual and cognitive configularity. Poster presented at the 2009 International Conference on Cognitive Modeling, Manchester UK.
- Mondloch, C. J., Maurer, D. & Ahola, S. (2006). Becoming a face expert. *Psychological Science*, 17(11), 930–934.

- Murdock, B. B. (1998). The mirror effect and attention-likelihood theory: A reflective analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 524–534.
- O'Toole, A. J., Wenger, M. J. & Townsend, J. T. (2001). Quantitative models of perceiving and remembering faces: Precedents and possibilities. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 1–38). Mahwah NJ: Erlbaum.
- Parks, T. E., Coss, R. C. & Coss, C. S. (1985). Thatcher and the cheshire cat: Context and the processing of facial features. *Perception*, *14*, 747–754.
- Rhodes, G., Brake, S. & Atkinson, A. P. (1993). What's lost in inverted faces. *Cognition*, *47*, 25–57.
- Richler, J. J., Gauthier, I., Wenger, M. J. & Palmeri, T. J. (2008). Holistic processing of faces: Perceptual and decisional components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 328–342.
- Richler, J. J., Mack, M. L., Gauthier, I. & Palmeri, T. J. (2007). Distinguishing between perceptual and decisional sources of holism in face processing. *Proceedings of the Twentieth Meeting of the Cognitive Science Society*.
- Schwarzer, G. (2002). Processing of facial and non-facial visual stimuli in 2-5-year-old children. *Infant and Child Development*, *11*(3), 253–269.
- Sekuler, A. B., Gaspar, C. M., Gold, J. M. & Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology*, *14*(5), 391–396.
- Sikstrom, S. (2001). The variance theory of the mirror effect in recognition memory. *Psychonomic Bulletin & Review*, *8*, 408–438.
- Sturzel, F. & Spillmann, L. (2000). Thatcher illusion: Dependence on angle of rotation. *Perception*, *29*(8), 937–942.
- Talati, Z., Rhodes, G. & Jeffery, L. (2010). Now you see it, now you don't: Shedding light on the thatcher illusion. *Psychological Science*, *21*(2), 219–221.
- Tanaka, J. W. & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, *46A*, 225–245.
- Tanaka, J. W. & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition*, *25*, 583–592.
- Thomas, R. D. (2001a). Characterizing perceptual interactions in face identification using multidimensional signal detection theory. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 193–228). Mahwah, NJ: Erlbaum.
- Thomas, R. D. (2001b). Perceptual interactions of facial dimensions in speeded classification and identification. *Perception & Psychophysics*, *63*, 625–650.
- Thomas, R. D. (2006). Processing time predictions of current models of perception in the classic additive factors paradigm. *Journal of Mathematical Psychology*, *50*, 441–455.
- Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, *9*, 483–484.
- Townsend, J. T. & Fific, M. (2004). Parallel versus serial processing and individual differences in high-speed search in human memory. *Perception & Psychophysics*, *66*, 953–962.
- Townsend, J. T., Hu, G. G. & Ashby, F. G. (1981). A test of visual feature sampling independence with orthogonal straight lines. *Psychological Research*, *43*, 259–275.
- Townsend, J. T., Hu, G. G. & Kadlec, H. (1988). Feature sensitivity, bias, and interdependencies as a function of intensity and payoffs. *Perception & Psychophysics*, *43*, 575–591.
- Townsend, J. T. & Thomas, R. D. (1994). Stochastic dependencies in parallel and serial models: Effects on systems factorial interactions. *Journal of Mathematical Psychology*, *38*, 1–34.
- Uttal, W. R. (1988). *On seeing forms*. Hillsdale, NJ: Erlbaum.
- Wenger, M. J., Copeland, A. M., Bittner, J. L. & Thomas, R. D. (2008). Evidence for criterion shifts in visual perceptual learning: Data and implications. *Perception & Psychophysics*, *70*, 1248–1273.
- Wenger, M. J. & Ingvalson, E. M. (2002). A decisional component of holistic encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 872–892.
- Wenger, M. J. & Ingvalson, E. M. (2003). Preserving informational separability and violating decisional separability in facial perception and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1106–1118.
- Wenger, M. J. & Townsend, J. T. (2001). Faces as gestalt stimuli: Process characteristics. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition* (pp. 229–284). Mahwah, NJ: Erlbaum.
- Wenger, M. J. & Townsend, J. T. (2006). On the costs and benefits of faces and words. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 755–779.
- Wickens, T. D. (1992). Maximum-likelihood estimation of a multivariate gaussian rating model with excluded data. *Journal of Mathematical Psychology*, *36*, 213–234.
- Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York: Oxford.
- Wild, H. A. & Busey, T. A. (2004). Seeing faces in the noise: Stochastic activity in perceptual regions of the brain may influence the perception of ambiguous stimuli. *Psychonomic Bulletin & Review*, *11*(3), 475–481.
- Young, A. W., Hellawell, D. & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*, 747–759.
- Zar, J. H. (1999). *Biostatistical analysis* (Fourth Ed.). Upper Saddle River, NJ: Prentice Hall.

Appendix A: Methods of the Numerical Simulations

The predictions of the example models considered in the introduction were obtained by numerical simulations. The parameters for all of the models considered in the introduction are presented in Table 7. Note that all simulations assumed equal marginal variances in all cases, with σ_E and $\sigma_I = 1.0$. All simulations were performed using Matlab version 7.6 running on a 1,000-node Linux cluster.

For of the N_R runs of each model, a total of n observations per stimulus were generated from the appropriate parameterized bivariate gaussian distribution. Identification/confusion matrixes were constructed by assigning each observation to a response region, using the appropriate parameterized response bounds. Marginal hit and false alarm rates were calculated from these identification/confusion matrices, along with the proportions required for the tests of Marginal Response Invariance (MRI) and Sampling Independence (SI). Values used in these two tests were transformed using an arcsin square root transform (Zar, 1999) whenever either or both values were less than or equal to 0.30, or greater than or equal to 0.70, to deal with heterogeneity of variances. Confidence intervals for the marginal sensitivity and bias measures were calculated according to Gourevitch and Galanter (1967).

Appendix B: Methods of Fitting and Evaluating the Bivariate Gaussian Models

The model-fitting and evaluation procedure used in this study follows the procedure outlined by Thomas (2001a), and applied in a previous study of the perception of hierarchical forms (Copeland & Wenger, 2006), and was done for each of the 96 pairwise dimensional comparisons involved in the marginal analyses (see Tables 4 and 5). For each of those comparison, a hierarchy of possible models was defined, always beginning with a null model (one in which PI, PS, and DS held everywhere). The hierarchy extended from this null model by systematically increasing the number of free parameters, thus representing different potential violations, including any suggested by the marginal analyses.

The specific models considered for each dimensional comparison for each observer also included two other sets of models: one set involving various violations of PI and the other instantiating a condition known as *mean shift integrality* (Ashby & Townsend, 1986; Maddox, 1992) (an interesting configuration of distributions involving a shift of means such that the marginal distances are maintained but the marginal locations are changed). Both conditions can be difficult to correctly identify in data (Ashby & Townsend, 1986; Kadlec & Townsend, 1992a,b) and so pose potential challenges to inferences, particularly with respect to PS and DS. However, and as expected from the models presented in the introduction, in none of the comparisons did a model instantiating any violations of PI provide the best account of the data. In addition, the mean shift integrality model failed in all comparisons to provide the best account of the data.

In order to limit the number of free parameters, the following assumptions and restrictions were made. First, equal (unit) variance was assumed for all models. Second, the location of one of the distributions (e.g., exterior upright, feature upright) was fixed to have marginal means of 0, 0. Third, only continuous (no violation of DS) or piece-wise (violation of DS) linear decision bounds were considered (e.g., see Figure 3(d)). Under these assumptions, for example, the null model (no violations of PI, PS, or DS) can be specified as follows. The four mean vectors would be specified as

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0 \\ \mu_{2,2} \end{bmatrix}, \mu_3 = \begin{bmatrix} \mu_{3,1} \\ 0 \end{bmatrix}, \mu_4 = \begin{bmatrix} \mu_{3,1} \\ \mu_{2,2} \end{bmatrix},$$

where the $\mu_{i,j}$ s are parameters to be estimated. Note here the equalities used to represent PS holding. Each of the four ($i = 1 \dots 4$) variance/covariance matrixes for this model would be specified as

$$\Sigma_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and the two decision bounds would be specified as two parameters to be estimated, γ_1 and γ_2 . Thus this null model requires four free parameters.

Table 7

Parameters of the simulation models used in the introduction. The values given for ρ for the PI model are the lowest and highest values used, along with the increment across that range. Note: PI = models used to examine the effects of all possible violations of PI, at all possible signs and magnitudes; the subscripts E and I refer to the external and internal dimensions, respectively; $\sigma_{i1} \dots \sigma_{i4}$ refer to the marginal variances for each of the four stimuli ($i = E, I$), with all assumed to be equal to 1.0

Parameter	Model					PI
	0	1	2	3	4	
N_R	1,000	1,000	1,000	1,000	1,000	1,000
n	250	250	250	250	250	250
μ_{E1}	0.0	0.0	0.0	0.0	0.0	0.0
μ_{I1}	0.0	0.0	0.0	0.0	0.0	0.0
μ_{E2}	0.0	0.0	0.0	0.0	0.0	0.0
μ_{I2}	2.0	2.0	2.0	2.0	2.0	2.0
μ_{E3}	2.0	2.0	2.0	2.0	2.0	2.0
μ_{I3}	0.0	-1.0	-2.0	0.0	0.0	0.0
μ_{E4}	2.0	2.0	2.0	2.0	2.0	2.0
μ_{I4}	2.0	3.0	2.0	2.0	2.0	2.0
σ_{i1}	1.0	1.0	1.0	1.0	1.0	1.0
σ_{i2}	1.0	1.0	1.0	1.0	1.0	1.0
σ_{i3}	1.0	1.0	1.0	1.0	1.0	1.0
σ_{i4}	1.0	1.0	1.0	1.0	1.0	1.0
ρ_1	0.0	0.0	0.0	0.0	0.0	[-1.00 : 0.01 : 1.00]
ρ_2	0.0	0.0	0.0	0.0	0.0	[-1.00 : 0.01 : 1.00]
ρ_3	0.0	0.0	0.0	0.0	-0.99	[-1.00 : 0.01 : 1.00]
ρ_4	0.0	0.0	0.0	0.0	0.99	[-1.00 : 0.01 : 1.00]
γ_{E1}	1.0	1.0	1.0	1.0	1.0	1.0
γ_{E2}	1.0	1.0	1.0	1.0	1.0	1.0
γ_{I1}	1.0	1.0	1.0	1.0	1.0	1.0
γ_{I2}	1.0	1.0	1.0	2.0	1.0	1.0

All models were fit to the empirical identification confusion matrixes (separately for each comparison for each observer) using methods developed by Macho (2007), extending and generalizing an approach developed by Wickens (1992), and implemented using *R*. Results of each model fit included the log likelihood for the model along with measures of fit (Akaike Information Criterion, AIC Akaike, 1983; Thomas, 2001a), adjusted for the number of free parameters. Models in a hierarchical relationship with respect to one another (i.e., in which one model generalizes another by relaxing a parameter constraint) were compared using a χ^2 statistic, with degrees of freedom equal to the difference in the number of free parameters. Models that did not exist in a hierarchical relationship to one another were compared using the AIC, the model selected as providing the superior description being the one with the lower value of AIC. A tutorial (with a complete data set) of this approach can be found in Thomas (2001a) and another example application of this approach can be found in Copeland and Wenger (2006).