Running Head: STRUCTURE OF INTEGRAL DIMENSIONS

The Structure of Integral Dimensions: Contrasting Topological and Cartesian Representations

Matt Jones
University of Colorado

Robert L. Goldstone
Indiana University

Address correspondence to:

Matt Jones
Department of Psychology and Neuroscience
University of Colorado
345 UCB
Boulder, CO 80309
Phone: 303.735.1351
Fax: 303.492.2967
mcj@colorado.edu

Abstract

Diverse evidence shows that perceptually integral dimensions, such as those composing color, are represented holistically. However, the nature of these holistic representations is poorly understood. Extant theories, such as those founded on multidimensional scaling or General Recognition Theory, model integral stimulus spaces using a Cartesian coordinate system, just as with spaces defined by separable dimensions. This approach entails a rich geometrical structure that has never been questioned but may not be psychologically meaningful for integral dimensions. In particular, Cartesian models carry a notion of orthogonality of component dimensions, such that if one dimension is diagnostic for a classification or discrimination task, another can be selected as uniquely irrelevant. This article advances an alternative model in which integral dimensions are characterized as topological spaces. The Cartesian and topological models are tested in a series of experiments using the perceptual-learning phenomenon of dimension differentiation, whereby discrimination training with integral-dimension stimuli can induce an analytic representation of those stimuli. Under the present task design, the two models make contrasting predictions regarding the analytic representation that will be learned. Results consistently support the Cartesian model. These findings indicate that perceptual representations of integral dimensions are surprisingly structured, despite their holistic, unanalyzed nature.

The Structure of Integral Dimensions: Contrasting Topological and Cartesian Representations

Much of perceptual processing can be characterized as identifying or imposing useful structure on sensory input (e.g., Schyns, Goldstone, & Thibaut, 1998). One way this is achieved is by decomposing rich, multidimensional sensations into separate psychological dimensions. For example, we readily perceive the shapes, sizes, and colors of objects around us. However, there are limits to this perceptual decomposition. A classic example is color space, which has three mathematical degrees of freedom but which can be perceptually decomposed into representations of individual dimensions only imperfectly and with cognitive effort (Garner & Felfoldy, 1970). Stimulus spaces with these properties—physically multidimensional, but psychologically difficult to analyze—are known as *integral dimensions*. A large body of research indicates that integral stimuli are primarily processed holistically, meaning as unitary percepts rather than as conjunctions of values on component dimensions (Attneave, 1950; Shepard, 1964; for a review see Kemler Nelson, 1993). However, despite a long history of research, basic aspects of the holistic representations of integral dimensions are still poorly understood. The goal of this study is to explore the structure of these representations.

There is a long history, founded in the tradition of multidimensional scaling (MDS; Torgerson, 1958) and continued in General Recognition Theory (GRT; Ashby & Townsend, 1986), of modeling multidimensional stimuli as points in a space endowed with a Cartesian coordinate system. Various cognitive tasks are taken to involve learning and decision processes within that space. This standard *Cartesian model* turns out to imply a rich geometrical structure that may not be psychologically meaningful for integral dimensions. As an alternative, we consider a *topological model* of integral

dimensions, in which similarity is defined in a local sense, but beyond that there is essentially no structure contained in the representation. This model is motivated by the claim that an integral space is best viewed as a single psychological dimension (i.e., information channel), even though mathematically it has multiple degrees of freedom (Lockhead, 1972). From this perspective, an integral stimulus space might be expected have little internal structure.

A principal difference between the Cartesian and topological models, which we explore here, concerns the relationships among component dimensions within the integral stimulus space. By *component dimension*, we mean any (mathematically) unidimensional component of the space, such as hue, saturation, or brightness within the integral space of color. According to the Cartesian model, there is a well-defined angle between any two component dimensions, and in particular there is a well-defined notion of whether two component dimensions are orthogonal (i.e., perpendicular). According to the topological model, these properties are not psychologically meaningful.

This question of orthogonality has important implications for learning with integral-dimension stimuli. Of particular interest here is the phenomenon of *dimension differentiation*, whereby observers learn to perceptually decompose an originally integral stimulus space into component dimensions. Using a training-transfer paradigm with (mathematically) two-dimensional stimulus sets, Goldstone and Steyvers (2001) showed that experimental participants learn not just the *primary dimension* that is diagnostic during training, but also a *complementary dimension* that captures the remaining variation in the stimuli. These and subsequent authors (e.g., Op de Beeck, Wagemans, & Vogels, 2003) have assumed that the complementary dimension is determined as being orthogonal to the primary dimension. This assumption seems so obvious that it has received no scrutiny, and in fact barely any explicit recognition. However, if the topological model is correct, then integral dimensions do not have the geometrical

structure necessary to determine orthogonality, and hence the complementary dimension must by determined by some other principle.

One natural alternative hypothesis is that the complementary dimension is determined as being statistically uncorrelated with the primary dimension, under the distribution of stimuli present in the task. Such a mechanism makes sense from the standpoint of information theory and has precedent in neural coding theory and vision research (Barlow & Foldiak, 1989; Simoncelli & Olshausen, 2001), because uncorrelated signals have no redundancy and hence maximize representational capacity. Thus, the question is whether the effect of dimension differentiation is to decompose integral dimensions into component dimensions that are statistically independent (*Independence hypothesis*) or that are orthogonal according to some preexisting geometry (*Orthogonal hypothesis*). The Independence hypothesis is compatible with both the topological and Cartesian models, whereas the Orthogonal hypothesis is only sensible within the Cartesian model. Therefore, evidence for the Orthogonal hypothesis would rule out the topological model and provide the first empirical support that psychological representations of integral dimensions have the geometrical structure implied by the Cartesian model.

The present experiments investigate the determinants of the complementary dimension learned in dimension differentiation, by manipulating stimulus distributions to distinguish the Orthogonal and Independence hypotheses. All three experiments support the Orthogonal hypothesis and, hence, the Cartesian model. We argue this is a surprising finding, despite the fact that it coincides with traditional modeling approaches. The geometrical structure assumed by the Cartesian model previously lacked logical or empirical support and was (we believe) implicitly assumed only because the alternative topological characterization had not been considered. Nevertheless, the Cartesian

geometry appears to be psychologically real, and hence integral dimensions have a significant amount of perceptual structure despite their holistic, unanalyzed nature.

Cartesian versus Topological Models of Integral Dimensions

With a stimulus space composed of psychologically separable dimensions, the Cartesian representation has strong logical support (see Figure 1). Because each constituent dimension has a single degree of freedom and a natural ordering, it is isomorphic to a subset of the real number line (e.g., the set of possible brightnesses can be mapped to an ordered set of numbers). Because each stimulus in the combined stimulus space can be represented by its values on the constituent dimensions (e.g., values for size and for brightness), the combined space is isomorphic to the Cartesian product of the individual dimensions. In the case of two dimensions, the result is the Cartesian plane.

--- Figure 1 about here ---

In modeling integral dimensions, it is common to assume a Cartesian coordinate system just as with separable dimensions (e.g., Ashby & Townsend, 1986; Shepard, 1962). Generalizations of the Cartesian approach that do not assume orthogonality between axes still assume the angle between the axes is an important psychological property of the representation (Carroll & Chang, 1972; Tucker, 1972). However, the logical justification for Cartesian representations breaks down for integral stimuli because their representations are not compositional. Consequently, we are left with only the left half of Figure 1, with the dotted line standing for an untested assumption.

One challenge in developing an alternative to the Cartesian model is that it is difficult to envision a continuous stimulus space, and nearly impossible to depict one graphically, without implicitly building in a geometry. Fortunately, mathematical tools exist for this type of problem, and for present purposes they are not too conceptually complex. We take as a starting point Lockhead's (1972) suggestion that integral

dimensions are best thought of as a single psychological dimension that happens to have multiple degrees of freedom (i.e., stimuli are arranged locally as in a plane or higher-dimensional space, rather than a line). Garner (1974, p. 119) expressed a similar view: "Psychologically, if dimensions are integral, they are not really perceived as dimensions at all…and do not reflect the immediate perceptual experience of the subject." This view is supported by classic findings showing that processing is usually determined by similarity in the joint space rather than similarity along component dimensions (Garner, 1974), and by evidence that manipulations affecting discriminability along one component dimension have a concomitant effect on the whole space (Goldstone, 1994b; Hinson, Cannon, & Tennison, 1998).

The natural mathematical characterization of a stimulus space that has no structure except for local similarity is that of a topological space (see, e.g., Bredon, 1995, for an introduction).[1] The starting point for defining a topological space is a set, in this case the set of all possible stimulus values. Mathematically, a set is completely unstructured, in the same sense as a nominal variable in measurement theory—each stimulus is given a label, and no relationships are assumed among different labels. The structure in a topological space comes from a *topology* defined on the set, which is a specification of all *open neighborhoods* (see Figure 2). An open neighborhood is a subset of elements in the space (i.e., a set of stimuli in the present context) that can be thought of intuitively as being similar to each other, or as constituting a local region of the space. Open neighborhoods are easy to understand in a metric space (i.e., a space with a well-defined distance between any two points). In a metric space, the open

[1] More technically, we assume the structure of a differentiable manifold, which is a topological space augmented with a notion of smoothness of paths or curves through the space. This smoothness assumption is not critical, but it simplifies the empirical analysis below.

neighborhoods are generated by all sets of the form $\{y: d(x,y) < \varepsilon\}$; that is, sets containing all elements within an arbitrary positive distance (ε) of a given element (x). Thus, the open neighborhoods of an element provide information about which other elements are arbitrarily close to it. In a topological space, there is no distance metric, so there is no well-defined notion of similarity at a large scale, but the topology (i.e., the open neighborhoods) can be thought of intuitively as conferring a notion of local similarity. (Global properties that are often of interest in topology, such as connectedness and orientability, emerge from this local similarity structure.)

--- Figure 2 about here ---

To understand better the structure that is and is not present in a topological representation, it is illustrative to ask the same question of a Cartesian representation. Consider the stimulus spaces depicted in Figures 3A and 3B. The arrangement of stimuli in these two figures is the same; only the scaling is changed. Thus the two figures could be taken as depicting exactly the same psychological representation, differing only in how the researcher chose to draw the diagram. Next, consider the orientation of the stimulus space. Under the dominant MDS model of integral dimensions, similarity between stimuli is determined by their Euclidean distance in the Cartesian space (Garner, 1974; Shepard, 1964; Torgerson, 1951). It is well understood that the Euclidean metric is unaffected by rotating the stimulus set relative to the coordinate system. This property is taken to have important psychological implications, specifically that "axes are arbitrary, and one set is as good as any other" (Ballasteros, 1989, p. 238).[2] Consequently, a diagram such as that in Figure 3C depicts exactly the

[2] Although there is evidence that certain axes of integral dimensions can be processed differently (e.g., Grau & Kemler Nelson, 1988), Kemler Nelson (1993) argues these privileged-axis effects are due to analytic representations that are secondary to the holistic (integral) representations

same psychological representation as does Figure 3A. In contrast to findings with separable dimensions (e.g., Ashby, Queller, & Berretty, 1999), rigid rotation of an integral stimulus space is a purely formal transformation, with no psychological implications.[3]

--- Figure 3 about here ---

The topological model takes this idea further, by asserting that continuous, non-rigid transformations have no psychological implications either (because they do not alter the open neighborhoods of the space). Thus, under the topological model, Figure 3D depicts exactly the same psychological representation as do Figures 3A-3C. Although the diagrams are physically different, their differences do not reflect different psychological commitments. In particular, Figure 3D illustrates how orthogonality of component dimensions is not a meaningful psychological property in the topological model. Whereas Figures 3A-3C seem to indicate an orthogonal pair of dimensions, this is an incidental property of how the diagrams are drawn (whereas in the Cartesian model

---

under investigation here. We return to the relationship between privileged axes and the present findings at the conclusion of this article.

[3] In fact, the finding that integral dimensions are best fit by a Euclidean metric in the MDS framework (e.g., Grau & Kemler Nelson, 1988; Handel & Imai, 1972; Hyman & Well, 1967, 1968; Torgerson, 1958) is consistent with the topological model. Because ordinal MDS procedures only enforce a monotonic relationship between psychological distance and observed data (e.g., similarity ratings), the metric is only determined up to monotonic transformation (e.g., Shepard, 1962). Thus, the implication of the Euclidean metric is primarily the negative conclusion that the space lacks a well-defined orientation. Because the Euclidean metric is the unique rotation-invariant metric among the family typically considered in MDS studies (the Minkowski r-metrics), the topological model predicts it to fare the best, by virtue of imposing less extraneous structure than the others.

it reflects a psychological commitment). The one constraint in the topological model is that if a transformation is not continuous, meaning it "tears" the stimulus space (and hence some of the open neighborhoods), then the psychological representation has been changed (Figure 3E). It is in this sense that local similarity is the sole meaningful form of structure in the representation.

Although the Cartesian model also has a topology (as noted above, any distance metric implies a topology), we use the term *topological model* to mean the assumption that topology is the only structure present in the representation, that is, that there is no Cartesian structure. An intuitive way to think of the contrast between the Cartesian model and the topological model is by analogy to sheet metal and rubber. A sheet of metal has a rigidity that confers a stable geometry. Any two lines have a well-defined angle of intersection, and given any one line (and a point of intersection), there is a unique other line that intersects it perpendicularly. In contrast, if one were to draw two intersecting lines on a sheet of rubber, the sheet could be stretched to make their angle of intersection take on any nonzero value. All the rubber has is a local similarity structure, in that it cannot be torn. The topological model thus constitutes a significant departure from extant models, in that it attributes far less structure to perceptual representations of integral dimensions. Thus the topological model is more parsimonious, and hence should be viewed as a viable alternative in the absence of direct evidence for the geometrical structure implied by the Cartesian model.

Design of Experiments 1 and 2

The experiments reported here test between the Cartesian and topological models of integral dimensions, by contrasting the Orthogonal and Independence hypotheses of dimension differentiation. Experiments 1 and 2 extend the dimension differentiation paradigm of Goldstone and Steyvers (2001, Expts. 3 & 4), which is illustrated in Figure

4A. The 16 points represent stimuli, arranged in a circle within a two-dimensional integral stimulus space (morphed faces or colors varying in brightness and saturation). During training, the stimuli were divided into two equal-sized categories, as indicated by the solid horizontal line. The participants' task was to learn to classify the stimuli, from corrective feedback. On each trial, a stimulus was presented, the participant responded with one of the two category labels, and then the correct response was displayed. Following training, each participant was given a transfer task, using the same stimuli but divided into new categories, as indicated by one of the two dashed lines in Figure 4A (participants were told the categories had changed). The orientation of the transfer boundary relative to the training boundary was manipulated between participants, as either 90 or 45 degrees. The critical finding was that transfer performance, defined as the proportion of correct classifications, was superior in the 90-degree condition.

--- Figure 4 about here ---

Goldstone and Steyvers concluded the superior transfer in the 90-degree condition arose because the dimension that was diagnostic at transfer was perfectly irrelevant during training. They argued that the training phase induces, at least temporarily, an analytic representation of the stimuli, composed of the primary (diagnostic) dimension and a complementary (irrelevant) dimension. When the complementary dimension becomes diagnostic at transfer, the transfer task can be directly solved using this newly learned dimension, thus facilitating transfer performance.

The question addressed by the present experiments is what determines the complementary dimension. According to the Orthogonal hypothesis, the complementary dimension is orthogonal to the diagnostic dimension in training, with respect to a psychologically intrinsic geometry of the stimulus space. According to the Independence hypothesis, the complementary dimension is uncorrelated with the diagnostic dimension, under the distribution of stimuli present in the task. With the circular stimulus distribution

used by Goldstone and Steyvers (2001), these two hypotheses make exactly the same prediction. To distinguish between them, Experiments 1 and 2 adopt an elliptical stimulus distribution, as shown in Figure 4B. This change enables us to construct separate Perpendicular and Uncorrelated conditions, in which the diagnostic dimension at transfer is either perpendicular to or uncorrelated with the diagnostic training dimension. Comparing transfer performance between these conditions provides a test between the Orthogonal and Independence hypotheses. Under the assumption that transfer performance will be greatest when the diagnostic dimension at transfer coincides with the complementary dimension learned from training, the Orthogonal hypothesis predicts superior transfer in the Perpendicular condition, whereas the Independence hypothesis predicts superior transfer in the Uncorrelated condition.

In interpreting Figure 4B, one must keep in mind that the diagnostic dimension in each task is not the boundary itself; it is the dimension that best separates the two categories. Conventionally, this dimension is treated as lying perpendicular to the boundary. For example, because the training boundary in Figure 4B is shown as horizontal, the diagnostic dimension would be referred to as the vertical dimension. This convention is technically inappropriate in the context of the topological model, because it relies on a well-defined geometry to the space, although a reader choosing to think in those terms will encounter no confusion.[4]

[4] A more rigorous definition of a component dimension is a mapping from the stimulus space to (a subset of) the real number line, giving the value of every stimulus on the dimension in question. This mapping can be identified with its isoclines, which are sets of stimuli sharing a fixed value of the dimension. The category boundary determines the diagnostic dimension because it is one such isocline. The conventional view of a dimension as running perpendicular to its isoclines identifies the dimension with its gradient, which is not defined in a topological manifold.

In the Perpendicular condition, the category boundaries in training and transfer are perpendicular, and hence so are the diagnostic dimensions. In the Uncorrelated condition, the diagnostic training and transfer dimensions are uncorrelated in the sense that, if all 24 stimuli were expressed in terms of their values on these two dimensions, those two variables would be uncorrelated across the stimulus set. One easy way to see this is to observe that the training boundary and the uncorrelated transfer boundary jointly partition the stimuli into four equally sized subsets. Therefore, knowing which side of the training boundary a stimulus lies on gives no information (even probabilistically) regarding which transfer category it belongs in.

The Cartesian model is logically consistent with both the Orthogonal and Independence hypotheses. Although the model assumes that integral stimulus spaces have meaningful geometry, this geometry would not necessarily have to play a role in dimension differentiation. However, if orthogonality between component dimensions has any psychological implications at all, it seems that it would have to contribute to determining the irrelevant dimension in tasks like the training tasks used here. Therefore the Cartesian model strongly favors the Orthogonal hypothesis.

According to the topological model, any two linearly independent dimensions are sufficient to parameterize the space, in the sense that specifying the values of any stimulus on both dimensions fully determines that stimulus. Therefore, any component dimension other than the diagnostic dimension is logically capable of serving as the complementary dimension. Assuming the dimension that is learned depends on experience (i.e., the new analytic representation is not chosen blindly), a natural expectation is that it should be driven by stimulus statistics, in particular as in the

Independence hypothesis.[5] There may be other plausible principles that are compatible with the topological model, but the Orthogonal hypothesis is not, because it relies on information that the model holds is absent from the perceptual representation.

In summary, the goal of Experiments 1 and 2 was to compare transfer performance between Perpendicular and Uncorrelated conditions. Transfer performance is taken as an index of how well (or whether) participants have developed a psychological representation of the diagnostic dimension, as a consequence of learning the training categorization. Superior transfer in the Uncorrelated condition would support the Independence hypothesis of dimension differentiation and lend support to the topological model of integral dimensions. Superior transfer in the Perpendicular condition would support the Orthogonal hypothesis and would rule out the topological model, by providing direct evidence for the intrinsic geometry entailed by the Cartesian model.

One modification to the logic of Figure 4B made in the actual experiments was that the training task was varied between subjects while the transfer task was held fixed (rather than the other way around), so that transfer performance could be directly compared across conditions. Experiments 1 and 2 achieved this control in two different ways. In Experiment 1, the training tasks in the Perpendicular and Uncorrelated conditions differed by a rotation of the stimulus set, so that the appropriate transfer boundary was the same in both conditions. In Experiment 2, the roles of training and

[5] Correlations between component dimensions are not necessarily well-defined in the topological model, because they can be altered by nonlinear transformations. However, under the smoothness assumption of the differentiable manifold (Footnote 1), we can assume nonlinear considerations are negligible as long as the range of the stimulus set is sufficiently restricted. This is a necessary assumption of the topological model for it to apply to our or Goldstone & Steyvers' (2001) studies.

transfer boundaries in Figure 4B were reversed, so that subjects were trained on different boundaries (dashed lines) and transferred to a common boundary (solid line). Details of both experiments are given below.

Experiment 1

Participants in Experiment 1 leaned to classify color patches varying in brightness and saturation. These two physical dimensions form a classic example of an integral perceptual space (Garner & Felfoldy, 1970). Every participant performed a training and a transfer task, which drew stimuli from the same region of color space but which differed in the particular stimuli presented and in how they were partitioned into categories. The category labels differed for the two tasks, and participants were instructed at transfer that they would now sort the colors in a new way.

There were six experimental conditions, illustrated in Figure 5. The conditions differed in the stimuli and the category boundaries used in training and in transfer. The conditions were grouped into three types, based on the relationship between training and transfer tasks: Perpendicular (Conditions 1 & 4), Uncorrelated (2 & 5), and Control (3 & 6). Conditions 1-3 used the same transfer task, as did Conditions 4-6. Contrasting transfer performance between conditions of different types using the same transfer task allowed two separate tests between the Orthogonal and Independence hypotheses. The predictions for these contrasts are shown in Table 1 (the Unsupervised hypothesis is introduced in the Discussion of Experiment 1).

--- Figure 5 and Table 1 about here ---

The first contrast is between the Perpendicular and Uncorrelated conditions, as described in the previous section. The diagnostic dimensions in training and transfer differed by 90 degrees in the Perpendicular conditions and 60 degrees in the Uncorrelated conditions. The Pearson correlation between these dimensions, taken

over the training stimuli, was .45 in the Perpendicular conditions and 0 in the Uncorrelated conditions. Therefore, the Orthogonal hypothesis predicts superior transfer performance in the Perpendicular conditions, and the Independence hypothesis predicts superior transfer performance in the Uncorrelated conditions. In each pair of conditions to be contrasted (1 vs. 2 and 4 vs. 5), the training tasks were isomorphic but differed by a rotation of 30 degrees, allowing the transfer tasks to be identical.

The second contrast is between the Uncorrelated and Control conditions. It provides a direct test of the Independence hypothesis, by testing whether the stimulus distribution in training can affect performance during transfer, when the diagnostic training dimension is held fixed. In each pair of conditions to be contrasted (2 vs. 3 and 5 vs. 6), the training tasks differed in stimulus distribution but had the same diagnostic dimension, and the transfer tasks were identical. The diagnostic dimensions in training and transfer differed by 90 degrees in both conditions, but their Pearson correlation (over the training stimuli) was 0 in the Uncorrelated conditions and .71 in the Control conditions. Therefore, the Orthogonal hypothesis predicts no difference in transfer performance between conditions, but the Independence hypothesis predicts superior transfer in the Uncorrelated conditions. The two Control conditions were derived by swapping the training tasks of the two Uncorrelated conditions (hence their name, because they control for any effects of training distribution). Thus, the four conditions form a 2×2 design, in which two training tasks are crossed with two transfer tasks. Under this view, the Independence hypothesis predicts an interaction between training and transfer tasks, such that transfer performance is better when the diagnostic training and transfer dimensions are uncorrelated under the training distribution.

In designing the stimulus sets, we used the coordinates of the Munsell Color System, an established standard for psychophysical scaling of color space (Newhall, Nickerson, & Judd, 1943). According to the Cartesian model, this coordinate system is

the best candidate for specifying the geometry of perceptual color space. Therefore, according to the Cartesian model, dimensions that are perpendicular in Munsell coordinates should be perceptually orthogonal. The use of these coordinates thus underpins the Cartesian model's predictions.

Method

Participants. Sixty-three undergraduates participated for course credit or $6. All participants could earn a $1 bonus in each phase of the experiment for performance above 65%. Normal color vision was verified using the color plates in Ishihara (1967).

Stimuli. Stimuli were circular color patches shown in the center of a CRT monitor on a black background. Each stimulus had a diameter of 5 cm. All stimuli were of Munsell hue 10PB (i.e., in the purple-blue region). Brightness ranged between 6.8 and 8.2, and saturation ranged between 4.2 and 7.8 (see Tables A1 & A2 for the complete set of values). Calibration of the monitor and accurate representation of the Munsell color system were achieved with a Photoresearch Spectrascan 704 Colorimeter and the relevant equations of Brainard (1989) and Travis (1991). All calculations of stimulus values and category boundaries (described next) were based on the assumption that one unit of brightness is perceptually equivalent to two units of saturation (e.g., Nickerson, 1936).

Design. Participants were randomly assigned to six conditions (*N*s = 12, 10, 10, 10, 11, & 10, respectively). The conditions differed in the stimuli and categories used for the training and transfer tasks (see Figure 5). Every task comprised 24 stimuli, forming a circle or ellipse in stimulus space and divided by a straight line into two equally sized categories. In all training tasks, the stimulus ellipse and category boundary were arranged so that the diagnostic dimension and the dimension with which it was uncorrelated differed by 60°. There were two transfer tasks, each using the same

circular stimulus set and differing in their category boundaries. Each transfer task was used in three conditions (Conditions 1-3 or 4-6).

In the Perpendicular conditions, the category boundaries in the training and transfer tasks (and hence the diagnostic dimensions) were perpendicular. In the Uncorrelated conditions, the diagnostic training and transfer dimensions were uncorrelated with respect to the distribution of training stimuli. The Control conditions were obtained by swapping the training tasks of the Uncorrelated conditions. Thus, Control Condition 3 used the transfer task from Uncorrelated Condition 2 and the training task from Uncorrelated Condition 5, and Control Condition 6 used the opposite pairing.

Procedure. Participants were instructed prior to the training task that they would learn to classify colors into two categories, labeled A and B. After training, participants were told they would see more colors "similar to the ones from before," which they would learn to classify into two new categories labeled X and Y. The mapping of categories to category labels was randomized for each participant and task. Participants were instructed at transfer, "there is NO RELATION between this task and the previous one. Knowing whether a color is A or B WILL NOT HELP YOU decide if it is X or Y." These instructions were intended to discourage simple, explicit strategies at the level of individual stimuli, such as hypothesizing that all As are Xs and all Bs are Ys. They were not expected to impede the perceptual-learning processes of dimension differentiation. The fact that all three experiments found consistent and systematic transfer differences among conditions supports the assumption that the training task affected transfer performance, despite the instructions that they were unrelated.

Each task lasted 240 trials, divided into 5 blocks of 48, separated by self-paced breaks. Every stimulus appeared exactly twice per block; otherwise presentation order was randomized. Each trial began with presentation of a stimulus in the center of the monitor. The participant indicated a category response by pressing A or B during

training and X or Y during transfer. Feedback was given below the stimulus as "Correct" (in green font) or "Wrong" (in red) followed by "That was a(n) A/B/X/Y" (in white). The stimulus and feedback remained on the screen together for 1500 ms. Trials were separated by 500 ms of blank display. The entire experiment lasted 30-50 min.

Results

Learning curves at transfer were constructed by computing the proportion correct for all subjects in each condition during each transfer block. Figure 6 displays these learning curves, and Table 2 presents mean transfer performance averaged over blocks.

--- Figure 6 and Table 2 about here ---

The difference in transfer performance between Perpendicular and Uncorrelated conditions was tested using a mixed-effects ANOVA, with condition type and transfer task as between-subjects variables and block as a within-subjects variable. This analysis revealed significant main effects of condition type ($F_{1,39} = 4.10$, $p < .05$), transfer task ($F_{1,39} = 19.63$, $p < .001$), and block ($F_{2.97,\ 115.94} = 9.18$ with Greenhouse-Geisser [GG] sphericity correction, $\varepsilon = .743$, $p < .001$). There were no significant interactions (ps > .16). Collapsing over transfer tasks, average transfer performance in the Perpendicular conditions (1 & 4) was 81.3%, compared to 74.6% in the Uncorrelated conditions (2 & 5).

The difference in transfer performance between Uncorrelated and Control conditions was tested using a mixed-effects ANOVA with training and transfer tasks as between-subjects variables and block as a within-subjects variable. This analysis revealed significant main effects of training task ($F_{1,37} = 4.16$, $p < .05$), transfer task ($F_{1,37} = 15.28$, $p < .001$), and block ($F_{3.26,\ 120.56} = 6.95$, GG $\varepsilon = .81457$, $p < .001$). No interactions involving block were significant (all Fs < 1), but, critically, there was a reliable interaction between training and transfer tasks ($F_{1,37} = 5.46$, $p < .05$). This interaction is logically equivalent to a main effect of condition type, with worse transfer

performance in the Uncorrelated conditions than the Control conditions (collapsed means: $M_{uncorrelated}$ = 74.6%, $M_{control}$ = 81.3%). This result indicates an effect of stimulus distribution on transfer, but one opposite that predicted by the Independence hypothesis.

Discussion

The results of Experiment 1 support the Orthogonal hypothesis over the Independence hypothesis. Comparison between the Perpendicular and Uncorrelated conditions shows transfer performance is better when the training and transfer dimensions are perpendicular, even if those dimensions are correlated under the distribution of training stimuli. This result suggests the complementary dimension extracted in dimension differentiation is determined by its geometrical relationship to the primary dimension, not by their statistical relationship. This finding is consistent with the Cartesian model of integral dimensions and is at odds with the topological model.

The finding that transfer performance was reliably better in the Control conditions than in the Uncorrelated conditions presents a puzzle, because it cannot be explained by either hypothesis under consideration. The Independence hypothesis predicts the opposite effect, and the Orthogonal hypothesis predicts no difference at all. Thus, dimension differentiation does not appear to offer an explanation, regardless of how integral dimensions are represented. However, there is a possible explanation rooted in unsupervised learning. The mathematics of the Control conditions is such that the major axis of the training stimulus distribution is the same as the diagnostic transfer dimension (i.e., is perpendicular to the transfer category boundary; see Figure 5). This correspondence suggests that participants learn the principal dimension of variation among the stimuli during training (regardless of the category structure), and that they can use that knowledge at transfer if this dimension is sufficiently aligned with the diagnostic transfer dimension. Thus, this *Unsupervised hypothesis* would predict the superior transfer in the Control conditions, as listed in Table 1.

Experiment 3 further explores the possibility of an unsupervised learning mechanism with integral dimensions that is complementary to the supervised mechanism of dimension differentiation. For now, we note that such a mechanism is only possible within the Cartesian model. In the Cartesian model, the notion of principle variation is well-defined because distance and hence covariance are meaningful. In the topological model, this type of information is not present in the perceptual representation. Although information about correlation may be present (see Footnote 5), information about covariance is not. In particular, a stimulus set forming a proper ellipse and one forming a circle have exactly the same perceptual characteristics under the topological model (they only appear different to the researcher because of how they are objectively parameterized). Therefore, to the extent that the transfer advantage of Control over Uncorrelated conditions found here is due to unsupervised learning, this finding also supports the Cartesian over the topological model.

Experiment 2

An important feature of Experiment 1 was that the transfer task was identical between contrasted conditions (i.e., Conditions 1-3 or 4-6). This was achieved by allowing the stimulus set in each condition to differ between training and transfer. One strength of this approach is that it demonstrates that the dimensional structure learned in training generalizes to new exemplars. However, one negative consequence is that any effects of training stimulus distribution, as predicted by the Independence hypothesis, might have been altered or diluted by the distribution of transfer stimuli. To address this possibility, Experiment 2 followed a modified design in which each participant saw the same set of stimuli at training and transfer (see Figure 7). The design still contrasted pairs of Perpendicular and Uncorrelated conditions that had matched transfer tasks. Two such pairs were tested, differing by a 90-degree rotation in stimulus space. The two

conditions composing each pair differed only in how the stimuli were divided into categories during training.

--- Figure 7 about here ---

Another possible concern with Experiment 1 is that color space is low-dimensional and densely sampled in participants' prior experience, and hence the statistics of this prior experience might overcome those of a fifteen-minute training task. To address this concern, stimuli in Experiment 2 were novel faces. Four photographs were used to generate a continuous two-dimensional space of face stimuli using Steyvers' (1999) morphing algorithm. Figure 8 shows the stimuli used in two of the conditions; the other stimulus set was drawn from the same space. To the extent that the Independence hypothesis is correct, these stimuli should provide a better opportunity for effects of the stimulus distribution to be observed.

--- Figure 8 about here ---

Together, the differences between Experiments 1 and 2 serve to make Experiment 2 a more stringent test of the Orthogonal hypothesis. The matched stimulus sets in training and transfer maximize the possibility for the stimulus distribution to influence learning (as predicted by the Independence hypothesis), as does the use of novel faces as stimuli. The matched stimulus sets between contrasted conditions also control for any possible unsupervised learning. Therefore, the Unsupervised hypothesis cannot predict any differences between conditions (see Table 1 for predictions from all three hypotheses). Superior transfer performance in the Perpendicular conditions in Experiment 2 would thus provide very strong evidence for the Orthogonal hypothesis and the Cartesian model.

Method

Participants. Twenty undergraduates participated for course credit.

Stimuli. Stimuli were images of faces, approximately 14 cm high and 12 cm wide, presented in the center of an LCD monitor on a black background. Stimuli were generated from photographs of four base faces (all bald Caucasian men) using Steyvers' (1999) morphing algorithm, which generates a stimulus image from input weights for the four base faces. For each stimulus, the weights for base faces A and B were constrained to sum to .5, as were the weights for base faces C and D. The stimuli varied in the relative weightings of A versus B and of C versus D. Images of the stimuli used in two of the experimental conditions, as well as the base faces, are displayed in Figure 8.

The four base faces were selected from a set of 104 candidates. A six-dimensional, non-metric (i.e., ordinal), Euclidean MDS solution for the 104 candidate faces was obtained using the method of Goldstone (1994a). A search was then performed to find the four faces for which the vectors A – B and C – D were as close as possible to being orthogonal and of equal length. These properties ensured proper psychophysical scaling of the stimulus set (according to the Cartesian model). For the four chosen faces, the two dimensions lie at an angle of 88.0 degrees in the MDS space, and C – D is longer than A – B by a factor of 1.14. The latter factor was accounted for in all stimulus-generation calculations (i.e., a unit along the CD dimension was equated to 1.14 units along the AB dimension).

Design. Participants were randomly assigned to four conditions (*N*s = 5, 5, 6, & 4, respectively). The stimulus values and category structures used in all conditions are displayed in Figure 7. One set of 24 stimuli was used for training and transfer in Conditions 1 and 2. During training, the stimuli were partitioned into categories differently for the two conditions, but both conditions used the same transfer categories. A second set of 24 stimuli was used for Conditions 3 and 4, which again used different training categories but had identical transfer tasks. Each of these pairs comprised one Perpendicular condition (Conditions 1 & 3) and one Uncorrelated condition (2 & 4).

Procedure. Participants were instructed prior to the training task that they would be shown faces of people who live in two fictitious towns, Bradford and Troy, and their job would be to learn which town each person lives in. After training, participants were told they would now learn to classify each person based on whether his last name is Smith or Jones. Participants were instructed that "knowing which town someone might live in WILL NOT HELP YOU decide whether they are a Smith or a Jones." The mapping of categories to category labels was randomized for each participant and task. Response keys were B and T in training, and S and J in transfer. The rest of the procedure was identical to that of Experiment 1.

Results

Figure 9 displays transfer learning curves by block, and Table 3 presents mean transfer performance in all four conditions. Collapsing over transfer tasks, transfer performance averaged 79.8% in the Perpendicular conditions and 76.1% in the Uncorrelated conditions. The reliability of this difference was confirmed by a mixed-effects ANOVA, which revealed main effects of condition type ($F_{1,16}$ = 5.10, p < .05), transfer task ($F_{1,16}$ = 34.65, p < .001), and block ($F_{2.93, 46.94}$ = 6.00, GG ε = .733, p < .001). None of the interactions was significant.

--- Figure 9 and Table 3 about here ---

Discussion

The results of Experiment 2 lend further support to the Orthogonal hypothesis over the Independence hypothesis. As in Experiment 1, participants exhibited better transfer performance when the diagnostic dimensions in training and transfer were perpendicular rather than uncorrelated. Unlike Experiment 1, the same stimulus distribution was used for training and transfer, eliminating the possibility that correlation between dimensions plays a role but only when defined with respect to the transfer distribution (e.g., that dimension differentiation somehow occurs during transfer).

Furthermore, this result was obtained using unfamiliar stimuli (faces) drawn from a vast, high-dimensional space, for which statistics from prior experience should play a minimal role. Finally, because both conditions within each contrasted pair used the same set of training stimuli (with different category structures), unsupervised learning cannot explain any differences in transfer performance.

Experiment 3

A final experiment was conducted to evaluate the unsupervised learning mechanism suggested by the results of Experiment 1. It was suggested above that the unexpected transfer advantage in the Control conditions of Experiment 1 arose because the dimension of maximal variation (i.e., the first principal component) in the training stimulus distribution coincided with the diagnostic transfer dimension. Central to this explanation is the proposal that participants engage in a form of unsupervised learning that extracts that principal component (regardless of the category structure), which in turn affects stimulus representations or attention during the transfer task. Because this Unsupervised hypothesis is incompatible with the topological model, direct evidence for this hypothesis would provide further support for the Cartesian model.

The strategy of Experiment 3 was to measure a particular rotational bias predicted by the Unsupervised hypothesis in participants' patterns of errors. In every condition, the diagnostic dimensions at training and transfer were perpendicular, but the first principle component of the training distribution was oblique to both of these (see Figure 10). The Unsupervised hypothesis predicts participants' attention will be biased away from the diagnostic dimension during transfer, in the direction of this unsupervised dimension. The experiment design contrasts pairs of conditions that differ only in the training distribution, and hence in the direction of the predicted bias. In the Clockwise conditions, the unsupervised dimension is situated clockwise from the diagnostic transfer

dimension, and hence the Unsupervised hypothesis predicts a clockwise bias in participants' errors (as elaborated below). The Counterclockwise conditions have a reverse relationship and hence lead to the opposite prediction.

--- Figure 10 about here ---

The logic of the predictions for Experiment 3 is illustrated in Figure 11 for the case of Condition 1. Figure 11A shows the training task, with lines indicating the diagnostic training dimension, the dimensions that are perpendicular to and uncorrelated with that dimension, and the unsupervised dimension. As elsewhere in this article, we indicate dimensions by boundaries or isoclines, rather than the conventional view of a dimension as lying perpendicular to its isoclines, because that convention is inappropriate in the context of the topological model. Thus, the unsupervised dimension is indicated by the minor axis of the stimulus distribution, because the first principal component is perpendicular to this boundary.

--- Figure 11 about here ---

The Unsupervised hypothesis predicts attention will shift to the unsupervised training dimension. The effect on the transfer task can be modeled as stretching the stimulus space along the attended dimension, as shown in Figure 11B (e.g., Nosofsky, 1986). Although we draw on extant theories to model the effects of selective attention, the current proposal differs markedly from previous theories in how and when selective attention operates. Existing theories assume that selective attention is feedback- or goal-driven and that it does not operate reliably with integral dimensions or in arbitrary directions in perceptual space (Garner, 1974; Nosofsky, 1992; but see Goldstone, 1994b). We return to the implications of the present proposal in the General Discussion.

The predicted perceptual representation of the transfer task, resulting from attention to the unsupervised training dimension, is shown in Figure 11C. The arrangement of stimuli under this representation leads to a prediction of asymmetric

rates of classification errors near the category boundary. In particular, the two critical border stimuli (boxed) should be misclassified more often than the other two border stimuli (circled). A similar, weaker asymmetry is predicted for stimuli further from the boundary. The prediction of asymmetric error rates can be understood in a number of different ways and is qualitatively the same for all standard theories of category representations. First, similarity to members of the opposite category is greater for the critical stimuli than for the other border stimuli. Therefore, exemplar-based models predict more errors for the critical stimuli (e.g., Nosofsky, 1986). Second, similarity to the opposite prototype (defined as the mean or centroid of all stimuli in each category) is greater for the critical stimuli. Therefore, prototype models make the same qualitative prediction (e.g., Smith & Minda, 1998). Third, the training category boundary is rotated clockwise in the attentionally altered representation of Figure 11C. Consequently, the orthogonal dimension under this representation (i.e., the predicted complementary dimension) corresponds to a decision bound that is rotated clockwise from vertical. Therefore, models that learn decision bounds with a tendency toward unidimensional rules make the same qualitative prediction as well (e.g., Ashby & Maddox, 2005).

Rather than simply comparing error rates to the border stimuli, we devised a more sensitive measure that takes into account responses to all stimuli. Specifically, a linear classifier was fit to each participant's transfer responses. The classifier is similar to classic decision-bound models of categorization (Ashby & Maddox, 1993) but is simpler and is intended merely as a data-analysis tool. The classifier is derived from a logistic regression with category response as the outcome and objective stimulus coordinates as the two predictors. The regression coefficients are then translated to an orientation, by taking the arctangent of their ratio. This process amounts to fitting each participant's responses using a two-dimensional logistic function (basically a smoothed step function) with degrees of freedom for its orientation and steepness. A participant

responding without rotational bias would produce an orientation perfectly aligned with the true category boundary, whereas the predicted asymmetry of classification errors would manifest as a deviation of the estimated orientation away from the true boundary. This deviation served as the dependent measure in Experiment 3. The Unsupervised hypothesis predicts the deviation to be clockwise in the Clockwise conditions and counterclockwise in the Counterclockwise conditions.

Notwithstanding that Experiments 1 and 2 appear to rule out the Independence hypothesis, this hypothesis also makes a prediction in Experiment 3, which is directly opposite the prediction of the Unsupervised hypothesis. In each Clockwise condition, the dimension that is uncorrelated with the diagnostic dimension at training is rotated counterclockwise relative to the transfer dimension (see Figure 11A). Therefore, following essentially the same reasoning as above, the Independence hypothesis predicts a counterclockwise bias in participants' classification errors. The reverse reasoning applies to the Counterclockwise conditions. Finally, because the training and transfer dimensions are perpendicular in all conditions (absent any modification of the perceptual space by unsupervised learning), the Orthogonal hypothesis by itself predicts no effects of training condition in this experiment. Table 1 lists the predictions of all three hypotheses. In summary, Experiment 3 provides a direct test of the Unsupervised hypothesis, as well as a contrast with the Independence hypothesis.

Method

Participants. Forty undergraduates participated for course credit.

Stimuli. Stimuli were generated as in Experiment 2, using the same four base faces. Because of differences in overall performance between the two transfer tasks of Experiment 2, the scaling factor obtained from the MDS solution (see Expt. 2 Methods) was omitted and the AB and CD dimensions were scaled equally, as a simple default assumption. Note that because the diagnostic dimensions in all tasks in Experiment 3

were aligned with one of these nominal dimensions, their relative psychological scaling does not affect whether they are orthogonal and does not affect any of the predictions.

Design. Participants were randomly assigned to four conditions (*N*s = 11, 9, 10, & 10, respectively). The stimulus values and category structures used in all conditions are displayed in Figure 10. Conditions 1 and 2 used the same transfer task, as did Conditions 3 and 4. Both transfer tasks used the same, circular set of stimuli, with diagnostic dimensions differing by 90 degrees. In all conditions, the diagnostic training dimension was perpendicular to the diagnostic transfer dimension. The two conditions associated with each transfer task differed only in their distributions of training stimuli. In each Clockwise condition (1 & 3), the major axis of the (elliptical) training distribution was 36.95 degrees clockwise from the transfer dimension (i.e., from being perpendicular to the transfer category boundary). This relationship was 36.95 degrees counterclockwise in the Counterclockwise conditions (2 & 4).

Procedure. The procedure of Experiment 3 was identical to that of Experiment 2.

Results

Transfer learning curves by block are shown in Figure 12. A mixed-effects ANOVA revealed significant effects of transfer task ($F_{1, 36}$ = 13.55, p < .001) and block ($F_{2.91, 104.86}$ = 16.71, GG ε = .72817). Neither the main effect of condition type (Clockwise vs. Counterclockwise) nor any of its interactions was significant (ps > .24), implying the manipulation of training distributions did not affect transfer performance, as expected.

--- Figure 12 about here ---

Prior to fitting the linear classifier to transfer responses, a grand-average learning curve was computed using a block size of 10 trials. Based on visual inspection, the first two points of this curve were markedly lower than the rest. Because the predictions for this experiment regard error patterns once the categories were reasonably well learned,

and because random behavior early in learning adds noise to the classifier fits, the first 20 transfer trials (out of 240) were omitted from the classifier analysis.

The linear classifier was estimated by fitting a logistic regression to the final 220 transfer responses from each participant, with the two objective stimulus dimensions (AB and CD) as predictors. The resulting regression coefficients are denoted β_{AB} and β_{CD}. The orientation of the classifier was then computed as $\theta = tan^{-1}(\beta_{AB}/\beta_{CD})$. The arctangent function was defined to take values between 0° and 180°, and 180° was added to θ if $\beta_{AB} < 0$. Under this definition, θ represents the orientation of the best-fitting linear bound separating the participant's category responses, measured in degrees clockwise from horizontal (with respect to the graphical scheme of Figure 10).

Table 4 presents the results of the classifier analysis. Mean deviation of classifier orientation for each condition is reported in terms of degrees clockwise from the optimal category boundary. Thus, the table reports $\theta - 90°$ for Conditions 1 and 2 and θ for Conditions 3 and 4. Collapsing across transfer conditions, transfer responses were biased an average of 11.2° (clockwise) in the Clockwise conditions and -13.5° (i.e., 13.5° counterclockwise) in the Counterclockwise conditions. This difference was confirmed by a 2×2 ANOVA, which revealed a main effect of condition type (Clockwise vs. Counterclockwise; $F_{1,36} = 10.44$, $p < .01$), a marginal effect of transfer task ($F_{1,36} = 4.10$, $p < .1$), and no interaction ($F_{1,36} = 1.75$, $p > .1$).

---Table 4 about here ---

Discussion

The results of Experiment 3 confirm the predictions of the Unsupervised hypothesis. In both Clockwise and Counterclockwise conditions, transfer responses were biased in the direction of the unsupervised training dimension. Regardless of how the categories are represented (by exemplars, prototypes, or decision bounds), this

effect is consistent with selective attention to that dimension. Because the experimental manipulation varied the training stimulus distribution while holding the diagnostic dimension fixed, the observed effects are due to the stimuli themselves and not the category structure (i.e., feedback), thus implicating unsupervised learning. A great deal of theoretical and empirical work has supported the proposal of supervised, feedback-driven selective attention among separable dimensions (i.e., attention to predictive or diagnostic dimensions; Jones, Maddox, & Love, 2005; Nosofsky, 1992; Sutherland & Mackintosh, 1971), but the present finding of unsupervised attention with integral dimensions is novel and not anticipated by extant category-learning models.

The results also provide further evidence against the Independence hypothesis, which predicts a pattern opposite what was observed. The Orthogonal hypothesis, taken alone, predicts no effect either way, but this is not a problem for that hypothesis because it is not in competition with the Unsupervised hypothesis. The Orthogonal and Independence hypotheses concern dimension differentiation (specifically, what determines the complementary dimension), whereas the Unsupervised hypothesis concerns a putative separate learning process, based only on the stimuli and not the category structure. One could try to save the Independence hypothesis by conjecturing that dimension differentiation contributed a bias that was opposite what was observed, but that this bias was overcome by the effect of unsupervised learning. However, Experiment 2 found no evidence for the Independence hypothesis when unsupervised learning was controlled. Taken together, the experiments support the Orthogonal over the Independence hypothesis as an explication of dimension differentiation, with unsupervised learning as an additional, separate mechanism. Because the Orthogonal and Unsupervised hypotheses are both incompatible with the topological model, the two learning mechanisms provide separate sources of support for the Cartesian model.

General Discussion

The goal of this study was to contrast the Cartesian and topological models of integral dimensions, by testing between the Orthogonal and Independence hypotheses. Experiments 1 and 2 both supported the Orthogonal hypothesis, by showing that when subjects learn to discriminate two categories of stimuli, they transfer best to a task in which the new diagnostic dimension is perpendicular to (rather than uncorrelated with) the original one. This finding in turn supports the Cartesian model, because it shows the geometrical structure that model attributes to integral perceptual spaces—in particular, angles between component dimensions—is psychologically meaningful.

The other primary finding from this study also supports the Cartesian model over the topological model. Experiments 1 and 3 found evidence for an unsupervised learning effect, whereby subjects learn or attend to the dimension of maximal variation in the stimulus distribution, regardless of the category structure. This effect was opposite the predictions of the Independence hypothesis. It is also incompatible with the topological model, because distance in perceptual space, and hence the dimension of maximal variation, are not psychologically well-defined according to that model.

These results come from using both colors and faces as stimuli. There is debate over whether faces are processed differently from other stimuli (e.g., Bukach, Gauthier, & Tarr, 2006; Grill-Spector, Knouf, & Kanwisher, 2004), but faces and colors nevertheless led to the same conclusions here. The convergence between two such different integral dimensions speaks to the generality of these conclusions.

We view the support for the Cartesian model as surprising, despite its traditional role in both MDS (Garner, 1974; Kruskal, 1964; Shepard, 1962, 1964; Torgerson, 1958) and GRT (Ashby & Lee, 1991; Ashby & Maddox, 1994; Ashby & Townsend, 1986). The strong geometrical structure assumed by the Cartesian model had received little acknowledgement or scrutiny, and it was adopted primarily because the alternative had

not been considered. Here we have shown how setting aside the assumptions of the Cartesian model leads to a model in which the representation of integral dimensions is much more primitive and unstructured, its only principle of organization being the local similarity that defines a topological space. The lack of internal structure noted by past researchers of integral dimensions (Garner, 1974; Lockhead 1972) strongly suggests something like the topological model considered here.

Nevertheless, the present results indicate that component dimensions (i.e., directions) in an integral stimulus space have well-defined angles, and distances in different directions can be meaningfully compared. Both of these properties are inconsistent with the topological model, but they are direct consequences of the core assumptions of the Cartesian model. Therefore, integral dimensions have an internal geometric structure of the type implied by Cartesian models. This metric structure is likely adaptable with sufficient experience (Goldstone, 1998; Schyns et al., 1998), but it appears to be a fundamental, if malleable, characteristic of the perceptual representation.

One possible objection is that interpretation of the present results depends on assuming the stimuli were correctly scaled (despite the careful calibration methods used in all three experiments). Indeed, the main effects of transfer task on transfer performance in all three experiments suggest that the CD face dimension was more discriminable than the AB dimension, and that discrimination between high-saturation–low-brightness and low-saturation–high-brightness was easier than between high-saturation–high-brightness and low-saturation–low-brightness (cf. Melara & Marks', 1990, finding of interaction between pitch and loudness). Likewise, selective attention might increase the salience of the diagnostic dimension in training or transfer (although this effect is known to be weak with integral dimensions), effectively stretching the perceptual space along that dimension (Nosofsky, 1986). However, neither of these

effects should be expected to alter the relationship between diagnostic training and transfer dimensions in the Perpendicular conditions, because the change in scaling would be aligned with those dimensions. Hence, the predictions of the Orthogonal hypothesis should not be altered. More critically, positing a different stimulus scaling cannot save the Independence hypothesis, because any linear transformation of stimulus coordinates cannot change the correlations between diagnostic dimensions at training and transfer. Likewise, the topological model holds that the choice of coordinate system is irrelevant (because psychologically, there is no coordinate system), so there is no way for it to predict a different result under a different choice of scaling.

Holistic and Analytic Representations

Our results are reminiscent of previous findings of privileged axes with integral dimensions (Foard & Kemler, 1984; Grau & Kemler Nelson, 1988; Melara, Marks, & Potts, 1993a), but the theoretical implications are quite different. Research has shown that classification or discrimination along certain component dimensions in integral spaces (e.g., brightness and saturation, or pitch and loudness) is easier than along other, rotated dimensions. These privileged axes are evidence for the presence of (weak) analytic representations of integral dimensions, which have been argued to be secondary to holistic representations (Kemler Nelson, 1993). Whereas this past work shows privileged axes exist, the current study can be viewed as addressing the principles guiding their acquisition. The results indicate that when new privileged axes are learned (either temporarily or permanently), they are chosen to be orthogonal with respect to an intrinsic geometry of the perceptual space. Critically, because subjects learned new, arbitrary dimensions (especially with the face stimuli), this geometrical structure must be a preexisting property of the holistic representation itself, before dimension differentiation takes place.

This conclusion also bears on the difference between analytic (separable) and holistic (integral) representations. According to the topological model, transitioning from a holistic to an analytic representation entails a radical reorganization that adds a great deal of new structure to the perceptual space. Instead, the present results indicate that much of that structure already exists; the only change is in selecting a particular orientation or set of axes. This shift might arise from a change in perceptual representation, enabling access to or explicit encoding of stimulus values on the separate dimensions (Goldstone & Steyvers, 2001). Alternatively, it may arise as a change in hypotheses regarding how concepts are distributed, as oriented randomly in stimulus space versus aligned with particular axes (Austerweil & Griffiths, 2010).

Learning of new analytic representations raises a number of open questions, having to do with the nature of the representation when dimension differentiation is partial and not permanent, such as in our experimental participants at the end of training. Does the representation lie somewhere on a continuum between integral and separable, with one (or more) axis systems partially dominant in an otherwise isotropic space; or are there parallel, competing representations, one integral and holistic and the other(s) separable and compositional (but somehow not fully established)? Likewise, can multiple sets of privileged axes exist simultaneously?[6] Research by Melara, Marks, and colleagues (Melara & Marks, 1990; Melara, Marks, & Lesko, 1992; Melara et al., 1993a;

[6] In the topological model an additional, analogous question arises in learning an individual set of privileged axes. If a diagnostic dimension remains fixed, but the stimulus distribution changes over time, are multiple complementary dimensions learned (each uncorrelated with the diagnostic dimension under a different experienced stimulus distribution), or is the distributional information somehow combined to produce a single complementary dimension? This issue does not arise in the Cartesian model (with the Orthogonal hypothesis) because a primary dimension determines a unique complementary dimension regardless of stimulus distribution.

see also Foard & Kemler Nelson, 1984) suggests that analytic representations exist independently from holistic representations and that task factors can moderate their relative influence, but the details of how these representations interact have yet to be settled (Kemler Nelson, 1993; Melara, Marks, & Potts, 1993b).

Unsupervised Learning with Integral Dimensions

This study began with the goal of testing between Cartesian and topological models by investigating the determinants of the complementary dimension learned in dimension differentiation. However, the manipulation of stimulus distributions in these experiments led to an additional, unanticipated effect, which appears to be a form of unsupervised learning, driven by the dimension of greatest variation within the stimulus set. This finding provides additional support for the Cartesian model, as explained above (because the dimension of greatest variation is not well-defined in the topological model), but it is also theoretically significant in its own right.

The unsupervised learning observed here can be viewed as a form of selective attention, but of a fundamentally different nature than the type of selective attention previously studied in category learning and related paradigms. Previous research has shown that category learning can induce a shift of attention to the diagnostic dimension (Nosofsky, 1986), affecting perceptual discrimination and generalization (Goldstone, 1994b; Jones et al., 2005). In contrast, the present effect appears to be driven by the distribution of stimuli, regardless of the category structure. Furthermore, extensive research comparing integral and separable dimensions shows that feedback- or goal-driven attention is weak with integral stimuli (in fact, this is generally taken as a defining property of integral dimensions; Garner, 1974; Shepard, 1964).

Other research on categorization has demonstrated unsupervised effects of stimulus distributions, but of a different form than found here. Pothos and Close (2008) found that subjects' preference for unidimensional versus multidimensional sorts in

spontaneous classification depends on how stimuli are clustered. Gureckis and Goldstone (2008) showed that when categories are composed of distinct clusters (separated by regions of low stimulus density), subjects subsequently show enhanced discrimination between stimuli in different clusters within the same category. This latter effect is anticipated by models of category learning that explicitly represent categories as unions of stimulus clusters (Anderson, 1991; Griffiths, Canini, Sanborn, & Navarro, 2007; Love, Medin, & Gureckis, 2004). Canini, Shashkov, and Griffiths (2010) demonstrate that transfer between categorization tasks can be improved when training and transfer categories are recombinations of a common set of clusters.

The unsupervised learning observed here appears closely related to the statistical procedure of principal components analysis (PCA). PCA works by computing the covariance matrix of some data distribution and then rank-ordering its eigenvectors according to their eigenvalues. Projecting out the lower-ranked eigenvectors produces a simpler representation of the data that can be more effective in problems of estimation and prediction (see, e.g., Joliffe, 2002). Principles related to PCA, for learning the most informative dimensions in a stimulus set, have been proposed as models for vision (Bell & Sejnowski, 1997), object recognition (Intrator & Gold, 1993), speech perception (Toscano & McMurray 2009), and lexical acquisition (Landauer & Dumais, 1997). PCA and similar algorithms have also been proposed as models for human face perception (Burton, Jenkins, Hancock, & White, 2005; Dailey, Cottrell, Padgett, & Adolphs, 2002; Furl, Phillips, & O'Toole, 2002; Turk & Pentland, 1991; Valentin, Abdi, & Edelman, 1997). The demonstration of unsupervised learning in the present study goes beyond this previous work by directly manipulating the covariance structure of the stimulus set, and by showing an effect of this manipulation in the course of a half-hour learning task (whereas previous theories have tended to focus on developmental timescales). Furthermore, the present results indicate this unsupervised learning also occurs in

colors, suggesting it is a generic principle of perceptual learning with integral dimensions, rather than being specific to face recognition.

A further question regarding this unsupervised learning mechanism is how it relates to the supervised mechanism of dimension differentiation. One possibility is that the two mechanisms operate independently, one driven by the stimulus distribution and the other by the category structure. Experiment 2 showed that transfer performance depends on the training category structure (in line with the Orthogonal hypothesis) when the stimulus distribution is held fixed, and Experiment 3 showed a corresponding effect of stimulus distribution (in line with the Unsupervised hypothesis) when the category structure is held fixed. Therefore, both sources of information play a role. However, it is also possible that supervised and unsupervised information are combined in learning a single analytic representation, which aims somehow to balance their contributions. This possibility has precedent in research on semisupervised learning, wherein areas of low stimulus density can guide learning of a category boundary (Kalish, Rogers, Lang, & Zhu, 2011; Zhu, Rogers, Qian, & Kalish, 2007), and in object segmentation following categorization, wherein the segments people learn are jointly influenced by the stimuli present and how they are divided into categories (Goldstone, 2003; Pevtzow & Goldstone, 1994). More research is needed to determine how supervised and unsupervised information interact in learning with integral dimensions.

Conclusion

The type of structure contained in a psychological representation is a subtle but fundamental question. We have shown here how standard Cartesian models of integral dimensions imply more structure than is commonly realized, and how mathematical constructs from topology allow alternative models that do not make these assumptions. Our experimental results indicate that perceptual representations of integral dimensions have a surprising amount of intrinsic structure, sufficient to determine angles between,

and to compare stimulus variation along, different component dimensions. This structure is consistent with the geometry induced by a Cartesian coordinate system. An important future question will be to investigate the sensory, developmental, or innate mechanisms that give rise to this geometrical structure.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409-429.

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, *120*, 150-172.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372-400.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*, 423-466.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-78.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*, 1178-1199.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154-79.

Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology, 63*, 516-556.

Austerweil, J. L., & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.*

Ballesteros, S. (1989). Some determinants of perceived structure: Effects of stimulus tasks. In B. E. Shepp & S. Ballesteros (Eds.), *Object perception: Structure and process* (pp. 235-266). Hillsdale, NJ: Erlbaum.

Barlow, H. B., & Foldiak, P. (1989). Adaptation and decorrelation in the cortex. In R. Durbin, C. Miall, & G. Mitchison (Eds.), *The Computing Neuron* (pp. 454-472). New York: Addison-Wesley.

Bell, A., & Sejnowski, T. (1997). The "independent components" of natural scenes are edge filters. *Vision Research, 37*, 3327-3338.

Brainard, D. H. (1989). Calibration of a computer controlled color monitor. *Color Research and Application, 14*, 23–34.

Bredon, G. E. (1995). *Topology & Geometry*. New York: Springer-Verlag.

Bukach, C., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: The power of an expertise framework. *TRENDS in Cognitive Sciences, 10*, 159-166.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*, 256-284.

Canini, K. R., Shashkov, M. M., & Griffiths, T. L. (2010). Modeling transfer learning in human categorization with the hierarchical Dirichlet process. *Proceedings of the 27th International Conference on Machine Learning.*

Carroll, J. D., & Chang, J. J. (1972, March). IDOSCAL (individual differences in orientation scaling): A generalization ofiNDSCAL allowing miosyncratic reference systems as well as analytic approximation to INDSCAL*. Paper presented at the meeting of the Psychometric Society*, *Princeton, NJ*.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that perceives and categorizes facial expressions. *Journal of Cognitive Neuroscience, 14*, 1158–1173.

Foard, C. F., & Kemler, D. G. (1984). Holistic and analytic model of processing: The multiple determinants of perceptual analysis. *Jounral of Experimental Psychology: General*, *113*, 94-111.

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26*, 797-815.

Garner, W. R. (1974). *The Processing of Information and Structur*e. New York: Wiley.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1,* 225-241.

Goldstone, R. L. (1994a). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*, 381-386.

Goldstone, R. L. (1994b). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123,* 178-200.

Goldstone, R. L. (1998). Perceptual Learning. *Annual Review of Psychology*, *49*, 585-612.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann, & C. Olson (Eds.) *Perceptual Organization in Vision: Behavioral and Neural Perspectives* (pp. 233-278). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130*, 116-139.

Grau, J. W., & Kemler Nelson, D. G. (1988). The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General, 117,* 347-370.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007) Unifying rational models of categorization via the hierarchical Dirichlet process. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

Grill-Spector, K., Knouf, N., Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience, 7*, 555–62.

Gureckis, T.M. and Goldstone, R.L.. (2008). The effect of the internal structure of categories on perception. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 843-848). Austin, TX: Cognitive Science Society.

Hinson, J. M., Cannon, C. B., & Tennison, L. R. (1998). Range effects and dimensional organization in visual discrimination. *Behavioural Processes, 43*, 275-287.

Intrator, N., & Gold, J. I. (1993). Three-dimensional object recognition using an unsupervised BCM network. *Neural Computation, 5*, 61-74.

Ishihara, S. (1967). *Tests of colour-blindness* (38 plate collection). Tokyo: Shuppan.

Jolliffe, I. T. (2002). *Principal Components Analysis*. Springer.

Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 1066-1071.

Kalish, C. W., Rogers, T. T., Lang, L., & Zhu, X.. Can semi-supervised learning explain incorrect beliefs about categories? *Cognition, 120*, 106-118.

Kemler Nelson, D. G. (1993). Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1105-1113.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1-27.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Lockhead, G. R. (1972). Processing dimensional stimuli: A note. *Psychological Review, 79,* 410-419.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

Melara, R. D., & Marks, L. E. (1990). Perceptual primacy of dimensions: Support for a model of dimensional interaction. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 398-414.

Melara, R. D., Marks, L. E., & Lesko, K. E. (1992). Optional processes in similarity judgments. *Perception and Psychophysics, 51*, 123-133.

Melara, R. D., Marks, L. E., & Potts, B. C. (1993a). Primacy of dimensions in color perception. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1082-1104.

Melara, R. D., Marks, L. E., & Potts, B. C. (1993b). Early-holistic processing or dimensional similarity? *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1114-1120.

Newhall, S. M., Nickerson, D., & Judd, D. B. (1943). Final report of the O.S.A. subcommittee on spacing of the Munsell colors. *Journal of the Optical Society of America, 33,* 385–418.

Nickerson, D. (1936). The specification of color tolerances. *Textile Research, 6,* 505–514.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39-57.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43, 25-53.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased, but not differentiated. *Journal of Experimental Psychology: General*, *132*, 491–511.

Pevtzow, R., & Goldstone, R. L. (1994). Categorization and the parsing of objects. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. (pp. 717-722). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Pothos, E. M. & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition, 107*, 581-602.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). Development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1-54.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part I. *Psychometrika, 27*, 125-140.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54-87.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193-1216.

Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology, 10,* 502-532.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1411–1436.

Steyvers, M. (1999). Morphing techniques for generating and manipulating face images. *Behavior Research Methods, Instruments, & Computers, 31,* 359-369.

Sutherland, N., & Mackintosh, N. (1971). *Mechanisms of Animal Discrimination Learning*. NY: Academic Press.

Torgerson, W. S. (1951). *A theoretical and empirical investigation of multidimensional scaling*. Ph.D. Thesis, Princeton University.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Toscano, J. C., & McMurray, B. (2009). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34*, 434-464.

Travis, D. (1991). *Effective color displays: Theory and practice*. London: Academic Press.

Tucker, L. R. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika, 37,* 3-28.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*, 71-86.

Valentin, D., Abdi, H., & Edelman, B. (1997). What represents a face: A computational approach for the integration of physiological and psychological data. *Perception, 26*, 1271-1288.

Zhu, X., Rogers, T. T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In R. C. Holte & A. Howe (Eds.), *Proceedings of AAAI 2007* (pp. 864–869). Menlo Park: AAAI Press.

Notes

Author note. Matt Jones, Department of Psychology and Neuroscience; Robert L. Goldstone, Department of Psychological and Brain Sciences. This research was supported in part by AFOSR grant FA9550-10-1-0177 to MJ and NSF REESE grant 0910218 and Department of Education IES grant R305A1100060 to RLG. Correspondence regarding this article should be addressed to Matt Jones, Department of Psychology and Neuroscience, University of Colorado, 345 UCB, Boulder, CO 80309. Email: mcj@colorado.edu.

Appendix: Stimulus Values

Stimulus values for all experiments were calculated using mathematical (trigonometric) functions for circles and ellipses. This appendix reports the generating functions, numerical values, and critical mathematical properties for each stimulus set. For all experiments and conditions, the training categories are denoted by 1 and 2, and the transfer categories by 3 and 4.

Experiment 1

Transfer stimuli for Experiment 1 are described first, because their mathematics is simpler and motivates the design of the training stimuli. All conditions used a common set of stimuli for transfer, arranged in a perfect circle with respect to the assumed scaling of the space. First, a set of abstract stimulus values were defined on the unit circle as

$$\begin{aligned} x &= \cos(\theta) \\ y &= \sin(\theta) \end{aligned} \tag{A1}$$

The parameter θ takes on 24 equally spaced values, from 7.5° through 352.5° in increments of 15°. Equation A1 thus defines 24 points, (x, y), arranged evenly around a circle. Next, the abstract values were scaled to Munsell Value (v) and Chroma (c):

$$\begin{aligned} v &= 7.5 + .7 \cdot x \\ c &= 6 + 1.4 \cdot y \end{aligned} \tag{A2}$$

Equations A1 and A2 define a circle centered on Value (brightness) 7.5 and Chroma (saturation) 6, with a radius of .7 value units or 1.4 chroma units. These two increments were assumed to be psychologically equivalent based on Nickerson's (1936) classical scaling work showing that one unit of value is perceptually equivalent to two units of chroma.

Table A1 presents the transfer stimulus values for Experiment 1, as generated by Equations A1 and A2. The table also shows how the stimuli were partitioned into categories. In Conditions 1-3, the category boundary was between stimuli corresponding to θ = 22.5° and θ = 37.5° and (at the opposite side of the circle) between θ = 202.5° and θ = 217.5°. This partition corresponds to a category boundary oriented 30° counterclockwise from the brightness axis (under the graphical arrangement of Figure 6). In Conditions 4-6, the partition was between θ = 142.5° and 157.5° and between θ = 322.5° and 337.5°, corresponding to a boundary 30° clockwise from the brightness axis.

To generate the training stimuli, abstract stimulus values were defined analogously to Equation A1, but this time in an ellipse:

$$\begin{aligned} x &= \cos(\theta) \\ y &= \tfrac{2}{\sqrt{3}}\sin(\theta) + \tfrac{1}{\sqrt{3}}\cos(\theta) \end{aligned} \qquad \text{(A3)}$$

with θ taking on the same 24 evenly spaced values as above. The ellipse defined by Equation A3 is shaped as in the training task for Conditions 2 and 6 shown in Figure 6 (except for scaling). The other training tasks were obtained by rotation and reflection. For Condition 1, the ellipse defined by Equation A3 was rotated 30° counterclockwise.

Condition 1:

$$\begin{aligned} x' &= \tfrac{\sqrt{3}}{2}x - \tfrac{1}{2}y \\ y' &= \tfrac{1}{2}x + \tfrac{\sqrt{3}}{2}y \end{aligned} \qquad \text{(A4)}$$

The coefficients $\frac{\sqrt{3}}{2}$ and $\frac{1}{2}$ are the cosine and sine of 30°, respectively. For Conditions 3 and 5, the ellipse of Equation A3 was flipped horizontally.

Conditions 3 & 5:
$$x' = -x$$
$$y' = y \tag{A5}$$

For Condition 4, the ellipse was rotated 30° counterclockwise and flipped vertically.

Condition 4:
$$x' = \tfrac{\sqrt{3}}{2} x - \tfrac{1}{2} y$$
$$y' = -\tfrac{1}{2} x - \tfrac{\sqrt{3}}{2} y \tag{A6}$$

No rotation or reflection was applied for Conditions 2 and 6.

Conditions 2 & 6:
$$x' = x$$
$$y' = y \tag{A7}$$

Lastly, the same scaling used for the transfer stimuli (Eq. A2) was applied to center each stimulus set on Value 7.5 and Chroma 6 and to equate the psychological scaling of the two dimensions.

$$v = 7.5 + .7 \cdot x'$$
$$c = 6 + 1.4 \cdot y' \tag{A8}$$

Each training task was thus defined in three steps: generation of the abstract ellipse in (x, y) coordinates (Eq. A3), rotation or reflection into (x', y') coordinates (Eqs. A4-A7), and scaling onto Value and Chroma (Eq. A8). Table A2 reports the resulting Munsell coordinates of the training stimuli.

In all conditions, the training category structure was defined so that the abstract variable x (i.e., prior to rotation or reflection) was the relevant dimension, by assigning stimuli with $x < 0$ ($90° < \theta < 270°$) to Category 1 and the remaining stimuli to Category 2. Setting $\theta = 90°$ or $270°$ in Equation A3 yields $x = 0$, which confirms that the intended

category boundary $x = 0$ perfectly bisects the border stimuli at $\theta = 82.5°$ and $97.5°$ and at $\theta = 262.5°$ and $277.5°$.

A critical design feature of the training tasks concerns the relationship between diagnostic training and transfer dimensions in the three condition types (Perpendicular, Uncorrelated, & Control). First, the diagnostic training dimension in Condition 1 is 30° counterclockwise from the brightness dimension, because of the rotation applied in Equation A4. This training dimension is parallel to the transfer category boundary for this condition, meaning it is perpendicular to the diagnostic transfer dimension. Likewise, the training dimension in Condition 4 is 30° clockwise from brightness (from Eq. A6), which is parallel to the category boundary (i.e., perpendicular to the diagnostic dimension) in the transfer task for that condition. Thus the training and transfer dimensions are perpendicular in both Conditions 1 and 4, justifying their designation as Perpendicular conditions.

Second, Equation A3 implies

$$\frac{\sqrt{3}}{2} y - \frac{1}{2} x = \sin(\theta), \tag{A9}$$

implying that $\frac{\sqrt{3}}{2} y - \frac{1}{2} x$ is uncorrelated with x (because $\sin(\theta)$ and $\cos(\theta)$ are uncorrelated). Stated differently, if the stimulus space were linearly transformed or reparameterized so that x and $\frac{\sqrt{3}}{2} y - \frac{1}{2} x$ were the coordinate dimensions, then the stimuli would form a perfect circle (of the same form as in Eq. A1). Therefore, the topological model predicts $\frac{\sqrt{3}}{2} y - \frac{1}{2} x$ as the complementary dimension learned in dimension differentiation. This dimension is rotated 30° counterclockwise from the y dimension. It coincides with the diagnostic transfer dimension (i.e., is perpendicular to the transfer category boundary) in both Conditions 2 and 5. Thus the training and

transfer dimensions in these conditions are uncorrelated in these two conditions, justifying their designation as Uncorrelated conditions.

Third, elementary calculus applied to Equation A3 shows that $x^2 + y^2$ attains its maximum at θ = 45° and 135°. These values correspond to *x*-*y* coordinates of $\pm\left(\frac{1}{\sqrt{2}}, \frac{\sqrt{3}}{\sqrt{2}}\right)$, which define the two extremal points on the abstract stimulus ellipse. The two points lie on a line through the origin (in *x*-*y* space) that is 30° clockwise from vertical. This line defines the major axis of the ellipse, and hence the principal dimension of variation of the stimuli. In Condition 6, no rotation was used in translating from (*x*, *y*) to (*v*, *c*) (see Eqs. A7 & A8), so the principal variation in the stimuli lies 30° clockwise from the saturation dimension. This direction coincides with the diagnostic transfer dimension (i.e., is perpendicular to the transfer boundary) in that condition. In Condition 3, the reflection applied by Equation A5 leads the principle dimension of variation in the training stimuli to lie 30° clockwise from the saturation dimension. Again, this direction coincides with the transfer dimension in that condition. These relationships corroborate the statement in the main text that, in Control Conditions 3 and 6, the dimension indicated by the Unsupervised hypothesis is identical to the diagnostic transfer dimension.

Experiment 2

Stimuli in Experiment 2 were defined using an ellipse equation similar to that used for the training stimuli in Experiment 1 (Eq. A3). Because the same stimulus set was used for training and transfer within each condition, the coefficients defining the abstract stimulus values *x* and *y* had to be modified slightly, so that all of the desired category boundaries would cross midway between adjacent pairs of stimuli.

$$\begin{aligned} x &= \cos(\theta) \\ y &= \sin(\theta) + \frac{1}{\sqrt{3}}\cos(\theta) \end{aligned} \tag{A10}$$

In all conditions, the category structure for the transfer task was defined so that *x* was the relevant dimension, by assigning stimuli corresponding to 90° < θ < 270° to Category 3 and the remaining stimuli to Category 4. Because θ = 90° or 270° implies *x* = 0, the category boundary defined by *x* = 0 bisects the border stimuli (at θ = 82.7° & 97.5° and θ = 262.5° & 277.5°) as desired.

In the Perpendicular conditions (1 & 3), the training categories were defined so that *y* was the relevant dimension, by assigning stimuli corresponding to 150° < θ < 330° to Category 1 and the remaining stimuli to Category 2. From Equation A10, θ = 150° or 330° implies *y* = 0, and therefore the category boundary defined by *y* = 0 bisects the border stimuli (at θ = 142.5° & 157.5° and θ = 322.5° & 337.5°), as desired. Because the abstract coordinates *x* and *y* were scaled directly onto the objective stimulus coordinates AB and CD (as described below), the training and transfer dimensions are approximately perpendicular according to the MDS fit of the 104 candidate base faces (which suggests that AB and CD are nearly perpendicular; see Experiment 2 Methods). This relationship justifies the designation of Conditions 1 and 3 as Perpendicular conditions.

Regarding the Uncorrelated conditions, Equation A10 implies

$$y - \frac{1}{\sqrt{3}} x = \sin(\theta). \qquad \text{(A11)}$$

Therefore $y - \frac{1}{\sqrt{3}} x$ is uncorrelated with the diagnostic transfer dimension *x* (because sin(θ) and cos(θ) are uncorrelated), under the stimulus distribution used for both phases of the experiment. Therefore, paralleling the argument above with Experiment 1 (see Eq. A9), the topological model predicts that if $y - \frac{1}{\sqrt{3}} x$ is the diagnostic dimension during training, then *x* will be the complementary dimension that is learned. This dimension lies

30° counterclockwise from *y*.[7] In the Uncorrelated conditions of Experiment 2 (Conditions 2 & 4), the training categories were defined so that $y - \frac{1}{\sqrt{3}}x$ was the diagnostic dimension, by assigning stimuli corresponding to $\theta < 180°$ to Category 1 and the remaining stimuli to Category 2. From Equation A11, $\theta = 0°$ or 180° corresponds to $y - \frac{1}{\sqrt{3}}x = 0$, and therefore the category boundary defined by $y - \frac{1}{\sqrt{3}}x = 0$ bisects the border stimuli at $\theta = 352.7°$ and 7.5° and at $\theta = 172.5°$ and 187.5°. In conclusion, the diagnostic training dimension in both Conditions 2 and 4 is uncorrelated with the transfer dimension, justifying the designation of these conditions as Uncorrelated.

In Conditions 1 and 2, the abstract coordinates *x* and *y* were scaled onto the objective dimensions CD and AB, respectively (where A, B, C, and D denote the four base faces used to generate the morph stimuli).

$$\text{Conditions 1 \& 2:} \quad \begin{aligned} AB &= .5 + r \cdot y \\ CD &= .5 + \frac{r}{1.1376} \cdot x \end{aligned} \tag{A12}$$

The scaling factor 1.1376 compensates for the discrepancy between the distances A – B and C – D in the MDS solution, to equate the scaling of the two objective dimensions. The joint scaling factor $r = \frac{\sqrt{3}}{4}$ serves to place all stimulus coordinates into the unit square [0, 1] × [0, 1]. Conditions 3 and 4 were rotated 90° counterclockwise from Conditions 1 and 2, by scaling *–x* to *AB* and *y* to *CD*.

---

[7] Note that $y - \frac{1}{\sqrt{3}}x$ is proportional to, and hence lies in the same direction as, the $\frac{\sqrt{3}}{2}y - \frac{1}{2}x$ dimension discussed above in relation to Experiment 1. Because the present analysis is only concerned with angles and correlations between component dimensions, their magnitudes do not matter; only their directions are important.

$$\text{Conditions 3 \& 4:}\quad \begin{aligned} AB &= .5 - r \cdot x \\ CD &= .5 + \tfrac{r}{1.1376} \cdot y \end{aligned} \tag{A13}$$

We do not report the numerical values of the stimuli for Experiment 2, both for space reasons and because the stimuli depend on the particular base faces used here (in contrast to the Munsell coordinates of Experiment 1), but they are presented graphically in Figure 8 and can be readily computed from Equations A10, A12, and A13.

Experiment 3

The same abstract ellipse from Experiment 2 (Eq. A10) was used for the training tasks of Experiment 3. In all conditions, x was defined as the diagnostic training dimension, by assigning stimuli corresponding to $90° < \theta < 270°$ to Category 1 and the remaining stimuli to Category 2 (note from Eq. A10, $\theta = 90°$ or $270°$ implies $x = 0$). This category structure was scaled onto the objective stimulus coordinates for the four conditions as follows (with $r = \frac{\sqrt{3}}{4}$ as above).

$$\text{Condition 1:}\quad \begin{aligned} AB &= .5 - r \cdot y \\ CD &= .5 + r \cdot x \end{aligned} \tag{A14}$$

$$\text{Condition 2:}\quad \begin{aligned} AB &= .5 + r \cdot y \\ CD &= .5 + r \cdot x \end{aligned} \tag{A15}$$

$$\text{Condition 3:}\quad \begin{aligned} AB &= .5 - r \cdot x \\ CD &= .5 - r \cdot y \end{aligned} \tag{A16}$$

$$\text{Condition 4:}\quad \begin{aligned} AB &= .5 - r \cdot x \\ CD &= .5 + r \cdot y \end{aligned} \tag{A17}$$

Consequently, CD was diagnostic in Conditions 1 and 2, whereas AB was diagnostic in Conditions 3 and 4.

The stimulus set for transfer in all conditions was a circle defined by

$$AB = .5 + r \cdot \cos(\theta)$$
$$CD = .5 + r \cdot \sin(\theta). \quad \text{(A18)}$$

In Conditions 1 and 2, stimuli corresponding to 90° < θ < 270° were assigned to Category 3 and the rest to Category 4. This partition defines a category boundary at *AB* = .5 (since *AB* = .5 when θ = 90° or 270°) and makes AB the relevant dimension. In Conditions 3 and 4, stimuli corresponding to θ < 180° were assigned to Category 3 and the rest to Category 4. This partition defines a category boundary at *CD* = .5 (since *CD* = .5 when θ = 0° or 180°) and makes CD the relevant dimension. Therefore, the training and transfer dimensions were approximately perpendicular in all conditions, according to the MDS solution of the 104 candidate base faces.

Although the predictions of the Unsupervised hypothesis for Experiment 3 are qualitative, it is still informative to determine the orientation of the predicted unsupervised dimension (i.e., the first principal component of the training stimulus distribution). Differentiation of $x^2 + y^2$ with respect to θ (using Eq. A10) reveals a maximum at $\theta = \tan^{-1}\left(\frac{\sqrt{13}-1}{2\sqrt{3}}\right)$. Inserting this value into Equation A10 to calculate the ratio of *y* and *x* yields

$$\frac{y}{x} = \frac{\sqrt{13}+1}{2\sqrt{3}}. \quad \text{(A19)}$$

This ratio represents the tangent of the angle between the extremal points on the ellipse and the *x* axis. The arctangent of this ratio is approximately 53.05°, meaning the predicted unsupervised dimension is 53.05° counterclockwise from the *x* dimension, or 90° – 53.05° = 36.95° clockwise from the *y* dimension. In Condition 1, this orientation

translates (via Equation A14) to 36.95° clockwise from AB, which is the diagnostic transfer dimension in that condition. Similar reasoning (using Eqs. A15-A17) concludes that the unsupervised training dimension differs by 36.95° from the diagnostic transfer dimension in all conditions, clockwise in the Clockwise conditions and counterclockwise in the Counterclockwise conditions.

Regarding the predictions of the Independence hypothesis, the same reasoning as used with Experiment 2 (see Eq. A11) implies that $y - \frac{1}{\sqrt{3}}x$ is uncorrelated with the diagnostic training dimension, *x*, under the distribution of training stimuli. As above, this dimension is oriented 30° counterclockwise from the *y* dimension. Working through the rotations and reflections of the scaling equations (A14-A17) reveals that the uncorrelated dimension (i.e., the complementary dimension predicted by the Independence hypothesis) differs by 30° from the diagnostic transfer dimension, counterclockwise in the Clockwise conditions and clockwise in the Counterclockwise conditions.

The numerical stimulus values for Experiment 3 are not reported, for the same reasons given for Experiment 2, but they are presented graphically in Figure 11 and can be readily computed from Equations A11 and A14-A18.

Tables

Table 1: Experiment predictions

Hypothesis	Experiment 1	Experiment 2	Experiment 3
Orthogonal	**P > U**	**P > U**	↻ = ↺
	U = C		
Independence	U > P	U > P	↻ < ↺
	U > C		
Unsupervised	**C > U**	U = P	**↻ > ↺**

Notes: Predictions for Experiments 1 and 2 compare transfer performance between conditions. Predictions for Experiment 3 compare rotational bias of transfer responses between conditions. Correct predictions are marked by boldface. P = Perpendicular. U = Uncorrelated. C = Control. ↻ = Clockwise. ↺ = Counterclockwise.

Table 2: Mean transfer performance in Experiment 1

Condition	Type	Performance
1	Perpendicular	85.6%
2	Uncorrelated	82.9
3	Control	83.9
4	Perpendicular	76.0
5	Uncorrelated	67.0
6	Control	78.8

Table 3: Mean transfer performance in Experiment 2

Condition	Type	Performance
1	Perpendicular	85.3%
2	Uncorrelated	83.3
3	Perpendicular	75.2
4	Uncorrelated	67.3

Table 4: Mean rotational bias in Experiment 3, clockwise from optimal boundary

Condition	Type	Deviation
1	Clockwise	8.3 °
2	Counterclockwise	-27.7
3	Clockwise	14.3
4	Counterclockwise	-0.8

Table A1: Transfer items in Experiment 1

	Munsell Coordinates		Category	
θ	Value	Chroma	Conds 1-3	Conds 4-6
7.5	8.194	6.183	4	3
22.5	8.147	6.536	4	3
37.5	8.055	6.852	3	3
52.5	7.926	7.111	3	3
67.5	7.768	7.293	3	3
82.5	7.591	7.388	3	3
97.5	7.409	7.388	3	3
112.5	7.232	7.293	3	3
127.5	7.074	7.111	3	3
142.5	6.945	6.852	3	3
157.5	6.853	6.536	3	4
172.5	6.806	6.183	3	4
187.5	6.806	5.817	3	4
202.5	6.853	5.464	3	4
217.5	6.945	5.148	4	4
232.5	7.074	4.889	4	4
247.5	7.232	4.707	4	4
262.5	7.409	4.612	4	4
277.5	7.591	4.612	4	4
292.5	7.768	4.707	4	4
307.5	7.926	4.889	4	4
322.5	8.055	5.148	4	4
337.5	8.147	5.464	4	3
352.5	8.194	5.817	4	3

Table A2: Training items in Experiment 1

		Cond 1		Conds 2 & 6		Conds 3 & 5		Cond 4	
θ	Category	Value	Chroma	Value	Chroma	Value	Chroma	Value	Chroma
7.5	2	7.848	7.571	8.194	7.012	6.806	7.012	7.848	4.429
22.5	2	7.719	7.829	8.147	7.365	6.853	7.365	7.719	4.171
37.5	2	7.575	7.963	8.055	7.625	6.945	7.625	7.575	4.037
52.5	2	7.425	7.963	7.926	7.775	7.074	7.775	7.425	4.037
67.5	2	7.281	7.829	7.768	7.803	7.232	7.803	7.281	4.171
82.5	2	7.152	7.571	7.591	7.708	7.409	7.708	7.152	4.429
97.5	1	7.047	7.205	7.409	7.497	7.591	7.497	7.047	4.795
112.5	1	6.972	6.758	7.232	7.184	7.768	7.184	6.972	5.242
127.5	1	6.933	6.258	7.074	6.790	7.926	6.790	6.933	5.742
142.5	1	6.933	5.742	6.945	6.343	8.055	6.343	6.933	6.258
157.5	1	6.972	5.242	6.853	5.872	8.147	5.872	6.972	6.758
172.5	1	7.047	4.795	6.806	5.410	8.194	5.410	7.047	7.205
187.5	1	7.152	4.429	6.806	4.988	8.194	4.988	7.152	7.571
202.5	1	7.281	4.171	6.853	4.635	8.147	4.635	7.281	7.829
217.5	1	7.425	4.037	6.945	4.375	8.055	4.375	7.425	7.963
232.5	1	7.575	4.037	7.074	4.225	7.926	4.225	7.575	7.963
247.5	1	7.719	4.171	7.232	4.197	7.768	4.197	7.719	7.829
262.5	1	7.848	4.429	7.409	4.292	7.591	4.292	7.848	7.571
277.5	2	7.953	4.795	7.591	4.503	7.409	4.503	7.953	7.205
292.5	2	8.028	5.242	7.768	4.816	7.232	4.816	8.028	6.758
307.5	2	8.067	5.742	7.926	5.210	7.074	5.210	8.067	6.258
322.5	2	8.067	6.258	8.055	5.657	6.945	5.657	8.067	5.742
337.5	2	8.028	6.758	8.147	6.128	6.853	6.128	8.028	5.242
352.5	2	7.953	7.205	8.194	6.590	6.806	6.590	7.953	4.795

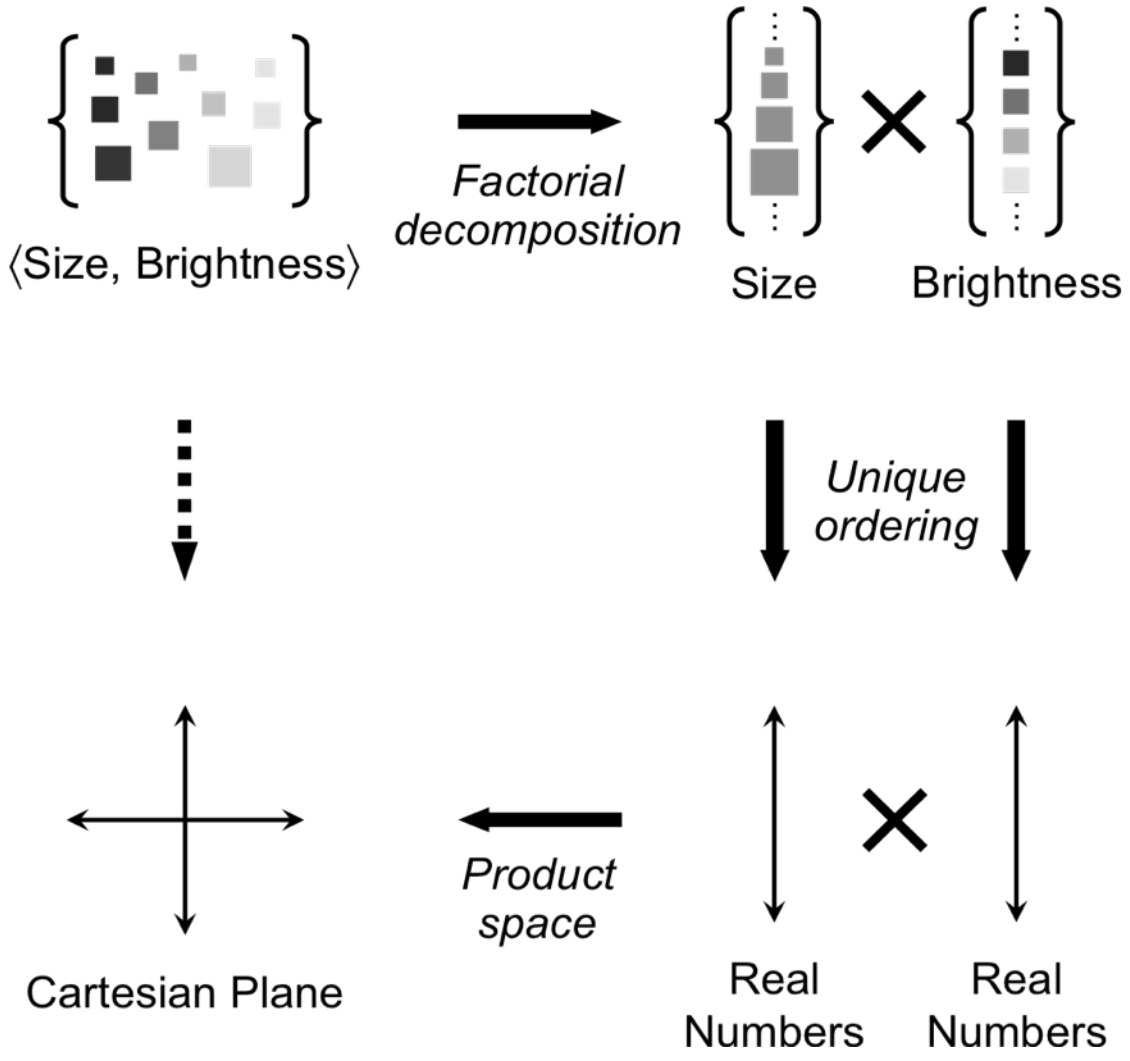

Figure 1. Illustration of logic behind Cartesian representation of separable dimensions. Upper left shows a set of stimuli varying in size and brightness. Because these are perceptually separable dimensions (e.g., Smith & Kemler, 1978), the stimulus space can be decomposed into separate representations of each (upper right). The component dimensions each have a single degree of freedom and a natural ordering, so they are both isomorphic to a subset of the real number line (lower right). This correspondence implies a correspondence between the perceptual representation of the joint stimulus space and the Cartesian plane (lower left).

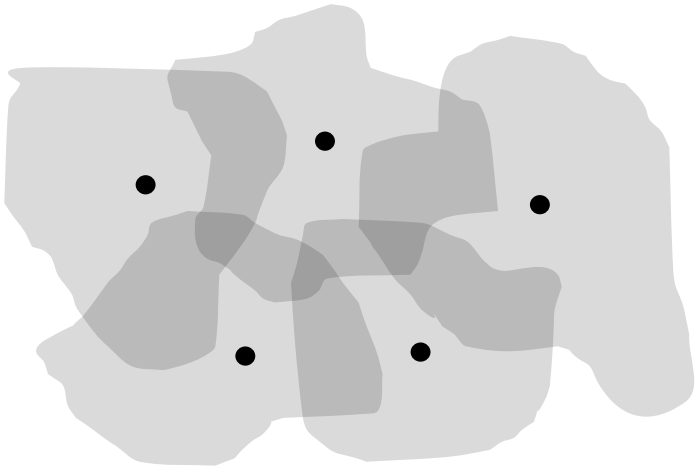

Figure 2. Illustration of a topological space. Points indicate example elements of the space, which correspond to stimulus values in the present model. Shaded regions indicate example open neighborhoods of those elements. In general, there are an infinite number of both elements and open neighborhoods (not shown). The structure of the topological space is determined by a specification of all of its open neighborhoods.

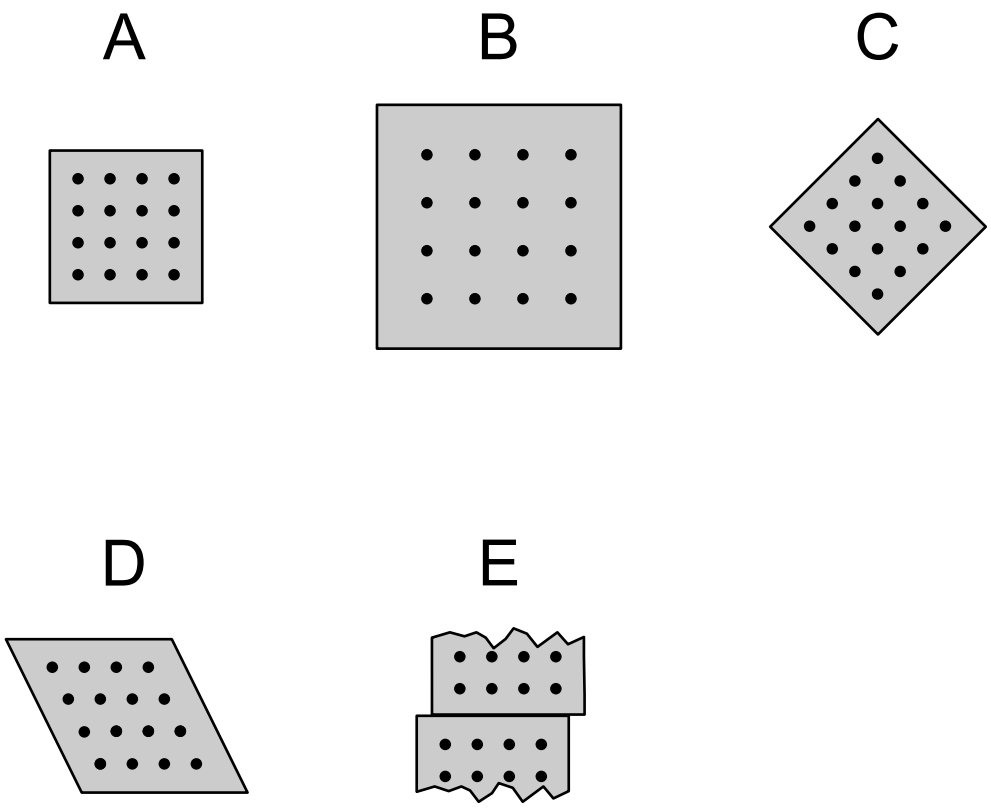

Figure 3. Illustration of the distinction between psychological commitments and incidental properties of models of perceptual representation. Each figure is a depiction of a continuous, integral perceptual stimulus space (grey region), with points indicating particular stimuli. Figures 3A and 3B differ only in overall scale, an incidental property of how the diagram is drawn. Thus, they can be interpreted as depicting exactly the same psychological representation. Figure 3C differs from 3A only in orientation. In Cartesian models of integral dimensions that assume Euclidean similarity metrics, this rigid rotation is also an incidental change with no psychological implications. Figure 3D differs from the others by a non-rigid transformation. Under the Cartesian model, it depicts a meaningfully different psychological representation (e.g., because the rows and columns of the highlighted stimuli are no longer orthogonal), but under the topological model, this too is an incidental transformation, and all four diagrams (3A-3D) depict exactly the

same psychological representation. Finally, Figure 3E depicts a psychological representation that is different from the others according to both Cartesian and topological models. This diagram differs from the others by a discontinuous transformation, in which the top and bottom halves of the stimulus space have been torn apart and rearranged.

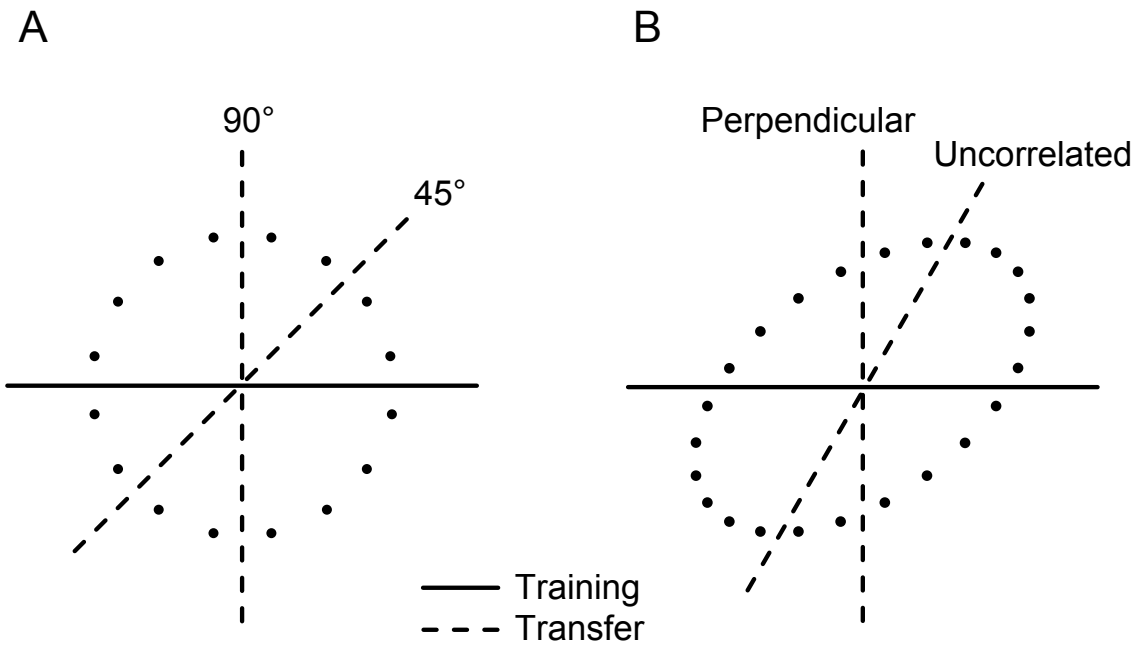

Figure 4. Illustration of dimension differentiation paradigm. Dots indicate stimuli. Solid and dashed lines indicate category boundaries for training and transfer tasks, respectively. Category boundaries determine which component dimension is diagnostic in each task. A: Conceptual design of Goldstone and Steyvers (2001). Orthogonal and Independence hypotheses both predict superior transfer performance in 90º condition, because diagnostic transfer dimension is both perpendicular to and uncorrelated with diagnostic training dimension in that condition. B: Conceptual design of Experiments 1 and 2. Elliptical stimulus distribution deconfounds whether diagnostic transfer dimension is perpendicular to and uncorrelated with diagnostic training dimension. Perpendicular transfer boundary corresponds to complementary dimension predicted by Orthogonal hypothesis; thus this hypothesis predicts superior transfer performance in Perpendicular condition. Uncorrelated transfer boundary corresponds to complementary dimension predicted by Independence hypothesis; this this hypothesis predicts superior transfer performance in Uncorrelated condition.

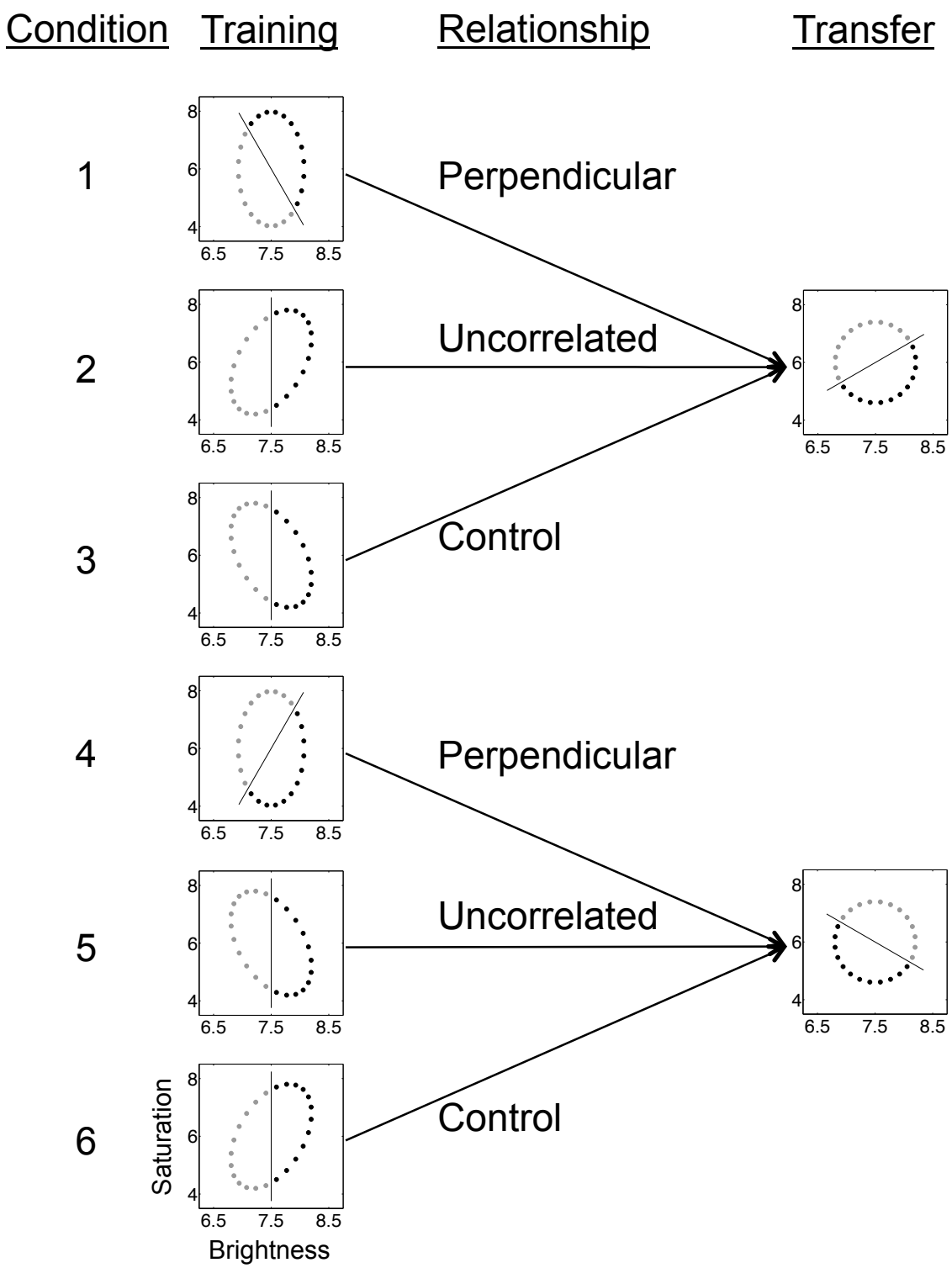

Figure 5. Design of Experiment 1. Each scatterplot shows the stimuli used within a particular phase and condition(s) of the experiment. Black and grey circles indicate

stimuli belonging to the two categories (training and transfer used different pairs of category labels). The boundary drawn through each stimulus set divides the two categories and is for illustrative purposes only. In Perpendicular conditions, the dimensions defining the training and testing categories were perpendicular. In Uncorrelated conditions, these dimensions were uncorrelated under the training stimulus distribution. Control conditions form a 2×2 design with Uncorrelated conditions, in which two training tasks (with the same diagnostic dimension but different stimulus distributions) were crossed with two transfer tasks.

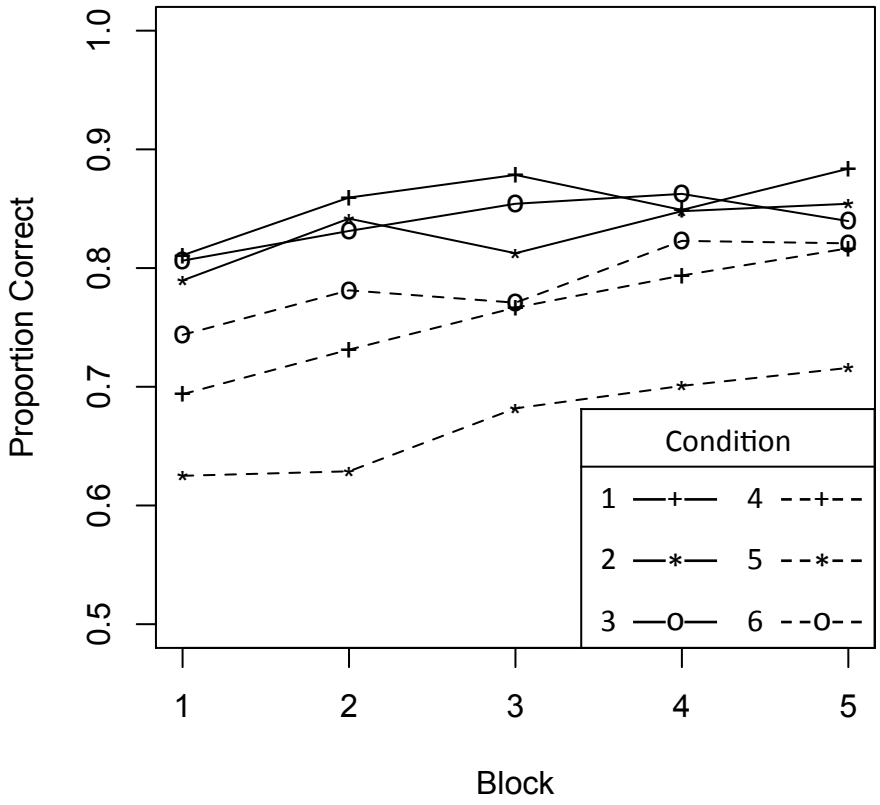

Figure 6. Learning curves for transfer phase of Experiment 1. Solid and dashed lines differentiate conditions using the two different transfer tasks. Perpendicular: +. Uncorrelated: ∗. Control: o.

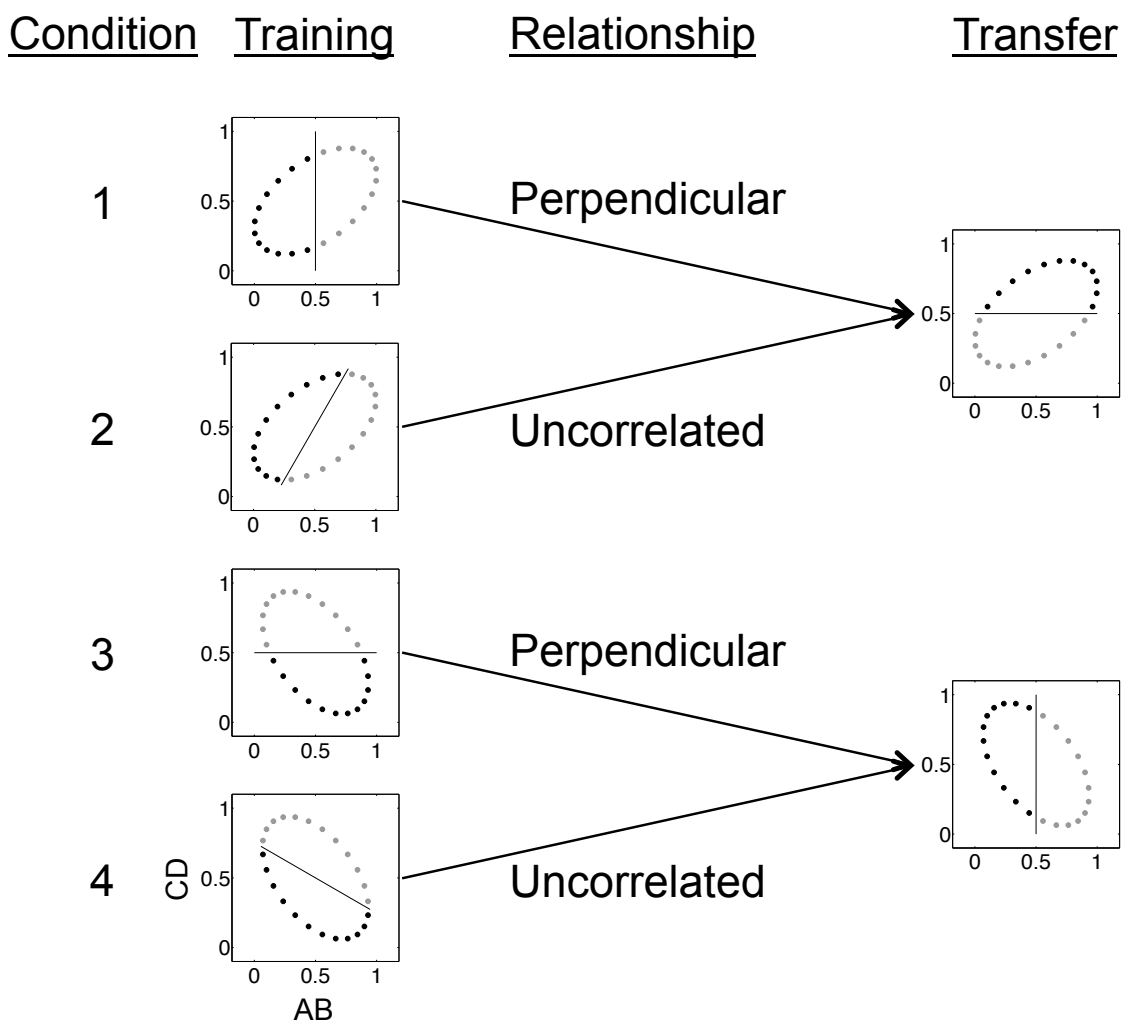

Figure 7. Design of Experiment 2, following same presentation scheme as Figure 5.

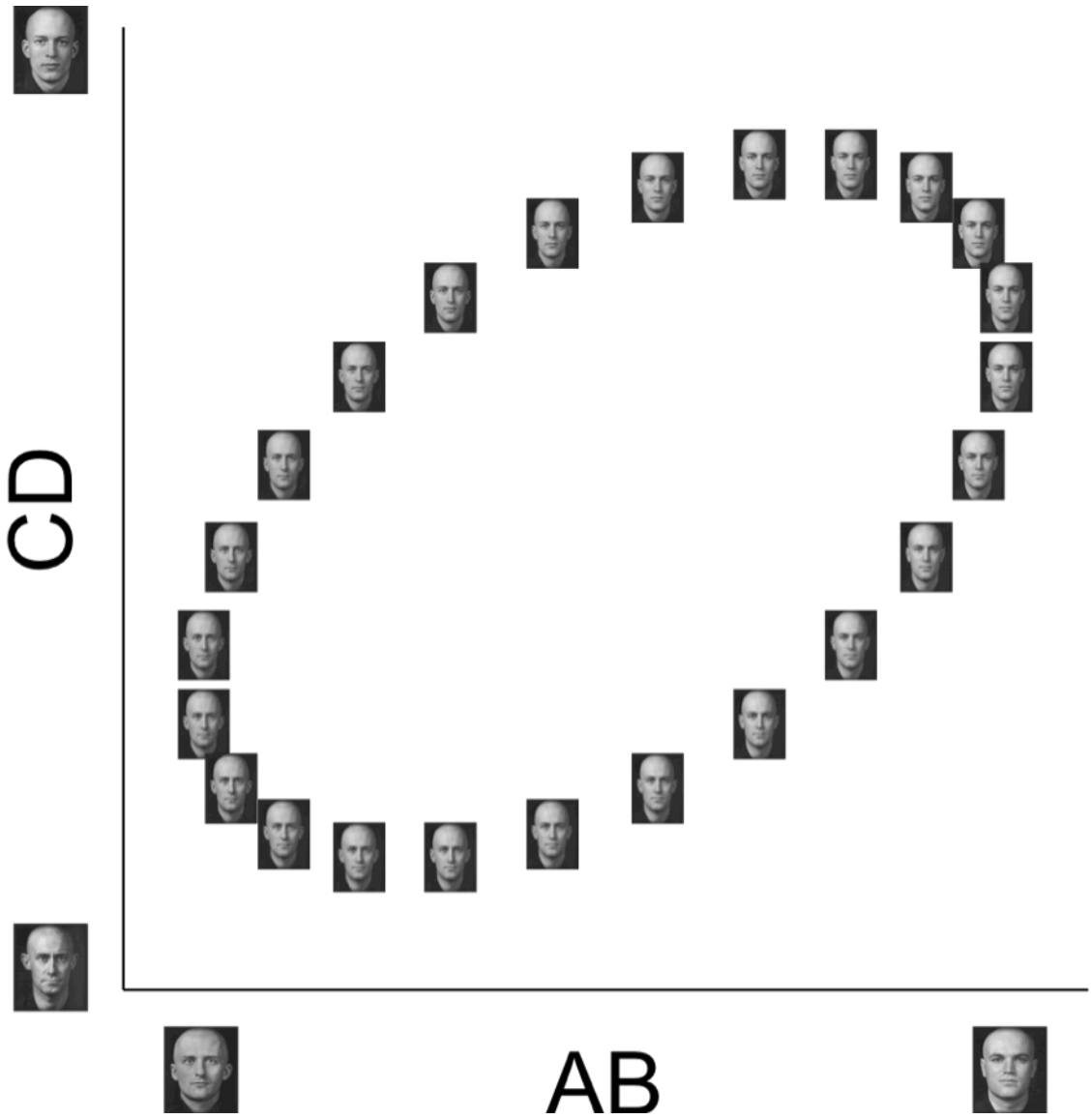

Figure 8. Stimuli used for Conditions 1 and 2 of Experiment 2. The images along the axes are the base faces from which the stimuli were generated. Stimuli for the other conditions and for Experiment 3 were drawn from the same stimulus space.

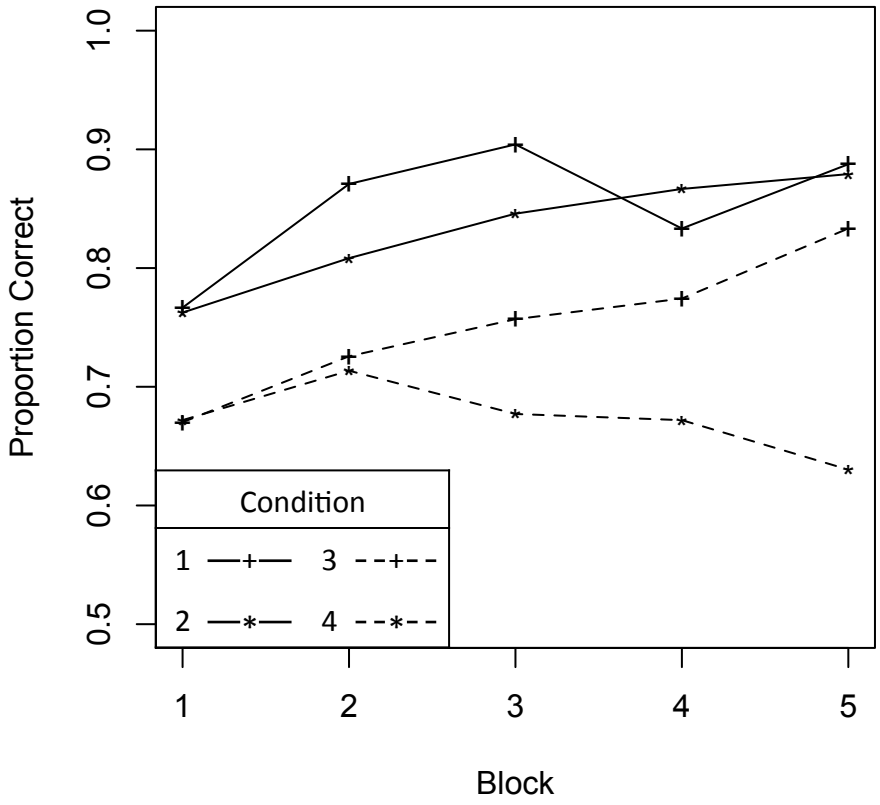

Figure 9. Learning curves for transfer phase of Experiment 2. Solid and dashed lines differentiate conditions using the two different transfer tasks. Perpendicular: +. Uncorrelated: ∗.

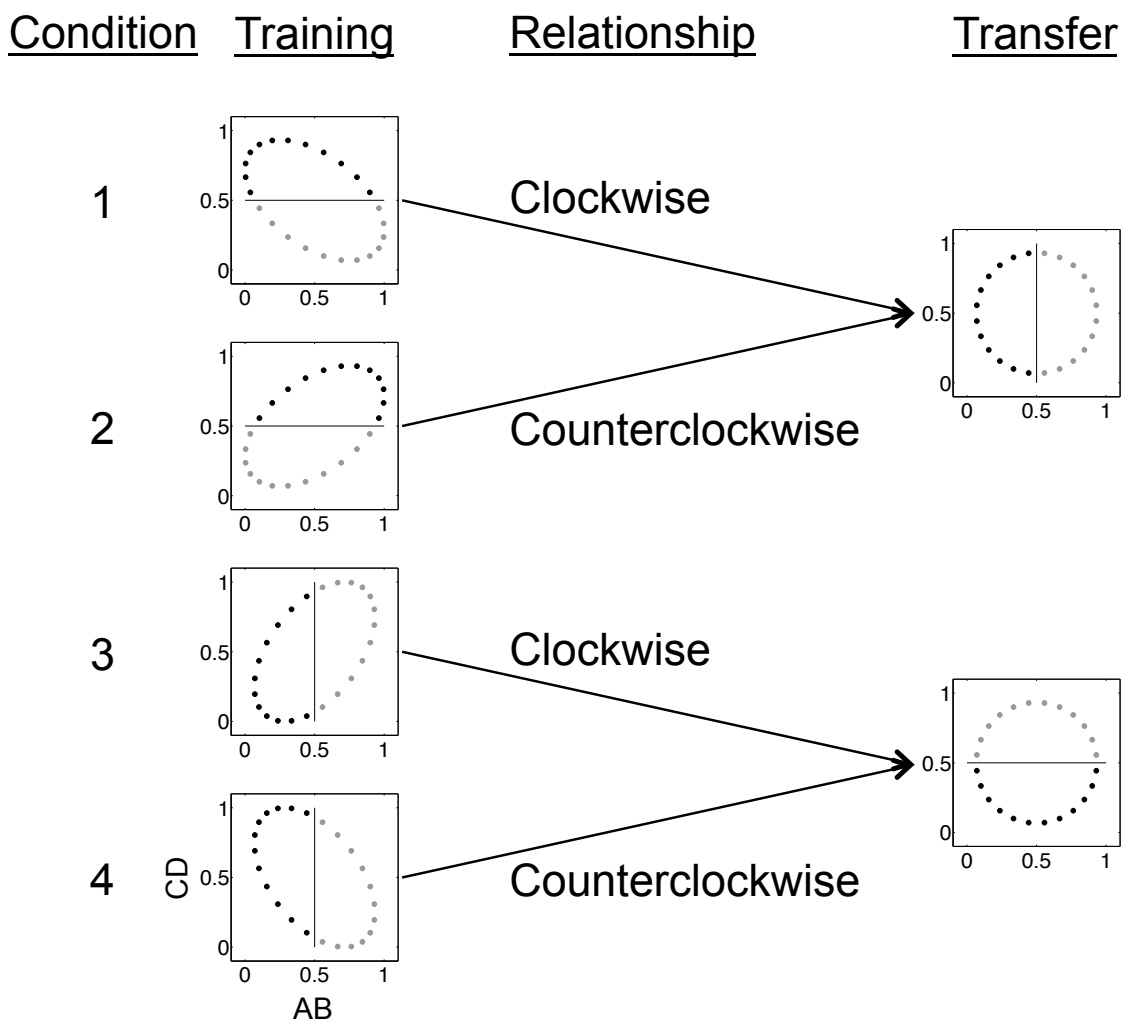

Figure 10. Design of Experiment 3, following the same presentation scheme as Figures 5 & 7. Clockwise and Counterclockwise refer to the orientation of the first principal component of the training stimulus distribution relative to the diagnostic transfer dimension. These labels also refer to the directions of predicted biases in participants' category judgments at transfer, according to the Unsupervised hypothesis.

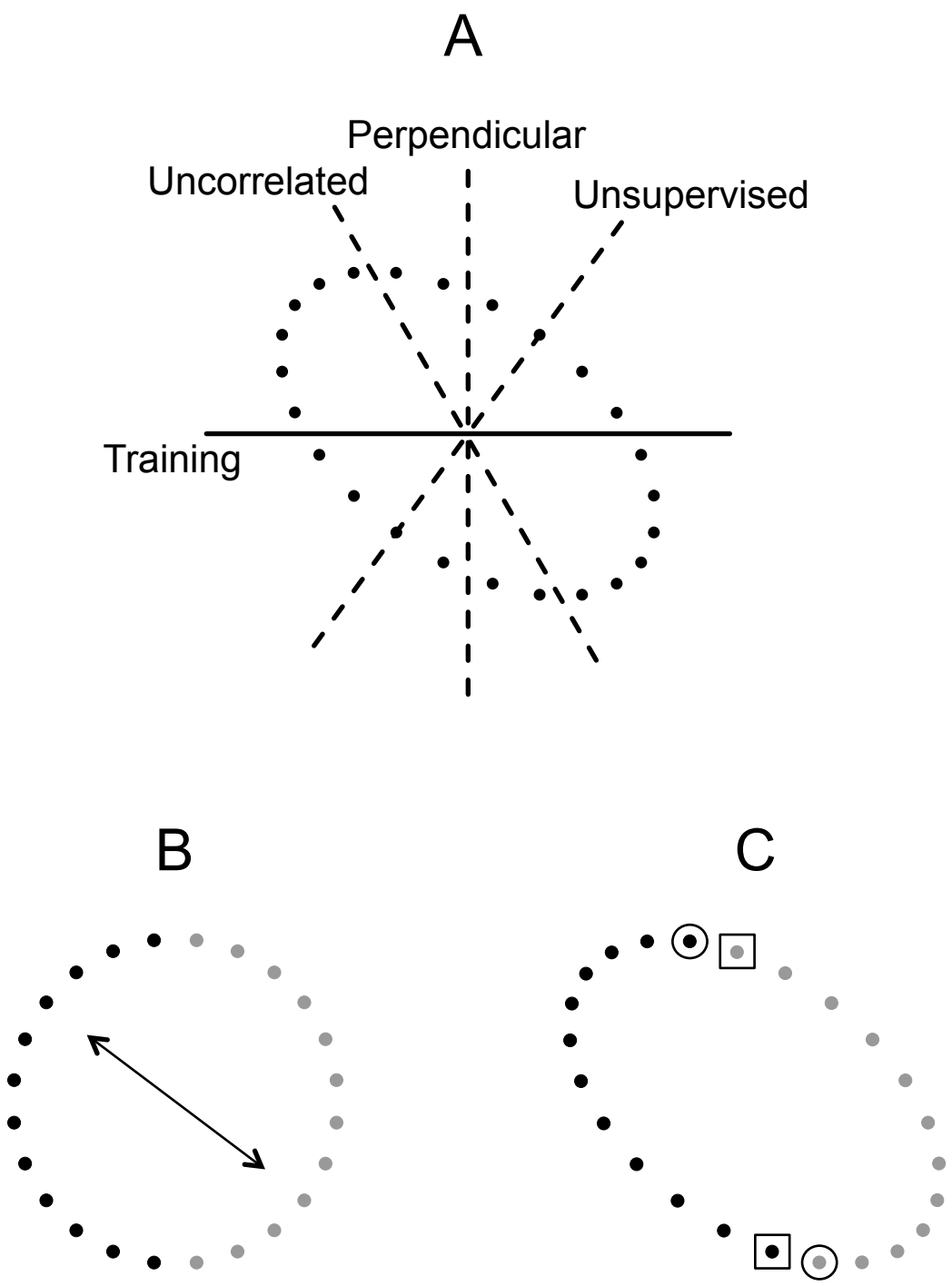

Figure 11. Illustration of predictions for Experiment 3, Clockwise Condition 1. A: Training task, with category boundary indicated by solid line. Dashed lines represent dimensions predicted to be learned by Independence (uncorrelated), Orthogonal (perpendicular),

and Unsupervised hypotheses (all dimensions indicated by isoclines). Perpendicular dimension matches the diagnostic dimension at transfer; hence Orthogonal hypothesis predicts good and unbiased transfer performance. Uncorrelated dimension is rotated counterclockwise from diagnostic transfer dimension; hence Independence hypothesis predicts counterclockwise bias in transfer classification errors. B: Transfer task, with arrow indicating effect of attention to unsupervised training dimension as predicted by Unsupervised hypothesis, which is modeled as stretching stimulus space. C: Resulting representation of transfer stimulus set. Unsupervised hypothesis predicts boxed border stimuli to be misclassified more often than circled border stimuli, and more generally a clockwise bias is predicted in errors over all stimuli. Counterclockwise condition is a mirror image; hence predictions of Unsupervised and Independence hypotheses are reversed in that condition.

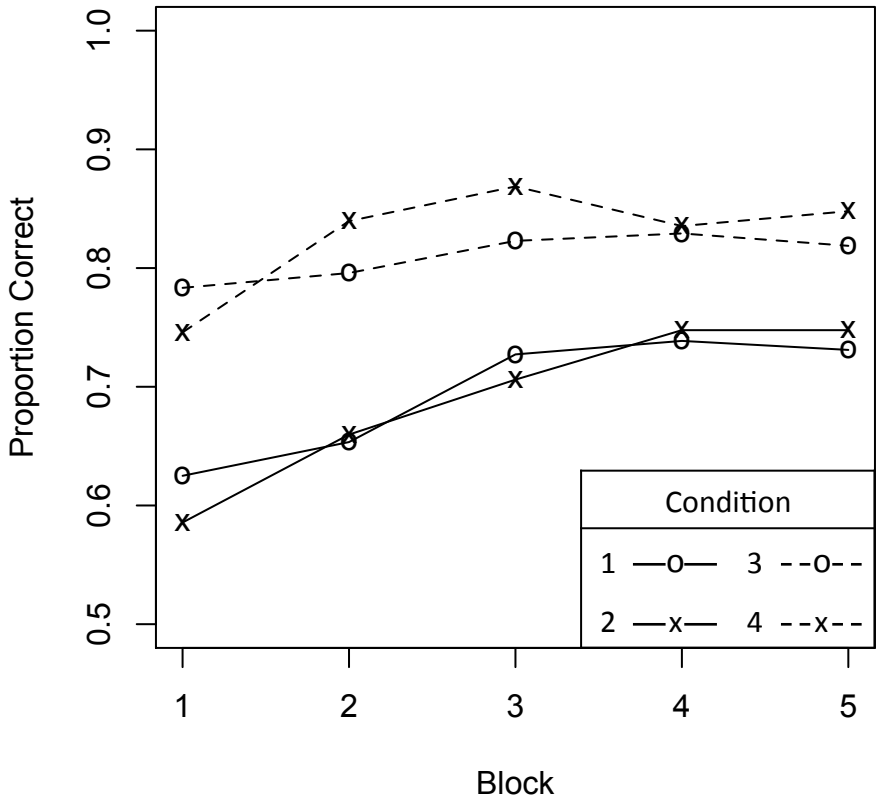

Figure 12. Learning curves for transfer phase of Experiment 3. Solid and dashed lines differentiate conditions using the two different transfer tasks. Clockwise: o. Counterclockwise: x.