Cognition 143 (2015) 163–177

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Why are some dimensions integral? Testing two hypotheses through causal learning experiments

Fabián A. Soto [a], Gonzalo R. Quintana [b], Andrés M. Pérez-Acosta [c], Fernando P. Ponce [b], Edgar H. Vogel [b,]*

[a] *Department of Psychology, Florida International University, United States*
[b] *Universidad de Talca, Chile*
[c] *Universidad del Rosario, Colombia*

ARTICLE INFO

Article history:
Received 30 December 2014
Revised 29 May 2015
Accepted 1 July 2015

Keywords:
Generalization
Configural and elemental processing
Separable and integral dimensions
Causal learning
Rational theory

ABSTRACT

Compound generalization and dimensional generalization are traditionally studied independently by different groups of researchers, who have proposed separate theories to explain results from each area. A recent extension of Shepard's rational theory of dimensional generalization allows an explanation of data from both areas within a single framework. However, the conceptualization of dimensional integrality in this theory (the direction hypothesis) is different from that favored by Shepard in his original theory (the correlation hypothesis). Here, we report two experiments that test differential predictions of these two notions of integrality. Each experiment takes a design from compound generalization and translates it into a design for dimensional generalization by replacing discrete stimulus components with dimensional values. Experiment 1 showed that an effect analogous to summation is found in dimensional generalization with separable dimensions, but the opposite effect is found with integral dimensions. Experiment 2 showed that the analogue of a biconditional discrimination is solved faster when stimuli vary in integral dimensions than when stimuli vary in separable dimensions. These results, which are analogous to more "non-linear" processing with integral than with separable dimensions, were predicted by the direction hypothesis, but not by the correlation hypothesis. This confirms the assumptions of the unified rational theory of stimulus generalization and reveals interesting links between compound and dimensional generalization phenomena.

1. Introduction

An important aspect of all forms of learning is generalization; that is, once we have learned something about the environment, to what extent do we generalize this knowledge to new situations, similar but not identical to the original learning events?

All fields in psychology dealing with learning and inference have explored one or more aspects of this problem. For example, *dimensional generalization*, or how learning about a stimulus is transferred to new stimuli that differ from the original along continuous dimensions, has been studied in animal instrumental conditioning (e.g., Blough, 1975; Guttman & Kalish, 1956; Soto & Wasserman, 2010; for a review of unidimensional generalization, see Ghirlanda & Enquist, 2003) and human identification and categorization (for reviews, see Nosofsky, 1992; Shepard, 1991). On the other hand, *compound generalization*, or how learning about one stimulus is transferred to new compounds comprising that stimulus, has been studied in Pavlovian conditioning (e.g., Myers, Vogel, Shin, & Wagner, 2001; Rescorla, 1997; Whitlow & Wagner, 1972) and human causal and contingency learning (e.g., Collins & Shanks, 2006; Glautier, 2004; Soto, Vogel, Castillo, & Wagner, 2009).

Unfortunately, these two lines of research have been pursued largely independently and researchers have shown little interest in developing a unified theoretical framework to understand both forms of generalization. Recently, Soto, Gershman, and Niv (2014) provided such unified framework by extending the rational theory of dimensional generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001) to the explanation of compound generalization phenomena. In the following two sections, we briefly review this theory, some of the relevant data that it attempts to explain and the open issues addressed by the present work. We then describe two experiments that aim to answer two of those open questions:

* Corresponding author at: Universidad de Talca, Facultad de Psicología, Talca, Chile.
E-mail address: evogel@utalca.cl (E.H. Vogel).

http://dx.doi.org/10.1016/j.cognition.2015.07.001
0010-0277/© 2015 Elsevier B.V. All rights reserved.

Why are some dimensions integral and others separable? Are the assumptions about integrality that are necessary to explain compound generalization also important to explain dimensional generalization?

1.1. Dimensional generalization and Shepard's rational theory

The most common basic result of a dimensional generalization experiment is that the response controlled by a stimulus orderly decreases as the value of the stimulus in one or more continuous dimensions is changed. An important insight in the study of dimensional generalization was the idea that re-scaling of stimulus dimensions to reveal "psychological dimensions" could lead to the discovery of fundamental principles of generalization and to stimulus representations that are useful for the study of other cognitive processes (for a review, see Nosofsky, 1992).

Indeed, two fundamental results about dimensional generalization have been found after such re-scaling. First, response probability decays as an exponential function of the psychological distance between a test stimulus and the original training stimulus (Shepard, 1965, 1987). Second, when stimuli are varied in two dimensions, the shape of the multidimensional generalization gradient varies depending on the exact dimensions under study (Cross, 1965; Shepard, 1987, 1991; Soto & Wasserman, 2010). Here the distinction between *separable* and *integral* dimensions becomes important (Garner, 1974; Shepard, 1991). Two dimensions are separable if it is possible to perceive or attend to only one dimension without attending to the other (e.g., size and orientation of a line). These dimensions produce diamond-shaped generalization gradients (see Fig. 1a), in which there is more generalization in the direction of the dimensions than in other directions of space. Diamond-shape gradients are equivalent to using a city-block metric to compute distances from coordinates in a spatial representation of the generalization data, such as that obtained from multidimensional scaling (MDS; Shepard, 1991). Two dimensions are integral if it is *not* possible to perceive or attend to only one dimension without attending to the other (e.g., saturation and brightness). These dimensions produce circular generalization gradients (see Fig. 1b), in which there is more or less the same generalization in any direction in the stimulus space. Circular gradients are equivalent to using an Euclidean metric to compute distances from coordinates in a spatial representation of the generalization data (Shepard, 1991).

Note that the fact that different sets of dimensions produce multidimensional generalization gradients with different shapes – or, equivalently, different metrics in a MDS representation – is an empirical result. The usual mechanistic explanation for this result is that different sets of dimensions interact differently during perception. Separable dimensions, but not integral dimensions, are processed independently and can be attended selectively (Garner, 1974).

A full account of generalization requires answering not only questions about mechanism, but also questions about function, such as: Why is the shape of unidimensional generalization gradients exponential instead of some other shape? Why do some dimensions seem to be processed separately and others integrally? Rational theories of cognition (Anderson, 1990) provide answers to such questions about function (Griffiths, Chater, Norris, & Pouget, 2012). Rational explanations propose hypotheses about what aspects of the task of generalization could have led, through adaptation, to the observable features of generalization behavior.

Shepard (1987) proposed a rational theory in which the properties of dimensional generalization are explained as resulting from probabilistic inference. The theory proposes that when an observer encounters a stimulus S1 followed by some significant consequence, S1 is represented as a point in a psychological space.[1] The stimulus is assumed to be a member of a natural class associated with the consequence. This class occupies a region in the observer's psychological space, called a *consequential region*. The only information that the observer has about this consequential region is that it overlaps with S1 in psychological space. After observing a new stimulus, S2, the inferential problem is to determine the probability that S2 belongs to the same natural kind as S1—the same consequential region—thus leading to the same consequence. This probability can be obtained by "hypothesis averaging," by taking all possible

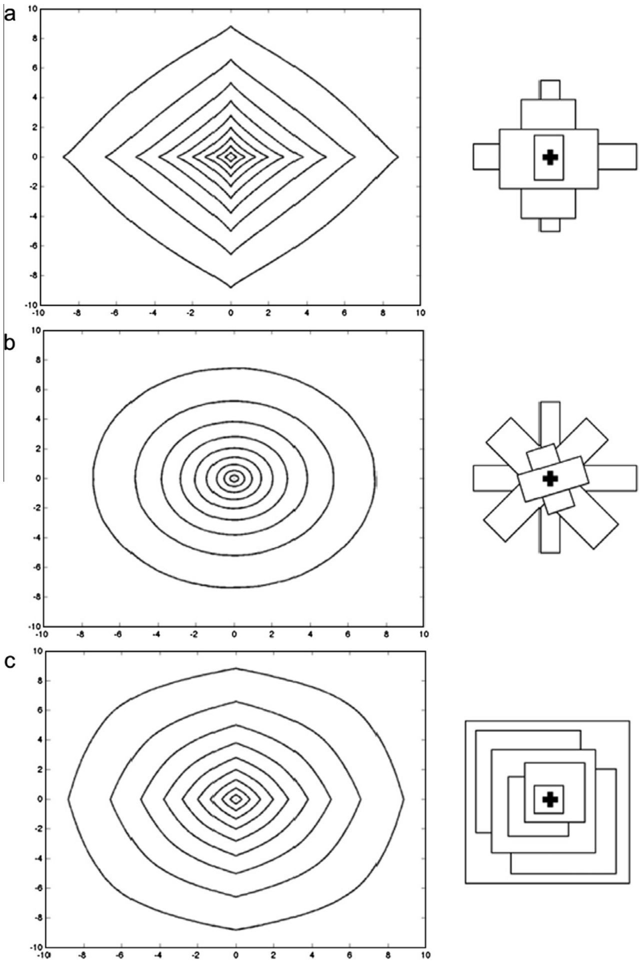


Fig. 1. Contour plots of multidimensional generalization gradients predicted by the consequential regions theory. The stimulus controlling a specific response is represented by the coordinates (0,0) and the scale in each axis represents distance from that stimulus along a specific perceptual dimension. Each line in a gradient represents the set of all points in the bidimensional space that have the same probability of generalization. These points of equal generalization probability assume the shape of a diamond for separable dimensions (a), and the shape of circles for integral dimensions using the direction hypothesis (b) and for integral dimensions using the correlation hypothesis (c). To the left of each gradient several examples of regions considered to evaluate the gradients are shown.

[1] There are two important points to clarify about Shepard's theory. First, when Shepard's first paper was published, Anderson's "rational" level (Anderson, 1990) had not yet been proposed. However, the most common interpretation of the theory is as a rational analysis of generalization (e.g., Soto et al., 2014; Tenenbaum & Griffiths, 2001). Second, despite being a rational analysis, the theory still makes representational assumptions that should be clearly separated from its assumptions about the generalization task (Fernbach & Sloman, 2011). The most important of these assumptions is that the observer represents stimuli as points in a psychological space. Importantly, explanations of generalization phenomena are a direct consequence of how the theory formalizes the inferential task of generalization, with the representational assumptions playing a minor role in such explanations.

consequential regions that contain S1, each with a different size and position, and computing the proportion of them that also contain S2.

When this process is repeated for many possible values of S2, the resulting probability falls approximately exponentially with distance between S1 and S2 in psychological space. More importantly for the present work, the shape and orientation of consequential regions in space have an important impact in the shape of multidimensional generalization gradients. Gradients shaped like a diamond, typical of separable dimensions, are obtained when the sides of each region are aligned with the stimulus dimensions and their sizes vary independently. That is, separable dimensions "correspond to uniquely defined independent variables in the world" (Shepard, 1987, p. 1322). Fig. 1a shows the contours of a generalization gradient obtained this way (for details, see the Supplementary Material) and examples of the kind of consequential regions that produce the gradient, which look like rectangles in which the size on one side is unrelated to the size on the other side. The plus sign in the figure represents S1, and each contour is the set of all points that have the same probability of coming from the same consequential region as S1. Gradients that are more circular, as those found with integral dimensions, are obtained by constraining regions to be squares, which have the same size in each dimension, as in Fig. 1c. That is, integral dimensions are those for which "there has been a positive correlation between the ranges of variation of stimuli associated with important consequences" (Shepard, 1991, p. 68). We will call this the *correlation hypothesis*.

As shown in Fig. 1b, circular generalization gradients can also be reproduced by assuming that rectangular regions have uncorrelated sizes along each dimension, but can have any orientation in space (Austerweil & Griffiths, 2010). According to this *direction hypothesis*, some dimensions are integral because natural classes can extend in any direction of the space created by such dimensions. The direction hypothesis deviates importantly from Shepard's theory in that it drops the assumption that natural classes must have similar extent on all integral dimensions (see Shepard, 1991, p. 67–68). As can be seen from Fig. 1, the shape of multidimensional generalization gradients cannot be used to distinguish between the correlation and direction hypotheses.

In summary, within Shepard's theory there are two ways to explain the distinction between integral and separable dimensions. According to the *direction* hypothesis, natural classes might extend only in the direction of the axes of stimulus space (separable dimensions) vs. in any direction of stimulus space (integral dimensions). According to the *correlation* hypothesis, natural classes might have different and independent extensions along the two axes of stimulus space (separable dimensions) vs. have the same extension along the two axes of stimulus space (integral dimensions). For Shepard, the most important distinction between separable and integral dimensions is captured by the correlation hypothesis (see Shepard, 1991, p. 68). Thus, two conceptually different versions of the rational theory of generalization exist, both of them capable to explain multidimensional generalization equally well, and there is no *a priori* reason to prefer one over the other.

1.2. Compound generalization and a unified theory

In the fields of associative and causal learning, much research has focused on examining to what extent learning about a stimulus generalizes to new compounds containing that stimulus. For example, in a typical *summation* experiment (e.g., Collins & Shanks, 2006; Rescorla, 1997; Soto et al., 2009; Whitlow & Wagner, 1972), participants might learn that both broccoli (stimulus A) and tomato (stimulus B) independently produce an allergic reaction of a certain intensity in an hypothetical patient, and then they are asked to predict the intensity of the allergic reaction to the compound "broccoli + tomato" (AB). A summation effect is found when there is a higher response to the compound than to its components.

Studies have found that compound generalization depends on the type of stimuli used as components, among other factors (for a review, see Melchers, Shanks, & Lachnit, 2008). For example, studies on the summation effect in Pavlovian conditioning that used stimuli of distinct sensory modalities, generally found evidence of summation (Rescorla, 1997; Whitlow & Wagner, 1972). On the contrary, absence of summation has been observed in studies using only visual stimuli (Aydin & Pearce, 1995; Rescorla & Coldwell, 1995). Kehoe, Horne, Horne, and Macrae (1994) examined this issue directly and confirmed that summation is found with stimuli of distinct sensory modalities (*i.e.*, tone-light and noise-light), but not with stimuli from the same modality (tone-noise).

To explain this pattern of results, most contemporary models of Pavlovian conditioning share the idea that the representation of a stimulus should be "nonlinear" (Shanks, Lachnit, & Melchers, 2008) or "context sensitive" (Wagner, 2003); that is, a stimulus presented in isolation is represented differently than when it is presented in compound with other stimuli. The theories of Wagner (2003), McLaren and Mackintosh (2002), and Harris (2006) assume that stimuli are represented by a set of elements whose activity, apart from depending on the stimulus they represent, depends on the presence or absence of other stimuli. Pearce (1987, 1994), suggests instead that compound stimuli should be represented as unique exemplars, and that the constituent elements only play a role in determining the degree of generalization between configurations.

Although elemental and configural approaches, such as those of Wagner (2003) and Pearce (1994), respectively, are often presented as radically distinct views of stimulus representation, it is important to recognize that they both permit non-linear stimulus processing,[2] although to varying degrees. The degree of non-linear (i.e., not additive: the whole is different than the sum of elements) processing assumed by a particular model has an influence on its predictions regarding the summation effect. If a markedly nonlinear processing (or highly context sensitive) is assumed, with very little of the associative strength of A and B generalizing to the compound AB, then it would be expected for the response to AB to be equal or lower than that to A and B. Conversely, if more linear processing (the whole approximates to the sum of elements) is assumed, then it would be expected that much of the associative strength of A and B is transferred to AB, obtaining greater responses to the compound than to its elements (i.e., summation).

To explain results like those of Kehoe et al. (1994), some models allow flexibility in the level to which they show nonlinear stimulus processing. The main determinant for the level of nonlinear processing in these models is stimulus similarity, or the level of overlap between the representation of two stimuli (Harris, 2006; McLaren & Mackintosh, 2002; Pearce, 1987, 1994). A related, but slightly different hypothesis is based on the distinction between integral and separable dimensions. Integral dimensions are thought to interfere with each other much more than separable dimensions do. Thus, in a stimulus formed by two integral dimensions, the identity of each dimension would tend to disappear, generating a new fusion, while with separable dimensions there would be a conservation of the identity of each dimension. Following this reasoning, several authors have suggested that components that

[2] In fact, Ghirlanda (2015) has provided a formal proof that any configural model can be translated into an equivalent elemental model and vice versa, so that they make identical predictions, providing that some conditions are met. Such conditions are met by current associative learning models.

lead to more nonlinear representations are analogous to integral dimensions, whereas components leading to more linear representations are analogous to separable dimensions (Lachnit, 1988; Melchers et al., 2008; Myers et al., 2001; Wagner & Vogel, 2008). However, this analogy has not been worked out quantitatively within the mechanistic framework of associative learning theories.

Recently, Soto et al. (2014) proposed a rational theory that formalizes this hypothesis and provides a unified framework to explain both dimensional and compound generalization. As in Shepard's theory, the model formalizes a generalization task in which an observer experiences some stimulus followed by an important consequence and assumes that there is a natural class of similar stimuli that will also produce the consequence. This set of stimuli form a consequential region in stimulus space whose extension must be inferred from data. When more than one stimulus is presented at the same time, as in compound generalization experiments, they could belong to a single consequential region or to multiple consequential regions. Thus, the learner must infer not only the size but the number of consequential regions involved and which regions produced which observed stimuli. Fig. 2 depicts an example in which four stimuli, varying along two dimensions, are assumed to belong to two different consequential regions. Region Z_1 includes stimuli A and B and is associated with an outcome, represented by the letter O. Region Z_2 includes stimuli C and D and is associated with no outcome, represented by $\sim$O.

Importantly, consequential regions with small sides, which represent more precise hypotheses, are weighted more heavily than other hypotheses (i.e., the model assumes the *size principle*, see Tenenbaum & Griffiths, 2001). This includes the case in which only one of the sides is very thin while the other side is elongated, as exemplified by region Z_2 from Fig. 2, which is elongated along dimension 1, but short along dimension 2.

To understand better how the model works, take the example of a summation experiment with stimuli that vary in two dimensions. After observing the compound AB, one possibility is that each component belongs to a different consequential region, as shown in Fig. 3a. If each region is also independently associated with an outcome O, then the model would predict that the outcome after AB is the sum of the outcomes after A and B; that is, a summation effect. A different possibility is that both A and B belong to the same region, as shown in Fig. 3b. In this case, the model would predict that the consequence after AB is the same as after either A or B; that is, no summation effect.

Whether the model infers one or two consequential regions for A and B depends on the kind of dimensions on which they vary. As shown in Fig. 3b, when dimensions are integral and consequential regions can be oriented in any direction, it is possible to find a single elongated region that includes both stimuli. This hypothesis would be weighted more heavily than other hypotheses that can explain the data. On the other hand, Fig. 3a shows that when dimensions are separable and consequential regions can only elongate along the axis of the stimulus space, there is no single elongated region that can produce both A and B. In this case, one hypothesis that would be weighted heavily is one in which each stimulus has been produced by separated and small consequential regions.

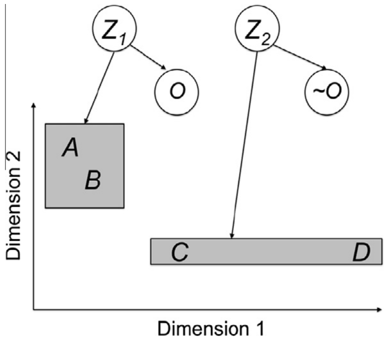


Fig. 2. Schematic representation of the consequential regions model used by Soto, Gershman and Niv (2014) to explain compound generalization phenomena.

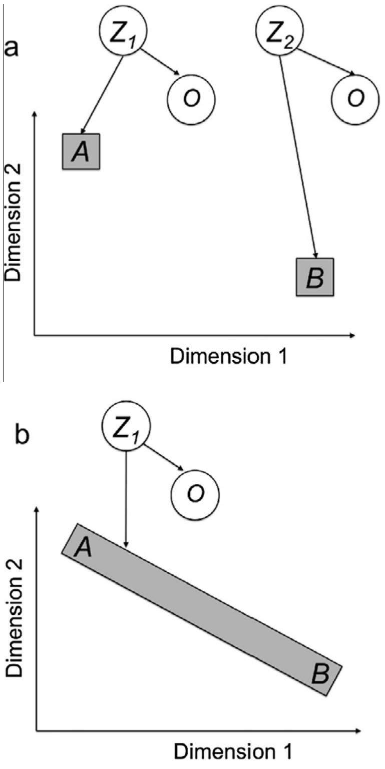


Fig. 3. Schematic representation of how different outcomes of a summation experiment can be explained within the framework of consequential regions theory.

According to this model, it is more likely to find summation with stimuli from different modalities because they are the quintessential example of separable dimensions (Garner, 1974; for a review of the evidence, see Marks, 2004). On the other hand, stimuli from the same modality, such as a noise and a tone, are more likely to vary along integral dimensions. As this example illustrates, the success of the theory to explain a variety of compound generalization phenomena strongly depends on the assumption that the direction hypothesis is the correct way to distinguish between integral and separable dimensions. If the correlation hypothesis were true, then consequential regions varying along integral dimensions would have the shape of squares aligned to the main dimensional axes, as shown in Fig. 1c. Elongated regions such as that shown in Fig. 3b would not be possible and, as was the case with separable dimensions, the hypothesis that A and B are generated by different regions would have much higher likelihood than the hypothesis that they are generated by a single region.

1.3. The present study

It remains to be tested whether the assumption made by Soto et al. (2014) to explain compound generalization (i.e., the direction hypothesis) will also prove to be successful in explaining dimensional generalization phenomena. As indicated earlier, both the direction hypothesis and the correlation hypothesis can reproduce generalization gradients from separable and integral dimensions, with the correlation hypothesis being favored by Shepard (1987, 1991). Thus, testing these two hypotheses in dimensional generalization experiments would deepen our understanding of both dimensional and compound generalization. In dimensional generalization, it would allow to determine what version of the consequential regions theory can explain the data best, leading to a better understanding of what distinguishes separable and integral dimensions. In compound generalization, it would confirm or disprove the validity of the assumptions used by the model of Soto and colleagues, providing a test of whether this model is truly a good candidate for a unified theory of stimulus generalization.

The goal of the present study is to test, in two experiments on dimensional generalization of causal learning, the predictions made by the direction and correlation hypotheses. The two experiments exploit an interesting aspect of the analogy between compound and dimensional generalization, first observed by Lachnit (1988): it is possible to take a design from compound generalization and translate it into a dimensional generalization experiment by treating dimensional values as stimulus components. Experiment 1 uses a design analogous to a summation experiment (as in Lachnit, 1988), whereas Experiment 2 uses a design analogous to a biconditional discrimination (Saavedra, 1975). More importantly, only the direction hypothesis makes predictions for such designs in which integral dimensions lead to results that are analogous to nonlinear processing and separable dimensions lead to results that are analogous to linear processing.

2. Experiment 1

In an attempt to connect the notions of integrality and separability with the type of stimulus processing in compound generalization, Lachnit (1988) adapted a summation design to dimensional generalization by exchanging the discrete stimuli A and B for values of a single stimulus in two dimensions. His experiments used a human Pavlovian conditioning preparation with stimuli varying in four values of each of two dimensions (Dimension A with values a1, a2, a3, and a4; Dimension B with values b1, b2, b3 and b4), as shown in Fig. 4a. Training comprised trials with two combinations that were always followed by an outcome (a1b2 → O and a3b4 → O) and two combinations that never were followed by the outcome (a2b1 → ~O and a4b3 → ~O). Upon the termination of training, summation was evaluated by testing with a novel stimulus formed by dimensional values previously followed by the outcome (a3b2) and a novel stimulus formed by dimensional values previously not followed by the outcome (a2b3).

It is easy to see from Fig. 4 that although the testing stimulus a3b2 shares dimensional values with training stimuli associated with the outcome (a1b2 and a3b4), it is closer in space to training stimuli associated with no outcome (a2b1 and a4b3). The opposite is true about testing stimulus a2b3. Lachnit (1988) reasoned that if separable dimensions are processed linearly, then a summation effect should occur and a3b2 should be strongly associated with the outcome whereas a2b3 should be strongly associated with no outcome. Consequential regions theory predicts the same: because regions extend in the direction of separable dimensions, hypotheses in which the stimulus a3b2 shares a consequential region with a1b2 or with a3b4, and hypotheses in which the stimulus a2b3 shares a consequential region with a2b1 or with a4b3, should be weighted heavily (Fig. 4b).

Lachnit (1988) also proposed that if integral dimensions are processed nonlinearly, then the opposite result should hold true: a3b2 should be associated with no outcome because of generalization from the closer stimuli a2b1 and a4b3, whereas a2b3 should be associated with outcome because of generalization from the closer stimuli a1b2 and a3b4. This prediction, however, is not in line with non-linear processing in current theories of Pavlovian conditioning (e.g., Harris, 2006; McLaren & Mackintosh, 2002; Pearce, 1987, 1994; Wagner, 2003), which predict that the association with an outcome acquired by the "components" a3 and b2 should generalize to a3b2.

Lachnit's (1988) prediction with integral dimensions aligns better with the direction hypothesis from consequential region theory: if consequential regions can be oriented in any direction of an integral space, then a strong hypothesis is that stimulus a3b2 shares a diagonally-oriented consequential region with the training stimuli a2b1 and a4b3, as shown in Fig. 4c.[3] Similarly, a2b3 is likely to share a region with a1b2 and a3b4. As we described above with the example of summation, the correlation hypothesis does not allow for such elongated and diagonally-oriented consequential regions, so it cannot make the same prediction as the direction hypothesis.

Fig. 5a shows the predictions of the model of Soto et al. (2014) for this design, when the direction hypothesis is implemented (see Section A of the Supplementary Material for details of the simulations). The most important prediction is of an interaction in which an effect analogous to summation (a3b2 > a2b3) is expected with separable dimensions, but the opposite effect (a3b2 < a2b3) is expected with integral dimensions. As shown in Fig. 5b, the correlation hypothesis does not predict an effect opposite to summation (a3b2 < a2b3) with integral dimensions.[4]

Lachnit's (1988) results were in line with his predictions and the direction hypothesis, suggesting that the direction hypothesis captures dimensional generalization in Pavlovian conditioning better than the correlation hypothesis.

Although the results of Lachnit (1988) were straightforward, his experiments included a confounding factor that is quite common in studies that compare integral and separable dimensions. Lachnit used completely different sets of dimensions for the separable and integral conditions and the scaling within each dimension could have favored a specific pattern of generalization to the critical test stimuli (see Section B of the Supplementary Material for a fuller description of this issue). One solution to this problem is pairing each dimension in the study with one dimension with which it forms an integral pair and with one dimension with which it forms a separable pair. In this balanced design, any artifacts should affect both conditions equally and be averaged out in the group results. Thus, to improve upon Lachnit's study, Experiment 1 followed the design shown in Fig. 4, but balancing the assignment of each dimension to the separable and integral conditions.

An additional novel contribution of the present experiment is that it used a causal learning task, allowing us to determine whether the results found by Lachnit in Pavlovian conditioning can be found in experiments involving the more "cognitive" tasks

[3] In Fig. 4 and in our simulations, stimuli are equally spaced along each dimension. With very large deviations from equal spacing, the hypothesis shown in Fig. 4c is not strong. However, thin and elongated diagonal consequential regions can still account for pairs of stimuli, including a testing stimulus and one of its closest training stimuli (such as a3b2 and a2b1). Such consequential regions can be very thin and thus also constitute strong hypotheses. Furthermore, such diagonally-oriented regions are shorter than vertically- and horizontally-oriented regions like those shown in Fig. 4b, thus being stronger hypotheses and leading to the prediction shown in Fig. 5a.

[4] The simulations presented in Figs. 5 and 8 assumed squared consequential regions. A different way to implement the correlation hypothesis is through circular regions (Shepard, 1987). We performed additional simulations using circular regions for integral dimensions; their results are presented in the supplementary material. In short, while circular regions can qualitatively capture the most important results of Experiment 1 for group integral, the predicted difference between a2b3 and a3b2 is so small that it would have been difficult to detect in our experiment. Furthermore, circular regions could not capture the results of Experiment 2.

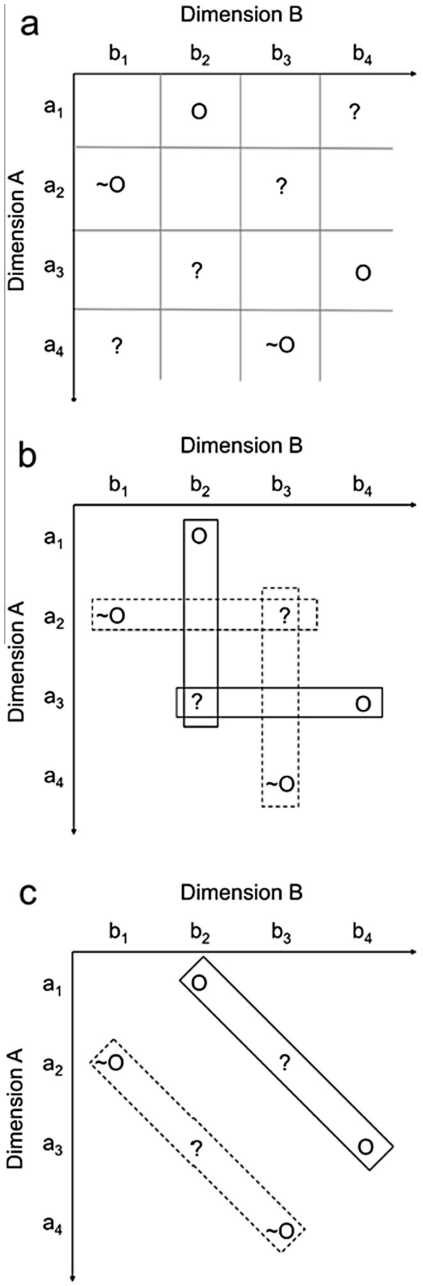


Fig. 4. Task used in Experiment 1 (a) and examples of regions heavily weighted by the consequential regions model implementing the direction hypothesis, both for separable dimensions (b) and for integral dimensions (c).

that are commonly used to study generalization in humans (e.g., Glautier, 2004; Soto et al., 2009).

2.1. Method

2.1.1. Participants

A total of 48 undergraduate psychology students at the University of Talca participated in the experiment for course credit. They had a mean age of 18.2 years ($S = 1.1$). They were tested individually and had no previous experience in similar research.

2.1.2. Materials

The stimuli were presented and data were collected using a HP Compaq personal computer connected to a 14-in. color screen and programmed using the E-prime software (Version 1.1; Psychology Software Tools, Inc., Pittsburgh, PA). The stimuli were constructed using a polygon and presented inside a white display screen in the shape of an irregular square (see Fig. C3 in the Supplementary Material). The stimuli were obtained by varying the same object in eight dimensions: brightness (hue 5RP varying in brightness value = 4:, 5:, 6:, and 7:, obtained from the Munsell color library of the Macromedia Freehand MX software), saturation (hue 5RP varying in chroma = :1,:4,:8, and:12, obtained from the Munsell color library of the Macromedia Freehand MX software), vertical position (5%, 25%, 50%, and 75% of the vertical axis of the display screen), horizontal position (5%, 25%, 50%, and 75% of the horizontal axis of the display screen), rectangle-height (15%, 20%, 25%, and 30% of the height of the display screen), rectangle-width (50%, 60%, 70%, and 80% of the width of the display screen), rotation around the *X*-axis (0°, 45°, 135°, 180°) and rotation around the *Y*-axis (0°, 45°, 135°, 180°). Note that the monitors were not calibrated, so although brightness and saturation were manipulated directly in the software used to create stimuli, we cannot guarantee that the Munsell chip values reported here were actually displayed. However, there is no reason to suspect that this technical limitation would do anything beyond adding noise to our results.

In order to equate intradimensional generalization between integral and separable conditions, each dimension was combined with one integral and one separable dimension. The integral dimensions were brightness and saturation (Integral 1; Lachnit, 1988) vertical and horizontal position (Integral 2; Garner & Felfoldy, 1970), rectangle width and height (Integral 3; Dunn, 1983; Monahan & Lockhead, 1977) and rotation around the *X*-axis and around the *Y*-axis (Integral 4; Soto & Wasserman, 2010). Stimuli in the separable sets were constructed with exactly the same values as those of the integral sets. The separable dimensions were Saturation and horizontal position (Separable 1), vertical position and brightness (Separable 2), height and rotation around the *Y*-axis (Separable 3), and rotation around the *X*-axis and width (separable 4). When not used, the dimensions were set at constant values. Since the participants had to solve two discriminations, they were randomly assigned to one of the following four subgroups ($n = 6$): integral 1 and 2, integral 3 and 4, separable 1 and 2, and separable 3 and 4. The specific stimuli involved in these integral and separable combinations are reproduced in the supplementary material (Figs. C1 and C2). Furthermore; in Section D of the supplementary material we present experimental evidence for the integrality and separability of this set of stimuli.

2.1.3. Procedure

At the beginning of the training phase, the participants were given instructions indicating that they would play the role of a dermatologist investigating what microorganisms would cause aging or rejuvenating effects in the skin (see part C of the supplementary material). Next, a series of 200 trials was presented to the participant (screenshots from training and testing trials are shown in Fig. C3 of the supplementary material). At the beginning of each trial, the sentence "the sample contains the following microorganism" appeared on the left-top portion of the screen simultaneously with the stimulus (microorganism). The presentation of the stimulus was followed 2 s later by the phrase, "What reaction you think this microorganism will cause in the rat's skin?" and the participants were required to answer "aging", "neutral", or "rejuvenation", by clicking the respective buttons. After the participant entered a response, feedback was provided on the bottom of the screen for 3 s. The feedback consisted of the words "CORRECT" or "INCORRECT", in yellow, over the word representing the programmed outcome, in white (i.e., "aging", "neutral", or "rejuvenation"). The trial terminated with a new screen of 1 s duration reporting the cumulative percent of correct responses.

Each participant was asked to learn simultaneously two discriminations of the type shown in Fig. 4a. One discrimination

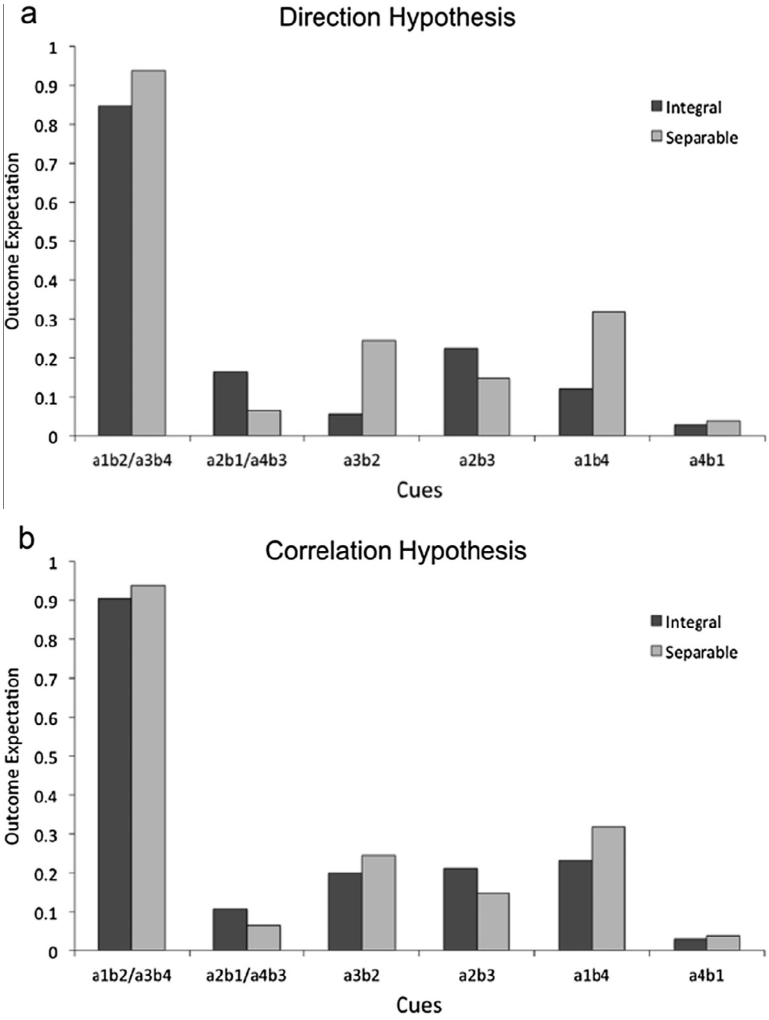


Fig. 5. Predictions of the direction (a) and correlation (b) hypotheses for stimuli tested in Experiment 1.

involved distinguishing between neutral and aging reactions and the other between neutral and rejuvenation reactions. This gives a total of 8 trial types that were presented in random order in 25 blocks of 8 trials each. A different set of stimuli was used in each discrimination. The assignment of aging or rejuvenation outcomes to each set of stimuli was counterbalanced across the participants of each subgroup.

Upon completion of the 200 training trials, the participants were presented with a series of testing trials. Each test stimulus appeared separately in the top center of the screen and the participants were asked to estimate the skin reaction to the presented microorganism by choosing a number from −5 to 5 in an 11-points scale going from “maximal aging” to “maximal rejuvenation,” with zero representing a neutral effect. The participants were required to rate the four training stimuli and the four novel compounds of each discrimination, totalizing 16 testing trials, presented in random order.

2.1.4. Data analysis

For each participant, the mean predictive ratings to each stimulus were averaged across the two discriminations. Since the ratings for the discrimination involving aging were expected to vary between 0 and −5 and for the discrimination involving rejuvenation to vary between 0 and 5, the former were multiplied by minus one prior to analysis.

Stimuli were classified according to two criteria and the resulting classes were included as factors in the ANOVA. All stimuli were classified according to their “outcome area” in “positive” (above the main diagonal in Fig. 4a) and “neutral” (below the main diagonal in Fig. 4a). Novel test stimuli were also classified according to stimulus type in “central” (a3b2 and a2b3) and “distal” (a1b4 and a4b1).

Separate ANOVAs were carried out for trained and novel test cues. Trained stimuli were analyzed through a 2 (outcome area) × 2 (group: integral vs. separable) mixed effects ANOVA. Novel test stimuli were analyzed through a 2 (outcome area) × 2 (stimulus type) × 2 (group: integral vs. separable) mixed effects ANOVA. Two additional 2 (outcome area) × 2 (group) ANOVAs were carried out, one for each level of stimulus type. Of these, the analysis of central stimuli is theoretically the most important. Only the direction hypothesis consistently predicted an outcome area × group interaction for this analysis, with higher responding to a2b3 than a3b2 in group integral. Higher responding to a3b2 than a2b3 in group separable is predicted by the model regardless of what hypothesis is adopted to explain integrality. Thus, we planned to run pairwise comparisons between a3b2 and a2b3 within each group, using the Bonferroni correction for multiple comparisons. Regarding the analysis of distal cues, both hypotheses predict a main effect of outcome area, with higher responding to a1b4 than a4b1 in both groups.

An alpha level of 0.05 was adopted for all tests of significance.

2.2. Results and discussion

Fig. 6 presents the mean predictive ratings for testing stimuli. Both groups learned the discrimination, as indicated by high mean causal judgements for stimuli that were followed by the outcome (a1b2/a3b4) and low mean causal judgements for stimuli that were not followed by the outcome (a2b1/a4b3). These observations were supported by the ANOVA on data from trained cues, which showed a significant main effect of outcome area, $F(1,46) = 1066.08$, $p < .001$, $\eta_p^2 = .96$, whereas the main effect of group and the interaction were not significant.

The pattern of causal judgements for test stimuli was quite similar to the predictions of the direction hypothesis shown in Fig. 5a. The pattern of responding was different for central stimuli and distal cues. With distal cues, both groups gave higher ratings to the positive stimulus (a1b4) than to the neutral stimulus (a4b1), whereas with central cues, only the integral group gave higher ratings to the positive stimulus (a2b3) than to the neutral stimulus (a3b2), with the separable group showing the opposite pattern. The ANOVA on data from test stimuli confirmed these observations, showing a significant group × outcome area × stimulus type interaction, $F(1,46) = 17.45$, $p < .001$, $\eta_p^2 = .28$. This ANOVA also showed significant main effects of group, $F(1,46) = 6.51$, $p < .05$, $\eta_p^2 = .87$, and outcome area, $F(1,46) = 17.45$, $p < .001$, $\eta_p^2 = .28$, and significant interactions of group × outcome area, $F(1,46) = 9.33$, $p < .001$, $\eta_p^2 = .17$, and outcome area × stimulus type, $F(1,46) = 57.6$, $p < .001$, $\eta_p^2 = .56$.

The group × outcome area ANOVA for central stimuli confirmed that the interaction predicted by the direction hypothesis and observed in Fig. 6 was significant, $F(1,46) = 24.01$, $p < .001$, $\eta_p^2 = .96$, whereas neither of the main effects were significant. The planned pairwise comparisons indicated that the rating to a3b2 was significantly higher than to a2b3 in group separable (Bonferroni-corrected $p < .001$) and the rating to a3b2 was significantly lower than to a2b3 in group integral (Bonferroni-corrected $p < .01$).

The group × outcome area ANOVA for distal stimuli revealed a completely different pattern of results, with significant main effects of group, $F(1,46) = 6.94$, $p < .05$, $\eta_p^2 = .13$, indicating higher ratings in group separable than in group integral, and outcome area, $F(1,46) = 113.62$, $p < .001$, $\eta_p^2 = .71$, indicating higher ratings to the stimuli formed by dimensions that were followed by the outcome in training, but a non-significant interaction between these factors. This pattern of results is captured by both the direction and correlation hypotheses (see Fig. 5).

On the other hand, there is one aspect of the results shown in Fig. 6 that is not captured by the consequential regions model, regardless of the hypothesis used to explain integrality: in group integral, causal judgements to stimulus a1b4 were higher than those to stimulus a2b3. As Fig. 5a shows, the direction hypothesis predicts the opposite effect (i.e., a1b4 < a2b3) and the correlation hypothesis predicts no difference. A post hoc paired-samples *t*-test indicated that the difference between ratings to a1b4 and a2b3 in group integral was not significant, $t(23) = -1.715$, $p = 0.10$, even without correction for multiple comparisons, which is more in line with the null effect predicted by the correlation hypothesis.

Note that other observed differences in mean rating between distal and central test stimuli (a1b4 > a2b3 in group separable; a4b1 < a3b2 in both groups) are correctly predicted by both hypotheses. One possibility is that all differences between distal and central test stimuli arose as the result of categorization, a phenomenon that is not covered by the rational model of generalization. For example, if people were using a diagonal bound to solve the task shown in Fig. 4a, then simple distance-to-bound effects could account for all observed differences between distal and central test stimuli. We will discuss this possibility in more detail in the General Discussion. For now, note that learning of a category bound could not explain the most interesting result of an interaction between condition and stimulus for the central test stimuli. Learning of a diagonal category bound leads to the prediction of higher responding to a3b2 than a2b3 in both integral and separable conditions.

To summarize, we found an interaction between group and outcome area for the central test stimuli that is in agreement with the results found by Lachnit (1988). That is, presenting a combination of dimensional levels previously associated with an outcome produces an effect analogous to summation when separable dimensions are used, but the opposite effect (more responding to a control combination than to the summation combination) is found with integral dimensions. This result is predicted by the direction hypothesis implemented by Soto et al. (2014).

The results show that the effect found by Lachnit was not an artifact of the specific stimuli used in each of the conditions of his experiment, as we made an attempt to avoid differences in generalization between the dimensions by using each dimension in both integral and separable conditions. Furthermore, the effect found by Lachnit is not restricted to simple forms of associative learning, but can be found in a task involving causal inference.

The direction hypothesis incorrectly predicted much lower responding to a1b4 in group integral than what was observed in the data. For this reason, it seems necessary to gather further

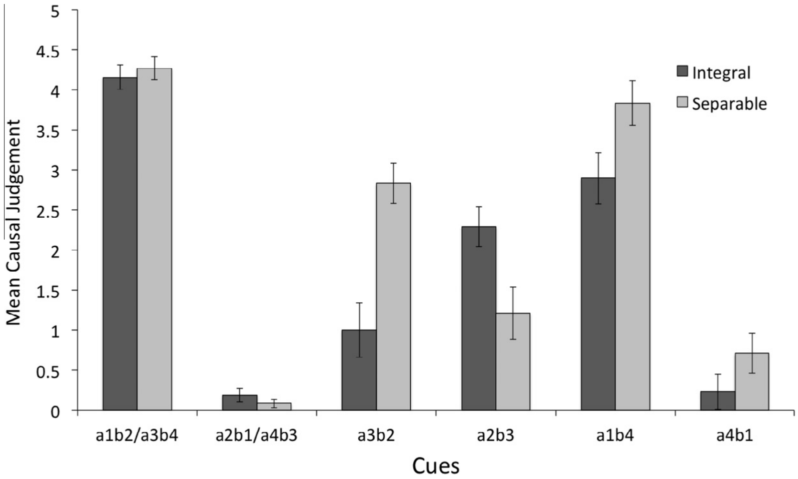


Fig. 6. Mean predictive ratings assigned to each experimental stimulus during testing in Experiment 1. Error bars are standard errors of the mean.

evidence as to whether the direction hypothesis can better capture data from dimensional generalization experiments than the correlation hypothesis. Experiment 2 was carried out with this goal in mind.

3. Experiment 2

Experiment 2 also exploits the analogy proposed by Lachnit (1988) between stimulus components in compound generalization and dimensional values in dimensional generalization. In a biconditional discrimination (Harris & Livesey, 2008; Harris, Livesey, Gharaei, & Westbrook, 2008; Saavedra, 1975) participants are presented with two compounds, AB and CD, followed by an outcome, and two compounds, AC and BD, followed by no outcome. Because each individual stimulus is followed by outcome and no outcome the same number of times, learning such a discrimination requires nonlinear stimulus representations.

When we replace stimulus components A, B, C and D for dimensional values a1, a2, b1 and b2, the biconditional discrimination takes the form shown in Fig. 7a. Fig. 7b shows that, according to the direction hypothesis, there is a very simple configuration of consequential regions that solves this discrimination when stimuli vary in integral dimensions. Here, one diagonal consequential region associated with the outcome generates a1b1 and a2b2, whereas a different diagonal consequential region associated with no outcome generates a1b2 and a2b1. There is no equivalent simple configuration with separable dimensions. In this case, the model has no alternative but to consider hypotheses in which specific stimuli are associated with single consequential regions, as in the example shown in Fig. 7c. The same is true for the case of integral dimensions and the correlation hypothesis: because consequential regions are aligned to the axes, the configuration in Fig. 7b is impossible.

For this reason, we expect that if the direction hypothesis is true, learning of the biconditional discrimination should be faster with integral dimensions. Fig. 8 shows the predictions of the model for this experimental design, both when the direction (panel a) and the correlation hypotheses (panel b) are implemented (for details about these simulations, see the supplementary material). It can be seen that the direction hypothesis predicts consistently better performance with integral dimensions across all training sessions. On the other hand, the correlation hypothesis does not predict a consistent difference between integral and separable dimensions in the learning speed of the biconditional discrimination.

Nonlinear discriminations such as that shown in Fig. 7 have been widely studied in Pavlovian conditioning (e.g., Harris et al., 2008; Saavedra, 1975), human causal learning (e.g., Harris & Livesey, 2008; Shanks, Charles, Darby, & Azmi, 1998) and category learning (e.g., Blair & Homa, 2001; Medin & Schwanenflugel, 1981). However, we know of no previous study directly comparing learning of a nonlinear discrimination with integral and separable stimuli.

Experiment 2 complements Experiment 1 in a very important way. Experiment 1 tested an effect analogous to summation in dimensional generalization. The summation effect is typically considered evidence of linear processing in compound generalization, and this kind of result was found only with separable dimensions. Thus, the previous experiment gave evidence for a relation between linear processing and separability, as proposed by several researchers (Lachnit, 1988; Melchers et al., 2008; Myers et al., 2001) and formalized by the consequential regions model (Soto et al., 2014). With integral dimensions, the results suggested an effect that was opposite to summation, for which it is impossible to find an analogy in the compound generalization literature. Thus, the previous experiment failed to link non-linear processing and integrality. On the other hand, the biconditional discrimination tested in the present experiment is usually considered evidence of non-linear processing in compound generalization. If the analogous discrimination tested here was solved faster with stimuli varying in integral dimensions, then there would be evidence linking separability with linear processing and integrality with non-linear processing.

3.1. Method

3.1.1. Participants

A total of 32 undergraduate psychology students at the University of Talca, Chile ($n = 16$) and at the University of Rosario, Colombia ($n = 16$) participated in the experiment for course credit. They had a mean age of 18.5 years ($S = 1.2$). They were tested individually and had no previous experience in similar research.

3.1.2. Materials and procedure

The same causal learning procedure and strategy for stimulus construction as in Experiment 1 were used. Training involved two simultaneous biconditional discriminations. In one of the discriminations there were two stimuli that always were followed by an aging reaction (a1b1 and a2b2) and two stimuli that were followed by a neutral reaction (a1b2 and a2b1). The other discrimination was identical to the first, but involved neutral versus rejuvenation outcomes and a different set of stimuli. The participants received 120 trials, including 15 presentations of each of the 8 trial types, with the restriction that each trial type was presented once in each block of eight trials in a random order within the block.

The participants were randomly assigned to one of the four sub-groups, defined in the same way as in Experiment 1. In each subgroup, the assignment of specific stimuli to the outcome and no outcome consequences and the assignment of one or another set of stimuli to the aging and rejuvenation outcomes were counterbalanced.

3.1.3. Data analysis

The effects of interest were assessed by computing the mean percent of correct responses in each block of training involving the 8 trial types. The statistical reliability of the effects was examined by a 15 (training block) × 2 (group: integral vs. separable) mixed design ANOVA.

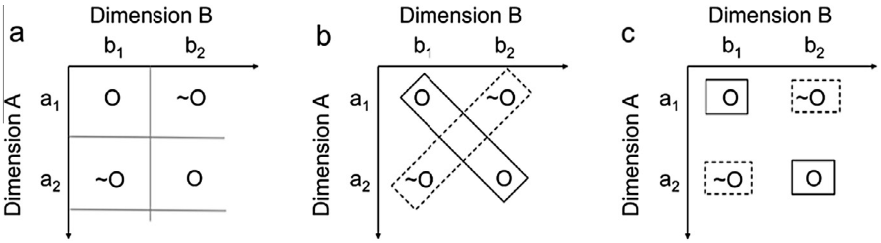


Fig. 7. Task used in Experiment 2 (a) and examples of regions heavily weighted by the consequential regions model implementing the direction hypothesis, both for integral dimensions (b) and for separable dimensions (c).

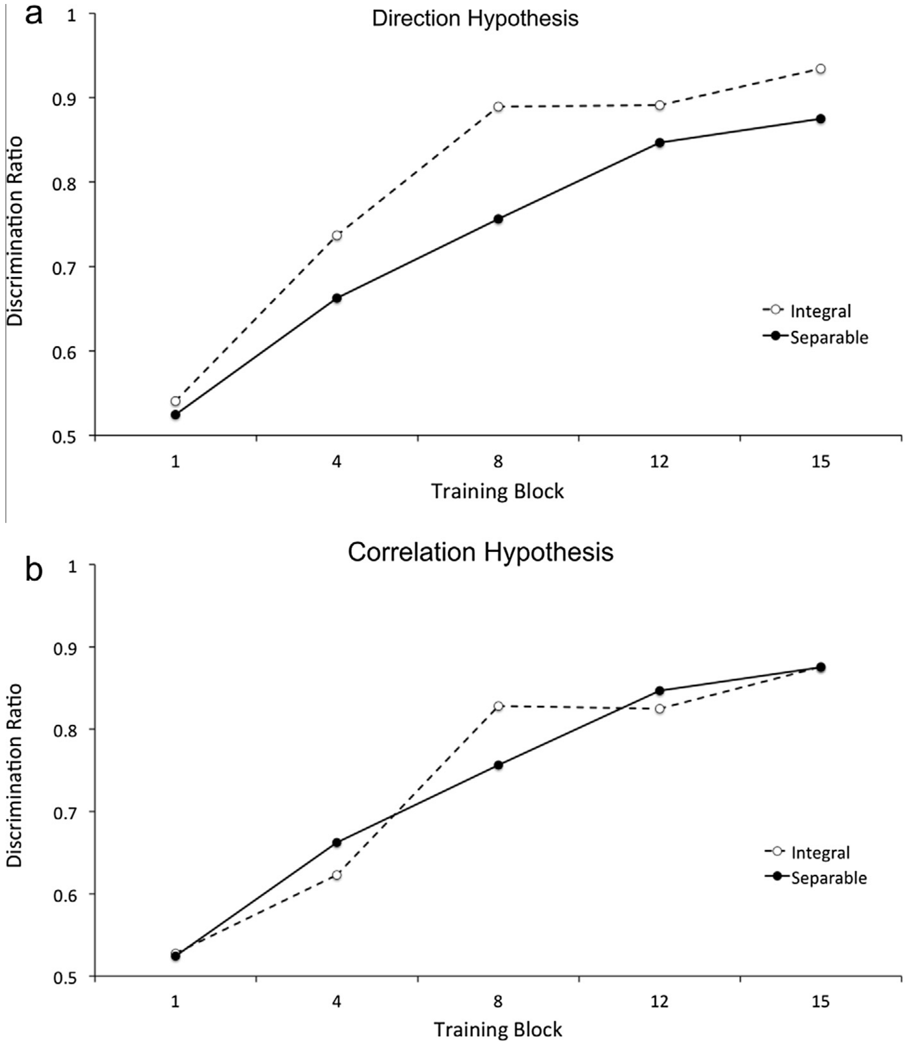


Fig. 8. Predictions of the direction (a) and correlation (b) hypotheses of accuracy across training in the biconditional discrimination of Experiment 2.

3.2. Results and discussion

Fig. 9 presents the percent of correct responses over training for the integral and separable groups. It is apparent from the figure that although in both groups the participants learned the biconditional discrimination, those belonging to group integral learned it faster than those in group separable. This was confirmed by reliable main effects of block (F (14,420) = 16.840; $p < 0.001$; $\eta_p^2 = 0.36$) and group (F (1,30) = 10.542; $p = 0.003$; $\eta_p^2 = 0.260$).

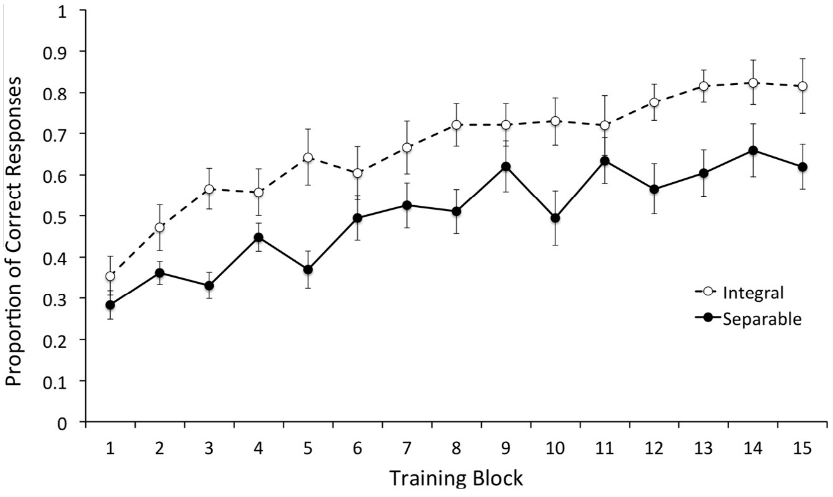


Fig. 9. Mean percent of correct responses over the 15 blocks of training of Experiment 2. Error bars are standard errors of the mean.

In summary, the direction of the differences among the groups was in agreement with the predictions of the direction hypothesis, which seems to provide a better way to distinguish between integrality and separability than the correlation hypothesis. Furthermore, the present results complement those of Experiment 1 in that they establish a link between non-linear stimulus processing in compound generalization and integrality in multidimensional generalization.

4. General discussion

Experiments 1 and 2 were designed as a test for the notion of integrality proposed by the direction hypothesis, implemented in a recent unified rational theory of stimulus generalization that explains both compound and dimensional generalization (Soto et al., 2014). The predictions of the direction hypothesis were contrasted to those of the correlation hypothesis, an alternative way to distinguish between separable and integral dimensions within consequential regions theory (Shepard, 1987, 1991).

Experiment 1 showed that an effect analogous to summation is found in dimensional generalization with separable dimensions, but the opposite effect is found with integral dimensions. These results are the first demonstration of this effect in causal learning and confirm the findings of Lachnit (1988) in Pavlovian conditioning. Furthermore, Experiment 1 ruled out the possibility that Lachnit's findings were due to using different dimensions for the integral and separable stimuli.

Experiment 2 showed that an analogue to a biconditional discrimination is solved faster by people when stimuli vary in integral dimensions than when stimuli vary in separable dimensions.

The results from both experiments were in line with the predictions of the direction hypothesis, but inconsistent with the predictions of the correlation hypothesis. Because assuming the correctness of the direction hypothesis was crucial in the model of Soto et al. (2014) to explain compound generalization within the framework of consequential regions theory, the experiments also support this model as a unified theory of stimulus generalization.

These results have implications both for fields studying dimensional generalization, such as instrumental conditioning and stimulus identification, and for fields studying compound generalization, such as Pavlovian conditioning and human causal learning. Regarding dimensional generalization, the present results show that the best way to conceptualize the distinction between integral and separable dimensions in consequential regions theory is the following: separable dimensions are special directions in stimulus space along which natural kinds extend, whereas integral dimensions are those for which natural kinds extend in any direction of stimulus space.

Regarding compound generalization, previous work has shown that assuming the direction hypothesis is crucial to explain compound generalization phenomena using consequential regions theory (Soto et al., 2014). The results reported here lend support to this assumption, showing that it is also crucial to explain dimensional generalization phenomena.

The results reported here also show that in some occasions an interesting link arises between dimensional and compound generalization: when a design from compound generalization is translated into a design for dimensional generalization by replacing discrete stimulus components with dimensional values, experimental results that are analogous to linear processing are found with separable dimensions and experimental results that are analogous to nonlinear processing are found with integral dimensions. To the best of our knowledge, the only theory currently capable of capturing such relations between component and dimensional interactions is the theory of Soto et al. (2014).

4.1. Levels of integrality in the rational theory of generalization

The main goal of a rational analysis of integrality is to understand why integral dimensions produce observed patterns of generalization. From a mechanistic perspective, a related question that can inform such rational analysis is this: How are integral dimensions represented and processed?

Smith and Kemler (1978) distinguished two possibilities regarding the representation of integral dimensions. Integral dimensions could be non-primary axes, having no special status compared to any other direction in psychological space, or they could be primary axes, being perceived hollistically but also sustaining a less preferred mode of processing in terms of component parts. A number of recent studies suggest that integral dimensions are primary axes, being psychologically meaningful despite the fact that they are usually processed in a holistic fashion (Foard & Kemler-Nelson, 1984; Grau & Kemler-Nelson, 1988; Jones & Goldstone, 2013; Melara & Marks, 1990; Melara, Marks, & Potts, 1993).

Kemler-Nelson (1993) concluded from a literature review that integral dimensions are real psychological dimensions (a similar conclusion was reached more recently by Jones & Goldstone, 2013) that are usually processed holistically, but in a small proportion of occasions are processed analytically. In this perspective, integrality and separability are two ends of a continuum, a conclusion on which most researchers would agree (e.g., Garner, 1974; Shepard, 1991; Smith & Kemler, 1978). On the other hand, the two mechanisms of holistic and analytic processing are a pure dichotomy. What determines where in the integrality–separability continuum lies a particular pair of dimensions is the extent to which their default mode of processing (holistic or analytic) is combined with instances of the opposite mode of processing.

The finding that integral dimensions can sometimes be primary axes seems problematic for the directionality hypothesis because, if natural kinds extend in any direction of psychological space, then there seems to be no reason to represent special dimensions at all. On the contrary, representing the dimensions might hinder learning about natural classes, because this would require an additional step of integration of information from separate representations. On the other hand, the correlation hypothesis seems to relate better to the idea of privileged primary axes. If natural kinds extend in specific directions of psychological space, but their extension is correlated, then there is reason to represent such directions as special dimensions and, at the same time, produce generalization gradients that are similar in any direction of space.

How can we reconcile the finding of primary integral dimensions with the directionality hypothesis, which was supported by our experimental results? In our model, the distinction between integral and separable dimensions is implemented as two different hypotheses spaces on the possible values taken by a parameter, θ, describing the orientation of consequential regions with respect to the dimensions of stimulus space. With separable dimensions, only hypotheses in which consequential regions are aligned with the axes of the space are considered ($\theta = 0°$). With integral dimensions, hypotheses in which consequential regions can extend in any direction of the space are considered (any θ from 0° to 360°). These two hypothesis spaces are two extremes in a continuum, just as the concepts of pure integrality and separability. As explained in the supplementary material, assigning a prior probability $v < 1$ to hypotheses in which $\theta = 0°$, and a uniform probability to all other hypotheses about θ allows the model to produce generalization patterns intermediate between integrality and separability, as shown in Fig. 10. Visual inspection of the gradients in Fig. 10 suggests that gradients predicted by the correlation hypothesis for integral dimensions are similar to those predicted by the direction hypothesis as intermediate cases between pure integrality and

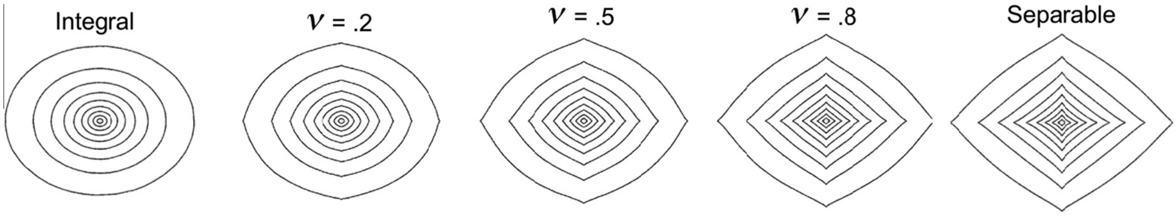


Fig. 10. Contour plots of generalization gradients predicted by the direction hypothesis for several cases intermediate between pure integrality (left) and pure separability (right).

pure separability (compare $v = 0.2$ and $v = 0.5$ in Fig. 10 with Fig. 1c). Such intermediate cases could involve natural classes that can extend in any direction of space with similar probability, but which are slightly more likely to extend in the direction of the main axes. These cases would provide a rational explanation to the findings of psychologically-privileged integral dimensions that is clearly related to Kemler-Nelson's (1993) mechanistic explanation in terms of combined modes of processing. That is, the proportion of holistic and analytic processing of two dimensions might be a function of the probability that natural classes extend in any direction of space versus in the direction of psychological dimensions.

4.2. Learning integral and separable dimensions

The version of our rational model discussed in the previous section can be easily extended to make predictions about what kinds of experiences with natural kinds should lead to learning of separable or integral dimensions. Although the gradients displayed in Fig. 10 were obtained by setting the parameter v to specific values, it is also possible to define a prior distribution on this parameter and learn, through Bayesian inference, what is the appropriate metric for a given pair of dimensions given past experience with categories varying in those dimensions.

In fact, this strategy was followed by Austerweil and Griffiths (2010) to predict what kinds of experiences should induce generalization patterns typical of integral and separable dimensions. As in the model of Soto et al. (2014), Austerweil and Griffiths assumed the direction hypothesis to distinguish between integral and separable dimensions. In an experiment testing their predictions, the authors found that after learning several categories aligned with the dimensions of width and height of a rectangle, generalization gradients along those dimensions became similar to those expected from separable dimensions. On the other hand, after learning categories with an extension indifferent to the original dimensions, generalization gradients became similar to those expected from integral dimensions.

This framework could also help to explain what conditions lead to a change in metric for a given pair of dimensions. A change from integrality to separability, or vice-versa, might be more or less difficult depending on a number of conditions. For example, the differentiation of novel dimensions that appear to be integral can happen after limited experience in a single categorization task (e.g., Goldstone & Steyvers, 2001; Soto & Ashby, 2015), and any direction in space can serve as a basis for such learned dimensions (Folstein, Gauthier, & Palmeri, 2012). On the other hand, similar experience with traditional integral dimensions can fail to produce evidence of differentiation (Goldstone, 1994), particularly when the relevant direction in space is misaligned with the integral axes (Foard & Kemler-Nelson, 1984). A possible explanation from our rational theory is that integral dimensions should be more difficult to differentiate when there is more previously accumulated evidence for their integrality. This previously accumulated evidence can be modeled as a distribution over the parameter v that is biased toward the integral metric. The stronger this bias is, the more new evidence is required to infer a separable metric.

Both of our experiments involved training with diagonally-oriented categories of stimuli (see Figs. 4c and 7b). Given the evidence that arbitrary directions in an integral space can be the basis for newly-learned dimensions (Folstein et al., 2012; Soto & Ashby, 2015), it could be argued that training in our tasks led to learning of diagonally-oriented separable dimensions in the integral conditions. If that was the case, then our results could be explained as arising from differently-oriented separable dimensions instead than from the integrality/separability of dimensions. There are two reasons that make this explanation of our results unlikely. First, learning of diagonal dimensions would be quite difficult with our stimuli. Traditional integral dimensions like those used here are privileged directions in stimulus space (Kemler-Nelson, 1993) and, unlike completely novel dimensions, do not seem to be differentiated after categorization training (e.g., Goldstone, 1994). Second, dimension differentiation seems to require much more extensive categorization training than that used here, both in terms of number of stimuli per category and duration of training (see Folstein et al., 2012; Goldstone & Steyvers, 2001; Soto & Ashby, 2015), even when completely novel dimensions are trained.

4.3. What is the best mechanistic explanation for our results?

The consequential regions theory explored here provides a rational analysis of generalization (Anderson, 1990). This type of explanation is fundamentally different from mechanistic theories of compound generalization, such as computational models of associative learning (e.g., Harris, 2006; McLaren & Mackintosh, 2002; Pearce, 1987, 1994; Wagner, 2003) and mechanistic theories of category learning (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Kruschke, 1992; Nosofsky, 1984). An open question is which of these models provides the best mechanistic explanation of our results.

The results of Experiment 2 could be explained straightforwardly by contemporary models of associative learning (e.g., Harris, 2006; McLaren & Mackintosh, 2002; Wagner, 2003). In these models, representing dimensional levels just as discrete stimulus components and assuming that integral components determine a higher level of nonlinear processing than separable components leads to the correct prediction: better learning of the biconditional discrimination with integral than separable dimensions.

Such straightforward application of associative models presents more difficulties in explaining the results from Experiment 1. One might assume that the experiment involves discrimination between 4 compounds formed by unique and common elements, a1b2X+, a3b4X+, a2b1X−, and a4b3X−, where X represents a constant contextual cue. Assuming an error-driven learning rule (e.g., Rescorla & Wagner, 1972), the uniquely reinforced elements a1, b2, a3 and b4 should become substantially excitatory, the partially reinforced X element becomes moderately excitatory, and the uniquely nonreinforced elements, a2, b1, a4 and b3, become inhibitory. In testing, differential responding to the novel compounds a3b2X and a2b3X would depend on how much excitation receives the former from a3 and b2, and how much inhibition from a2 and b3 receives the latter. If the integral compounds are processed in a more nonlinear fashion than the separable compounds, it is expected a greater loss of excitation (i.e., less responding to a3b2) and greater loss of inhibition (more responding to a2b3) in the integral that in the separable condition. This might lead to the correct predictions that group separable should show greater responding to a3b2 than group integral and that group integral should show greater responding to a2b3 than group separable. However, this reasoning cannot account for the fact that in the group integral a combination of two dimensional levels never followed by an outcome (a2b3) led to higher causal ratings than a combination of two dimensional levels that were always followed by an outcome (a3b2).

Other mechanistic models of associative learning can describe compound and dimensional generalization through a unified theory (e.g., Blough, 1975; Ghirlanda, 2005). However, such models were developed to explain only unidimensional generalization phenomena, and they cannot capture basic results from multidimensional generalization, such as the different shapes of generalization gradients along separable and integral dimensions (e.g., Soto & Wasserman, 2010). Thus, it seems like mechanistic models of associative learning will require important modifications before they can account for results such as those presented here.

The rational theory of Soto et al. (2014) suggests that the inferential task imposed by compound generalization is simply an extension of the task of dimensional generalization, in which more than a single stimulus can be presented at the same time. Thus, the inferential task of compound generalization contains all elements of the task of dimensional generalization, plus the additional problem of inferring whether different stimulus components have been generated by the same latent cause. What this suggests is that a successful model of compound generalization should start by implementing mechanisms to explain dimensional generalization. The development of such a model is possibly one of the most important theoretical challenges ahead of us in the field of stimulus generalization.

Now consider how models of category learning might account for the present results. Exemplar models of categorization have been used in the past to explain differences between integral and separable dimensions in learning of categorization tasks (Nosofsky et al., 1994; Nosofsky & Palmeri, 1996), based on the assumption that selective dimensional attention is more easily deployed to separable than integral dimensions. The same assumption could be used to explain the results of Experiment 1. Selective attention to one dimension essentially collapses the psychological space onto that dimension, making stimuli that share a value in the dimension very similar to one another. Generalization between such similar stimuli could lead to the pattern found for separable dimensions. Without selective attention, generalization from the closest training stimuli to the test stimuli (e.g., from a1b2 and a3b4 to the test stimulus a2b3) should favor a pattern of results similar to that found for integral dimensions.

Differences in selective attention are less likely to explain the results of Experiment 2. In this case, there is faster learning of the task with integral dimensions, meaning that the *lack* of selective attention would have led to better performance. It is unclear why selective attention to separable dimensions would be deployed in a way that hinders performance in a task.

Exemplar theory could explain the results observed in Experiment 2 as arising from the metric of the spatial space used to represent integral and separable dimensions. If it is assumed that integral dimensions are associated with an Euclidean metric and separable dimensions with a city-block metric, then there should be more intra-class generalization with integral than with separable dimensions in Experiment 2, facilitating learning with integral dimensions relative to separable dimensions. However, this is not an assumption that exemplar theorists have consistently made (see Maddox & Ashby, 1998), and thus the theory can explain the present results only in a post-hoc fashion. Furthermore, it is unclear whether this would constitute a proper mechanistic explanation. As mentioned in the introduction, different metrics of spatial models work as a re-description of the shape of generalization gradients. Taking this into account, an explanation of generalization data that resorts only to the metric in a spatial model is circular.

Regarding the predictions of prototype (e.g., Smith & Minda, 1998) and decision-bound (e.g., Ashby & Maddox, 1993) theories of categorization, both of them would make the correct prediction for the integral condition of Experiment 1, as they would divide the space into "outcome" and "no outcome" regions through a diagonal bound and predict higher judgments of causality to a2b3 than to a3b2. However, these theories would fail to make the opposite prediction for separable dimensions, where a similar diagonal bound should be learned.

Prototype theory cannot explain learning of non-linearly separable categorization tasks such as that used in Experiment 2 (Medin & Schwanenflugel, 1981). Although decision bound theory could explain learning of such structures through non-linear bounds, it is unclear how this theory could explain the observed difference in learning speed between the integral and separable conditions.

Overall, it seems clear that exemplar theories of category learning are in a better position than competing models to explain the results presented here. On the other hand, it is unclear whether these models can be used to explain the results of studies in compound generalization that contemporary models of associative learning can explain. One of the appeals of the rational model developed by Soto et al. (2014) is that it allows to explain both dimensional and compound generalization phenomena within the same framework. We hope that work with the rational theory of generalization can provide insights on how to extend mechanistic models in this direction.

5. Conclusion

In sum, a consequential regions model in which the main distinction between integrality and separability is the direction of consequential regions can: (1) explain data from dimensional generalization experiments better than Shepard's correlation hypothesis, as shown by the present experiments, (2) explain data from compound generalization experiments, as shown by Soto et al. (2014), (3) produce generalization gradients intermediate between pure integrality and separability and explain why a pair of dimensions lies at a particular point in the integrality–separability continuum, (4) provide insights about the conditions under which separable or integral dimensions can be learned from experience, and (5) provide insights about how to extend current mechanistic models of learning so that they can explain both dimensional and compound generalization phenomena.

Acknowledgments

E.H. Vogel was supported by Grants No. 1120265 from FONDECYT and No. 74150009 from CONICYT. Andrés M. Pérez-Acosta was supported by Foundation "Rubén Ardila" for the Advancement of Psychology (Colombia). Fabian Soto was supported by a Sage Center Junior Research Fellowship.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.07.001.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442–481. http://dx.doi.org/10.1037/0033-295X.105.3.442.

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

Austerweil, J. L., & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 73–78).

Aydin, A., & Pearce, J. M. (1995). Summation in autoshaping with short- and long-duration stimuli. *Quarterly Journal of Experimental Psychology, 48B*(3), 215–234. http://dx.doi.org/10.1080/14640749508401449.

Blair, M., & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition, 29*(8), 1153–1164.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 104*(1), 3–21. http://dx.doi.org/10.1037/0097-7403.1.1.3.

Collins, D. J., & Shanks, D. R. (2006). Summation in causal learning: Elemental processing or configural generalization? *Quarterly Journal of Experimental Psychology, 59*(9), 1524–1534. http://dx.doi.org/10.1080/17470210600639389.

Cross, D. V. (1965). Metric properties of multidimensional stimulus generalization. In D. I. Mostofsky (Ed.), *Stimulus generalization* (pp. 72–93). Palo Alto, CA: Stanford University Press.

Dunn, J. C. (1983). Spatial metrics of integral and separable dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 9*(2), 242–257.

Fernbach, P. M., & Sloman, S. A. (2011). Don't throw out the Bayes with the bathwater. *Behavioral and Brain Sciences, 34*(4), 198–199. http://dx.doi.org/10.1017/S0140525X11000264.

Foard, C. F., & Kemler-Nelson, D. G. (1984). Holistic and analytic modes of processing: The multiple determinants of perceptual analysis. *Journal of Experimental Psychology: General, 113*(1), 94–111.

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: Not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4). http://dx.doi.org/10.1037/a0025836. 807–802.

Garner, W. R. (1974). *The processing of information and structure*. New York: Lawrence Erlbaum Associates.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1*(3), 225–241. http://dx.doi.org/10.1016/0010-0285(70)90016-2.

Ghirlanda, S. (2005). Retrospective revaluation as simple associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 31*(1), 107–111.

Ghirlanda, S. (2015). On elemental and configural models of associative learning. *Journal of Mathematical Psychology, 64–65*, 8–16. http://dx.doi.org/10.1016/j.jmp.2014.11.003.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour, 66*, 15–36. http://dx.doi.org/10.1006/anbe.2003.2174.

Glautier, S. (2004). Asymmetry of generalization decrement in causal learning. *Quarterly Journal of Experimental Psychology, 57B*(4), 315. http://dx.doi.org/10.1080/02724990344000169.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*(2), 178–200.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130*(1), 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*(4), 347–370.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138*(3), 415–422. http://dx.doi.org/10.1037/a0026884.

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology, 51*(1), 79–88. http://dx.doi.org/10.1037/h0046219.

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review, 113*(3), 584–605. http://dx.doi.org/10.1037/0033-295X.113.3.584.

Harris, J. A., & Livesey, E. J. (2008). Comparing patterning and biconditional discriminations in humans. *Journal of Experimental Psychology: Animal Behavior Processes, 34*(1), 144–154. http://dx.doi.org/10.1037/0097-7403.34.1.144.

Harris, J. A., Livesey, E. J., Gharaei, S., & Westbrook, R. F. (2008). Negative patterning is easier than a biconditional discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 34*(4), 494–500. http://dx.doi.org/10.1037/0097-7403.34.4.494.

Jones, M., & Goldstone, R. L. (2013). The structure of integral dimensions: Contrasting topological and Cartesian representations. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 111–132. http://dx.doi.org/10.1037/a0029059.

Kehoe, E. J., Horne, A. J., Horne, P. S., & Macrae, M. (1994). Summation and configuration between and within sensory modalities in classical conditioning of the rabbit. *Animal Learning and Behavior, 22*(1), 19–26. http://dx.doi.org/10.3758/BF03199952.

Kemler-Nelson, D. G. (1993). Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception and Performance, 19*(5), 1105–1113.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44.

Lachnit, H. (1988). Convergent validation of information processing constructs with Pavlovian methodology. *Journal of Experimental Psychology: Human Perception and Performance, 14*(1), 143–152. http://dx.doi.org/10.1037/0096-1523.14.1.143.

Maddox, W. T., & Ashby, F. G. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 301–321.

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 85–105). Cambridge, MA: MIT Press.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30*(3), 177–200. http://dx.doi.org/10.3758/BF03192828.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*(5), 355–368.

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*(2), 398–414.

Melara, R. D., Marks, L. E., & Potts, B. C. (1993). Primacy of dimensions in color perception. *Journal of Experimental Psychology: Human Perception and Performance, 19*(5), 1082–1104.

Melchers, K. G., Shanks, D. R., & Lachnit, H. (2008). Stimulus coding in human associative learning: Flexible representations of parts and wholes. *Behavioural Processes, 77*(3), 413–427. http://dx.doi.org/10.1016/j.beproc.2007.09.013.

Monahan, J. S., & Lockhead, G. R. (1977). Identification of integral stimuli. *Journal of Experimental Psychology: General, 106*(1), 94–110. http://dx.doi.org/10.1037/0096-3445.106.1.94.

Myers, K. M., Vogel, E. H., Shin, J., & Wagner, A. R. (2001). A comparison of the Rescorla-Wagner and Pearce models in a negative patterning and a summation problem. *Animal Learning and Behavior, 29*(1), 36–45. http://dx.doi.org/10.3758/BF03192814.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 104–114.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43*, 25–53. http://dx.doi.org/10.1146/annurev.ps.43.020192.000325.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*(3), 352–369.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review, 3*(2), 222–226.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94*(1), 61–73. http://dx.doi.org/10.1037//0033-295X.94.1.61.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101*, 587–607. http://dx.doi.org/10.1037/0033-295X.101.4.587.

Rescorla, R. A. (1997). Summation: Assessment of a configural theory. *Animal Learning & Behavior, 25*(2), 200–209. http://dx.doi.org/10.3758/BF03199059.

Rescorla, R. A., & Coldwell, S. E. (1995). Summation in autoshaping. *Animal Learning & Behavior, 23*(3), 314–326. http://dx.doi.org/10.3758/BF03198928.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Prokasy & W. F. Black (Eds.), *Classical conditioning II: Current research and theory* (pp. 126–134). New York: Appleton Century Crofts.

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning & Motivation, 6*(3), 314–326. http://dx.doi.org/10.1016/0023-9690(75)90012-0.

Shanks, D. R., Charles, D., Darby, R. J., & Azmi, A. (1998). Configural processes in human associative learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1353–1378.

Shanks, D. R., Lachnit, H., & Melchers, K. G. (2008). Representational flexibility and the challenge to elemental theories of learning: Response to commentaries. *Behavioural Processes, 77*(3), 451–453. http://dx.doi.org/10.1016/j.beproc.2007.09.005.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323. http://dx.doi.org/10.1126/science.3629243.

Shepard, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. In D. I. Mostofsky (Ed.), *Stimulus generalization* (pp. 94–110). Palo Alto, CA: Stanford University Press.

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. Pomerantz & G. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53–71). Washington, DC: American Psychological Association.

Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology, 10*(4), 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*(6), 1411–1436. http://dx.doi.org/10.1037/0278-7393.24.6.1411.

Soto, F. A., & Ashby, F. G. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition, 139*, 105–129.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review, 121*(3), 526–558. http://dx.doi.org/10.1037/a0037018.

Soto, F. A., Vogel, E. H., Castillo, R. D., & Wagner, A. R. (2009). Generality of the summation effect in human causal learning. *Quarterly Journal of Experimental Psychology, 62*(5), 877–889. http://dx.doi.org/10.1080/17470210802373688.

Soto, F. A., & Wasserman, E. A. (2010). Integrality/separability of stimulus dimensions and multidimensional generalization in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 36*(2), 194–205. http://dx.doi.org/10.1037/a0016560.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629–640. http://dx.doi.org/10.1017/S0140525X01000061.

Wagner, A. R. (2003). Context-sensitive elemental theory. *Quarterly Journal of Experimental Psychology, 56B*(1), 7–29. http://dx.doi.org/10.1080/02724990244000133.

Wagner, A. R., & Vogel, E. H. (2008). Configural and elemental processing in associative learning: Commentary on Melchers, Shanks and Lachnit. *Behavioural Processes, 77*(3), 446–450. http://dx.doi.org/10.1016/j.beproc.2007.09.011.

Whitlow, J. W., & Wagner, A. R. (1972). Negative patterning in classical conditioning: Summation of response tendencies to isolable and configural components. *Psychonomic Science, 27*, 299–301. http://dx.doi.org/10.3758/BF03328970.

Why are some dimensions integral? Testing two hypotheses through causal learning experiments

Soto, F. A., Quintana, G. R., Pérez-Acosta. A.M., Ponce, F. P., Vogel, E. H.

Online Supplementary Material

A. The consequential regions model of Soto et al. (2014)

Here we briefly describe the consequential regions model of Soto et al. (2014), the way in which it was extended to implement the direction and correlation hypotheses and the procedures followed to obtain the simulated results presented in Figures 1, 5 and 8 of the main article.

The model

This model proposes that the task of the animal during an associative learning situation, and the task of humans in a causal learning or contingency learning experiment, is to infer the latent causes that have produced observable stimuli and an outcome. On each trial t the learner observes (1) one or more stimuli indexed by i and described by a vector $\mathbf{x}_{ti} = \{x_{ti1}, ..., x_{tiJ}\}$ of J continuous variables or stimulus dimensions, and (2) an outcome magnitude represented by the scalar r_t. For example, assume that on the first trial of one of our experiments a participant observes stimulus #1, with values 2 in dimension 1 and 4 in dimension 2. This stimulus is represented in the model as $\mathbf{x}_{11} = \{2, 4\}$, as shown in the left part of Figure A1. The participant also observes the outcome "rejuvenation," which is represented by $r_1 = 5$, as shown in the right part of Figure S1.

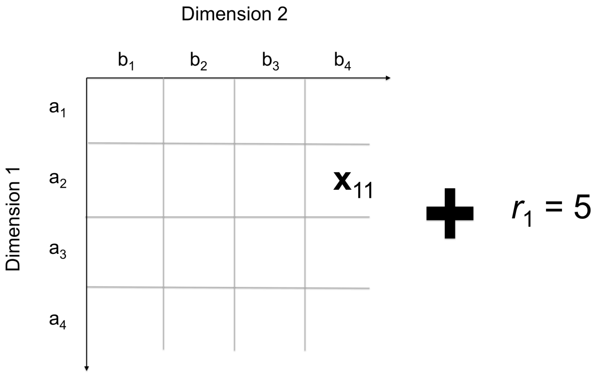


Figure A1. Example of the data observed by a participant on a given trial according to our rational model.

The generative process that produces these data is as follows. On each trial, a number of latent causes are sampled to be active or inactive. These binary variables are collected in the matrix **Z**, in which each column represents a latent variable and each row represents a trial. **Z** is distributed according to the Indian Buffet Process with parameter α (see Griffiths & Ghahramani, 2011):

$$\mathbf{Z} \sim IBP(\alpha)\,. \tag{1}$$

The top of Figure A2 shows three of these binary variables (Z_1, Z_2 and Z_3). Only the shaded variable Z_1 has been sampled to be active during the first trial. Given that this latent variable is active, it generates a number of stimuli n_{kt}, a variable that follows a geometric distribution with parameter π:

$$n_{kt} \sim Geometric(\pi)\,. \tag{2}$$

This distribution favors a small number of stimuli generated by each latent cause on a given trial. Accordingly, the latent cause Z_1 is shown in Figure A2 generating a single stimulus, $\mathbf{x}_{11}$. Any stimulus generated by the latent cause Z_1 has a value in each of the dimensions of psychological space. In the present example, there are two values: one for each dimension shown in Figure A2. The specific values that a stimulus can obtain, which are equivalent to the position of the stimulus in space, are constrained by the consequential region associated with a latent cause. In Figure A2, the consequential region associated with Z_1 is represented by the shaded rectangle. Note that the consequential regions associated with different latent variables can overlap in space.

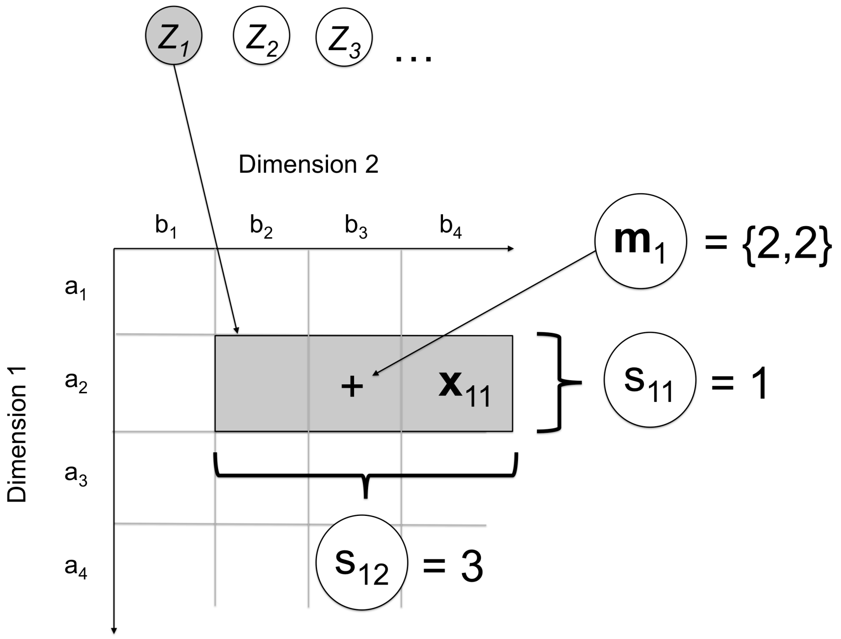


Figure A2. Schematic representation of the process generating stimuli with values in a number of dimensions in the model of Soto et al. (2014).

Two vectors of parameters are associated with each of these rectangular consequential regions. The parameter m_{kj} represents the position of consequential region k in dimension j. In the example shown in Figure A2, there are two of these parameters (one for each dimension) that can be arranged in the vector {2,2}, which determines the position of the rectangle in space. The parameter s_{kj} represent the size of consequential region k in dimension j. In the example shown in Figure A2, there are two of these parameters (one per dimension), representing the width (equal to 3 units) and height (equal to 1 unit) of the rectangular consequential region.

Stimuli are generated by just sampling from all possible values inside an active consequential region. Thus, the likelihood of sampling stimulus $\mathbf{x}_{ti}$ from the consequential region c_k is given by a uniform distribution:

$$x_{tij} \sim Uniform\left(m_{kj} - \frac{s_{kj}}{2}, m_{kj} + \frac{s_{kj}}{2}\right). \tag{3}$$

The model also assumes the existence of a *consequence distribution* over the possible positions and sizes of regions, which is shared by all latent causes. The position parameter is normally distributed:

$$m_{kj} \sim Normal\left(\mu_m, \sigma_m^2\right) \tag{4}$$

The size parameter is uniformly distributed:

$$s_{kj} \sim Uniform\left(a, b\right) \tag{5}$$

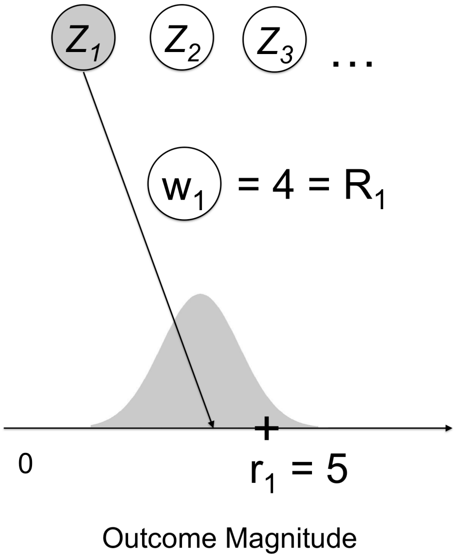


Figure A3. Schematic representation of the process generating outcomes of a certain magnitude in the model of Soto et al. (2014).

Figure A3 illustrates the process that generates the magnitude of an outcome. Each latent cause *k* has associated with it a weight parameter w_k that represents its contribution to the outcome. In the example shown in Figure A3, the parameter w_1 associated with the active latent cause Z_1 has a value of 4. The value of w_k is normally distributed:

$$w_k \sim Normal\left(\mu_w, \sigma_w^2\right) \qquad (6)$$

On any given trial, one or more latent causes may be active. The strength with which all latent causes together contribute to the outcome is the sum of their individual weights w_k:

$$R_t = \sum_{k=1}^{K} z_{tk} w_k \qquad (7)$$

In Figure A3, because the only active latent cause during trial 1 is Z_1, the value of R_1 is the same as the weight associated with this latent cause, which is 4. The actual magnitude of the outcome in trial *t*, or r_t, is sampled from a normal distribution with mean R_t, and standard deviation σ_t:

$$r_t \sim Normal\left(R_t, \sigma_t^2\right). \qquad (8)$$

This is represented in Figure A3 by a normal distribution centered at 4, from which the actual value of $r_1 = 5$ has been sampled.

Note that the same generative process is assumed for stimuli followed by a specific outcome and by "no outcome" in this model. In the "no outcome" case, latent causes can generate multidimensional stimuli through their consequential regions and also an outcome of magnitude zero.

Implementation of the direction and correlation hypotheses

In consequential regions theory, the difference between integral and separable dimensions is explained as arising from the use of two different hypotheses spaces of regions (Austerweil & Griffiths, 2010; Shepard, 1987, 1991). For both the direction and correlation hypotheses, separable dimensions arise by considering all consequential regions with sides aligned to the axes of stimulus space and all possible sizes of those sides. Equations (4) and (5) from the previous section are enough to describe the hypothesis space of separable dimensions.

The difference between the direction and correlation hypotheses lies in the hypothesis space believed to underlie dimensional integrality. According to the correlation hypothesis, this hypothesis space includes all consequential regions with sides aligned to the axes of stimulus space and with all sides of equal size. Again, Equations (4) and (5) can describe this hypothesis space, with the constraint that $s_{k1} = s_{k2} \ldots = s_{kJ}$. Thus, according to the correlation hypothesis, the hypothesis space for integral dimensions is a sub-space of the one considered for separable dimensions.

According to the direction hypothesis, the hypothesis space includes regions of all possible sizes in each side and also all possible orientations with respect to the main axes of space. Because each region has the same dimensionality as the stimulus space, Equations (4) and (5) are still useful to describe the positions and sizes of regions. However, we must now also consider a parameter that determines degrees of rotation of the region around its mid-point:

$$\theta_k \sim Uniform(0,360) \qquad (9)$$

Thus, according to the direction hypothesis, the hypothesis space for separable dimensions is a sub-space of the hypothesis space for integral dimensions. An interesting possibility arises if we discretize the variable θ_k, allowing it to take only some values between 0 and 360. A probability p is assigned to each of these values and the discretized variable θ_k follows a categorical distribution:

$$\theta_k \sim Categorical(V, \mathbf{p}), \qquad (10)$$

where V is the number of values that θ_k could take and $\mathbf{p}$ is a vector of probabilities $p_1, p_2, \ldots p_V$, one for each of such values. Let ν represent the probability of a value of θ equal to zero and let $(1-\nu)/(V-1)$ be the probability for all other possible values of θ. Then the separable case is obtained with $\nu = 1$, the integral case is obtained with $\nu = 1/V$, and a number of intermediate cases can be obtained for $1/V < \nu < 1$, as illustrated in Figure 10 of the main article.

Inference algorithm and simulation procedures

Inferences in our model are aimed at determining the expected value of the outcome on test trial t, or r_t, given the current observation of the compound of stimuli $\mathbf{x}_{t:}$, and the data observed on previous trials. That is, inference is focused on finding $E(r_t \mid \mathbf{X}, r_{1:t-1}, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is a vector of all variables describing the model's prior, which are fixed in all simulations to the following values $\boldsymbol{\psi} = \{\alpha = 5, \pi = 0.9, \lambda = 0.99, a = 0, b = 5, \mu_m = 0, \sigma_m = 10^{1/2}, \mu_w = 0, \sigma_w = 1, \sigma_r = 0.01^{1/2}\}$, $\mathbf{X}$ is a matrix of observed stimuli (both those observed so far and the current observation), and $r_{1:t-1}$ is a vector of previously observed outcome values. In order to calculate the distribution $p(r_t \mid \mathbf{X}, r_{1:t-1}, \boldsymbol{\psi})$ and from it the expected value of r_t, a number of hidden variables in the model need to be integrated out:

$$p\left(r_t \mid \mathbf{X}, r_{1:t-1}, \psi\right) = \int \sum_{\mathbf{Z}} \sum_{\boldsymbol{\theta}} p\left(r_t, \mathbf{Z}, \mathbf{m}, \mathbf{s}, \mathbf{w}, \boldsymbol{\theta} \mid \mathbf{X}, r_{1:t-1}, \psi\right) d\left(\mathbf{m}, \mathbf{s}, \mathbf{w}\right) \qquad (11)$$

Since this integral is not tractable, we approximate it using a set of L samples $\{r_{t1:L}, \mathbf{Z}_{1:L}, \mathbf{w}_{1:L}, \mathbf{m}_{1:L}, \mathbf{s}_{1:L}, \boldsymbol{\theta}_{1:L}\}$ drawn from the posterior distribution using a Markov Chain Monte Carlo (MCMC) procedure. Our MCMC algorithm involves a combination of Gibbs and Metropolis-Hastings sampling (Gilks, Richardson, & Spiegelhalter, 1996). The general strategy is to use a Gibbs sampler to cycle repeatedly through each variable, sampling them from its posterior distribution conditional on the previously sampled values of all the other variables. In the cases in which the conditional posterior is itself intractable, we use Metropolis-Hastings to approximate sampling from the posterior.

The approximated expected value of r_t is the average of the sampled values of this variable.

A complete description of the inference algorithm can be found in Appendix B of (Soto et al., 2014), which was used without modification for simulations involving separable dimensions. For simulations involving integral dimensions using the correlation hypothesis, the only modification to the algorithm was that a single size s_k was sampled at each iteration for each consequential region. For simulations involving integral dimensions using the direction hypothesis, the only modification to the algorithm was the addition of a step in the Gibbs cycle in order to sample θ (in all other cases, θ was always equal to zero). For this, we used an independence Metropolis-Hastings sampler with the prior distribution defined in Equation (9) as the proposal distribution. Values of θ from 0° to 360° in steps of 15° were considered. At each iteration and for each consequential region in the current sample, a candidate θ'_k was sampled from the prior and accepted with probability:

$$p\left(\theta_k^{\ell+1} = \theta'_k\right) = \min\left\{1, \frac{p\left(\mathbf{X} \mid \mathbf{Z}, \mathbf{s}, \mathbf{m}, \boldsymbol{\theta}^{\ell}_{-k}, \theta'_k\right)}{p\left(\mathbf{X} \mid \mathbf{Z}, \mathbf{s}, \mathbf{m}, \boldsymbol{\theta}^{\ell}\right)}\right\}, \tag{12}$$

which was achieved by sampling *u* from Uniform(0,1) and setting $\theta_k^{\ell+1} = \theta'$ if
$p\left(\theta_k^{\ell+1} = \theta'_k\right) > u$
, and $\theta_k^{\ell+1} = \theta_k^{\ell}$ otherwise.

For the simulations presented in Figure 5 of the main article, the MCMC sampler was run for 10,000 iterations so as to converge on the correct posterior distribution ("burn in"). Then, the algorithm was run for another 2,000 iterations, from which every 20th iteration was taken as a sample, for a total of 100 samples. This sampling interval was used because successive samples produced by the MCMC sampler are not independent from each other. Twenty-five independent chains were run for each condition, but we only kept samples from chains in which the expected value of r_t was larger than 0.7 for training stimuli followed by the outcome and lower than 0.3 for training stimuli followed by no outcome. This ensured that the model had learned the training discrimination, although the main predictions of each hypothesis did not change much when samples from all chains were used or when other cutoff values were used.

For the simulations presented in Figure 8, we followed a procedure proposed by Courville (2006) to obtain learning curves from the model: the MCMC algorithm was run repeatedly with an increasing proportion of all training data in each case (e.g, all data up to block 1, all data up to block 4, etc.). Each run involved 6,000 "burn in" iterations and an additional 2,000 iterations, from which every 20th iteration was taken as a sample, for a total of 100 samples. Each simulated data point in Figure 8 is the average of thirty independent simulations (chains).

The generalization gradients shown in Figure 1 and 10 were obtained by sampling 100,000 regions from the prior distribution, each containing the point $\{0,0\}$. Values of θ from 0° to 360° in steps of 5° were considered. A discretized 201 × 201 grid of equidistant points was used to evaluate the height of the gradient from the proportion of sampled consequential regions containing each point.

We did not perform a sensitivity analysis trying to determine whether the predictions reported here are robust across changes in the parameters of the prior. Unfortunately, each simulation (i.e., chain) with the current implementation of our model takes very long to run (from several hours to days, depending on the simulation), making a sensitivity analysis computationally infeasible.

However, we note that the parameters in the prior should have little influence in our predictions, compared to the influence of the data likelihood, for two reasons. First, most of these parameters were chosen so that they would have little influence on inferences, providing "flat" priors. This is the case for the parameters governing the extent, location and orientation of consequential regions, which are the only aspects of the model that differ across competing hypotheses. Second, the number of "training" data points is large in most of our simulations (the exception being the first blocks of Experiment 2), and when such large samples are involved the influence of the prior in Bayesian inference is overwhelmed by the likelihood of the data.

Furthermore, the parameters of the prior are the same across all the simulated models, meaning that differences in the predictions of different models are not due to differences in such parameters. Our approach was to use the same parameters in the prior as used in all our previous simulations with the model, presented in Soto et al., (2014). This strategy of attempting to explain all available experimental evidence using a single set of parameter values has a long tradition in learning theory, and it provides a strong test for the specific model being used.

Simulations with circular consequential regions

The results of simulations presented in the main article were obtained assuming rectangular consequential regions. Although this is a common assumption in work with the rational theory of generalization (e.g., Austerweil & Griffiths, 2010; Tenenbaum & Griffiths, 2001), Shepard (1987) originally proposed that for integral dimensions "psychological space should have no preferred axes. The consequential region is then most reasonably assumed to be circular" (p. 1322). Circular regions have the same size in any direction of psychological space, including the main axes, implementing the correlation hypothesis but also the idea that integral axes are not primary.

It is possible that implementing the correlation hypothesis through circular regions would correctly predict the results of our experiments. If this was the case, then it would be incorrect to claim that our study supports the direction hypothesis over the correlation hypothesis. To examine this possibility, we performed additional simulations of experiments 1 and 2 using circular regions for integral dimensions.

As in the original version of the correlation hypothesis, each consequential region k in this model is characterized by a vector of position parameters $\mathbf{m}_k$ and by a single size parameter s_k. As before, the position vector represents the center of the consequential region, but now the size parameter represents the diameter of the circular consequential region instead than its size in any particular direction of space. Stimuli are still assumed to be sampled uniformly from the consequential region, so the probability of sampling stimulus $\mathbf{x}_{\mathrm{ti}}$ from region c_k is equal to:

$$p\left(\mathbf{x}_{ti} \mid c_k\right) = \frac{1}{\pi\left(\frac{s_{kj}}{2}\right)^2}, \tag{13}$$

for stimuli that fall within the consequential region $\left(\left\|\mathbf{x}_{ti} - \mathbf{m}_k\right\| < \frac{s_k}{2}\right)$ and zero otherwise.

To perform simulations using circular regions, we used the same sampling algorithm used for squared regions. The only difference was that the stimulus likelihood was computed according to Equation (12) instead of Equation 4 from Soto et al. (2014). The simulation procedures were also the same as previously explained.

The results of simulations of Experiments 1 and 2 using a circular regions model are presented in the top and bottom panels of Figure A4, respectively. The top panel shows that the model correctly predicts higher responding to a2b3 than to a3b2 in Experiment 1. However, the predicted difference is very small compared to the predictions of the direction hypothesis (Figure 5a in the main article) and the difference actually found in the generalization data (Figure 10 in the main article). Given the error in our measurements, such a small predicted difference should have been very difficult to detect in our experiment. Still, the correlation hypothesis implemented with circular regions can at least qualitatively capture the results of Experiment 1.

The bottom panel of Figure A4 shows the predictions of the circular regions model for Experiment 2. In this case, the model fails to capture the correct pattern of results of a faster learning rate for the integral condition than for the separable condition.

Unlike the predictions of the direction hypothesis (Figure 8a in the main article) and the observed results (Figure 9 in the main article), the circular regions model does not predict a consistent better performance in the integral condition throughout training.

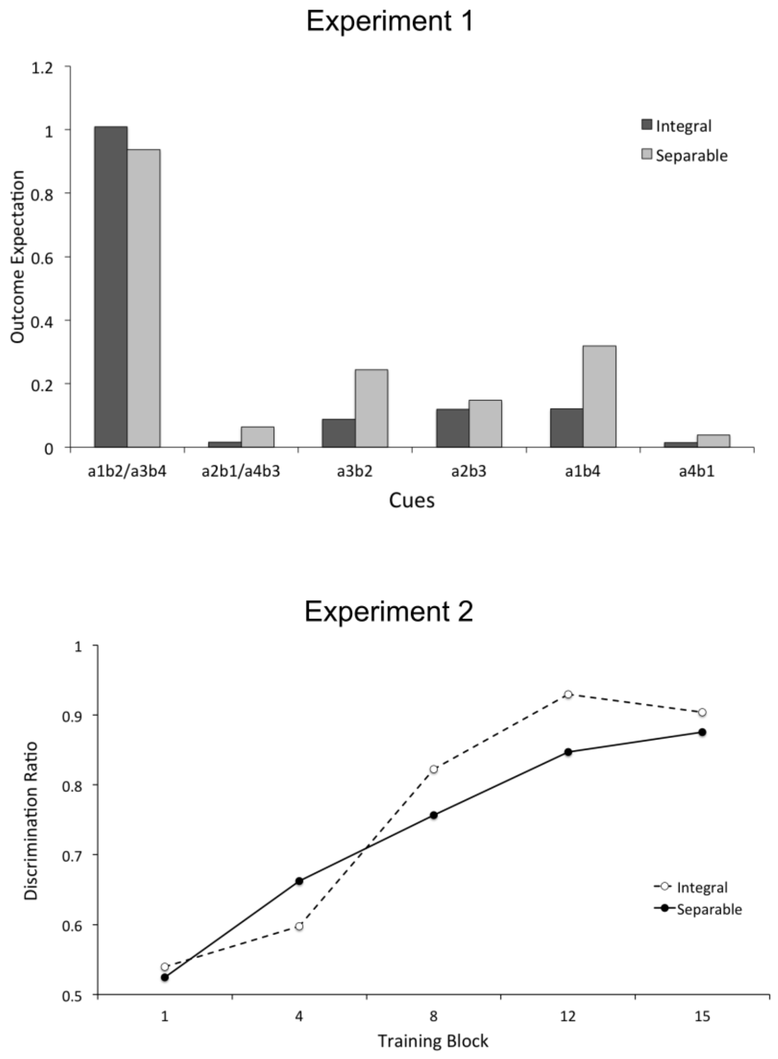


Figure A4. Simulations of Experiments 1 and 2 using circular consequential regions.

In summary, our simulations suggest that a model implementing the correlation hypothesis through circular regions could qualitatively capture the results of Experiment

1–although predicting a very small effect that would have been difficult to detect in our experiment–, but could not capture the results of Experiment 2. Thus, the results of our experiments support the direction hypothesis as the best explanation of integrality over the two versions of the correlation hypothesis proposed in the literature. Furthermore, the direction hypothesis offers a number of additional advantages over both versions of the correlation hypothesis, which are summarized in the general discussion section of the main article. Importantly, circular consequential regions cannot explain the results of compound generalization phenomena (Soto et al., 2014) and cannot explain why integral dimensions are psychologically privileged (see discussion and references in main article).

B. Possible confounds in Lachnit's (1988) experiment

As shown in Figure B1a, if stimuli along one of the dimensions used by Lachnit (1988; Dimension A in the figure) were more similar to each other than stimuli along the other dimension (Dimension B in the figure), then this would favor the pattern of results expected for separable dimensions, because each critical test stimuli would share a dimensional value with the closest training stimulus (see inside dotted ellipses in Figure B1b). As shown in Figure B1b, if the two central values of one dimension (Dimension A in the figure) were very close to each other, but far from the two extreme values in the dimension, then this would favor the pattern of results expected for integral dimensions, because each critical test stimulus would not share any dimensional value with the closest training stimulus (see inside dotted ellipses in Figure B1b). Together, the patterns of similarity shown in Figure 6 could explain the results obtained by Lachnit without resorting to the distinction between integral and separable dimensions.

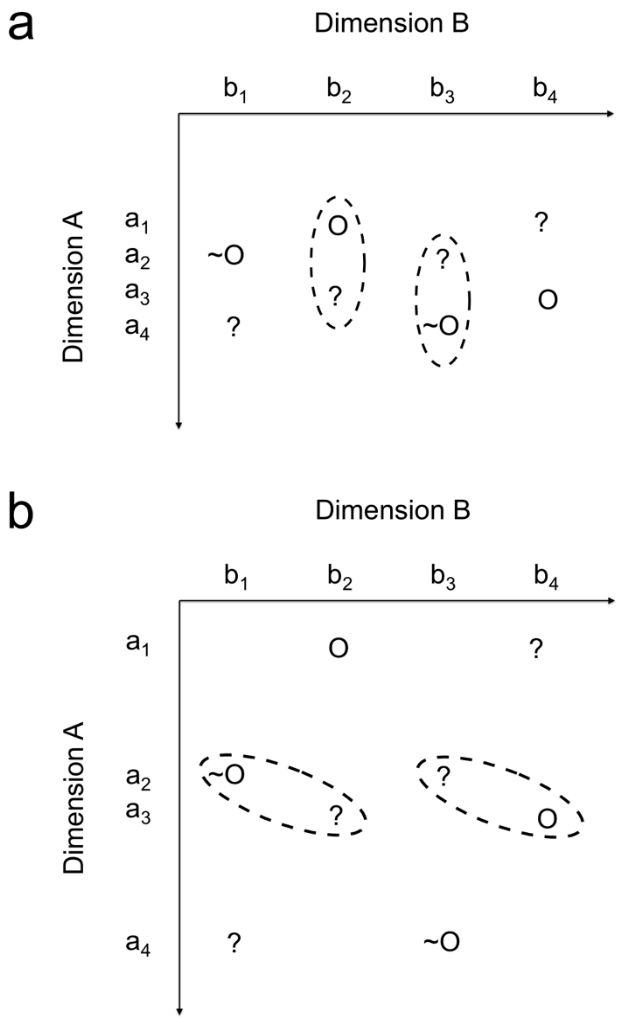


Figure B1. Schematic representation of how scaling along a single dimension could have affected the results of Lachnit's (1988) experiments. When one of the dimensions is compressed in relation to the other (a), each test stimulus and its closest training stimulus (grouped through an ellipse) share the same value in one dimension. When the two middle values in one of the dimensions are very similar (b), each test stimulus and its closest training stimulus do not share a value in either dimension.

C. Supplemental methods

Instructions to participants

The experimenter told the participants that all necessary instructions would be presented on the "instruction screens" included in the computer program. They then were left to complete the experiment in a private room. The following instructions were presented to the participants at the beginning of the training session (in Spanish):

Thank you for agreeing to participate in this research. In this experiment we are studying the learning mechanisms of humans. Your participation is anonymous and voluntary, and all your answers will be kept completely confidential.

We would like you to imagine that you are a dermatologist, that is, you are someone who studies the skin of people. Suppose that you were hired by a company that is developing new organic sunscreens, composed of certain kinds of microorganisms. You suspect that some of these microorganisms may provoke maleficent effects (aging), beneficent effects (rejuvenation), or neutral effects in the skin. In an attempt to discover which microorganism causes beneficent, maleficent or neutral reactions, you tested several sunscreens in laboratory animals (rats) and observed their reaction.

The results of each test will be shown to you on a series of screens. You will see a separate screen for each test. In each screen you will be shown a microscope view of each microorganism tested in a given animal. Next you will be asked to predict whether the animal will have an aging reaction, a rejuvenation reaction or a neutral reaction. Simply, click "aging" if you believe the animal will have an aging reaction, click "rejuvenation" if you believe the animal will have a rejuvenation reaction or click "neutral" if you believe the animal will have no reaction. After you make your *prediction, the computer will inform you on the reaction the animal actually had. You will have to guess at first, but with the aid of the feedback, your predictions should soon start to become more accurate. Please, pay attention to the different microorganisms because they are very similar to each other. Remember that your goal as an allergist is to learn which of these microorganisms is causing an allergic reaction.*

You might see this experiment as a game and try to score as many points (correct predictions) as you can. You will see the percentage of correct predictions you have made near the bottom of the screen during the tests.

Later in the experiment, you will be asked to rate to what extent each of the microorganisms cause aging, neutral or rejuvenation effects, based on the information you have seen so far.

In summary, your task is to learn which microorganisms produce allergic reactions in the animals.

The following instructions were presented to the participants at the beginning of the testing session:

Next, we would like you to rate the degree to which various types of microorganism will have aging, neutral or rejuvenation effects in the rats. To rate the effect of each microorganism, use a scale from -5 to +5 points, where -5 means maximal aging, 0 means neutral effect, and+5 means maximal rejuvenation. Please, click here to continue

Stimuli and Task

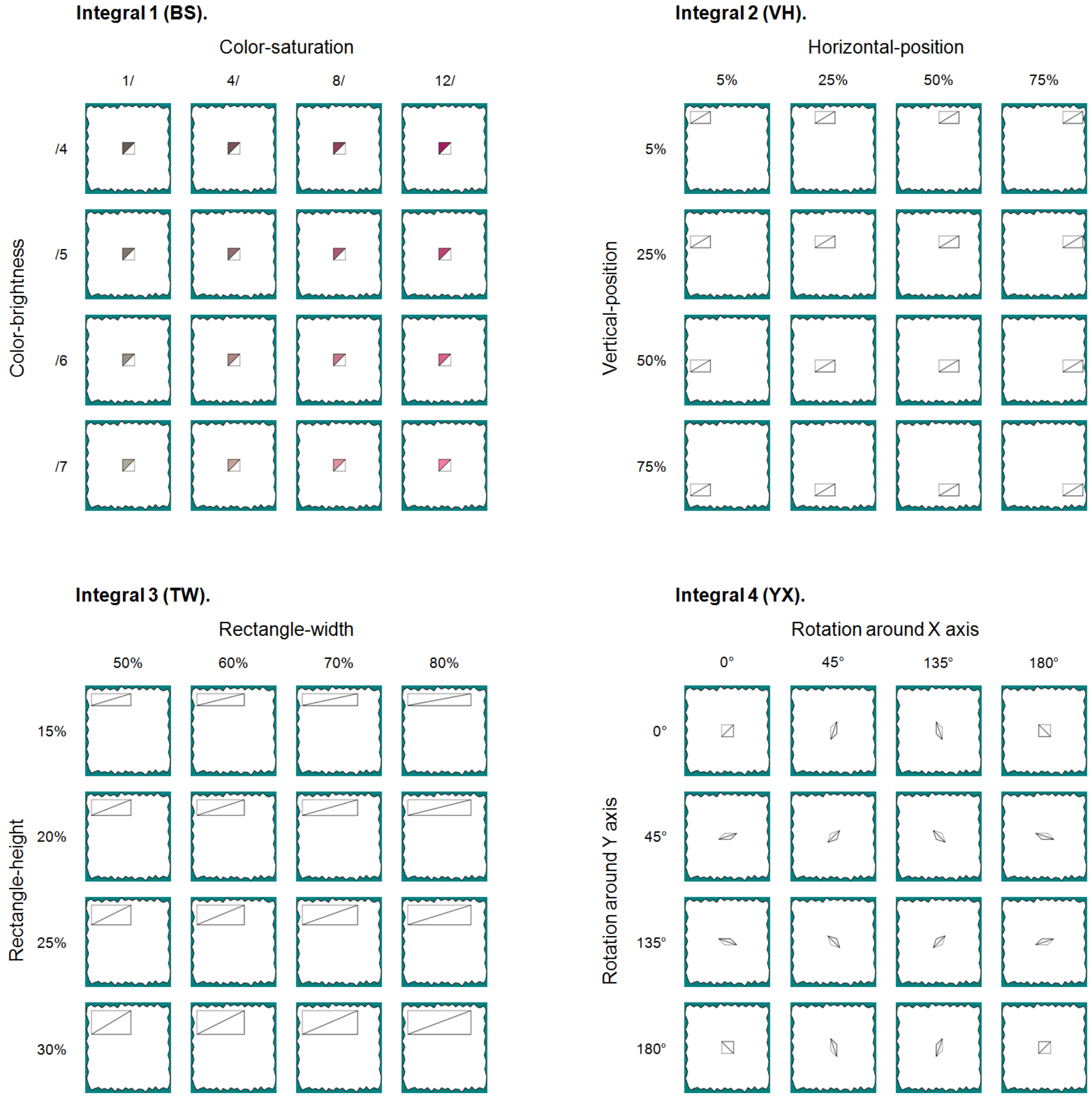


Figure C1. Stimuli used in the integral condition of our experiments.

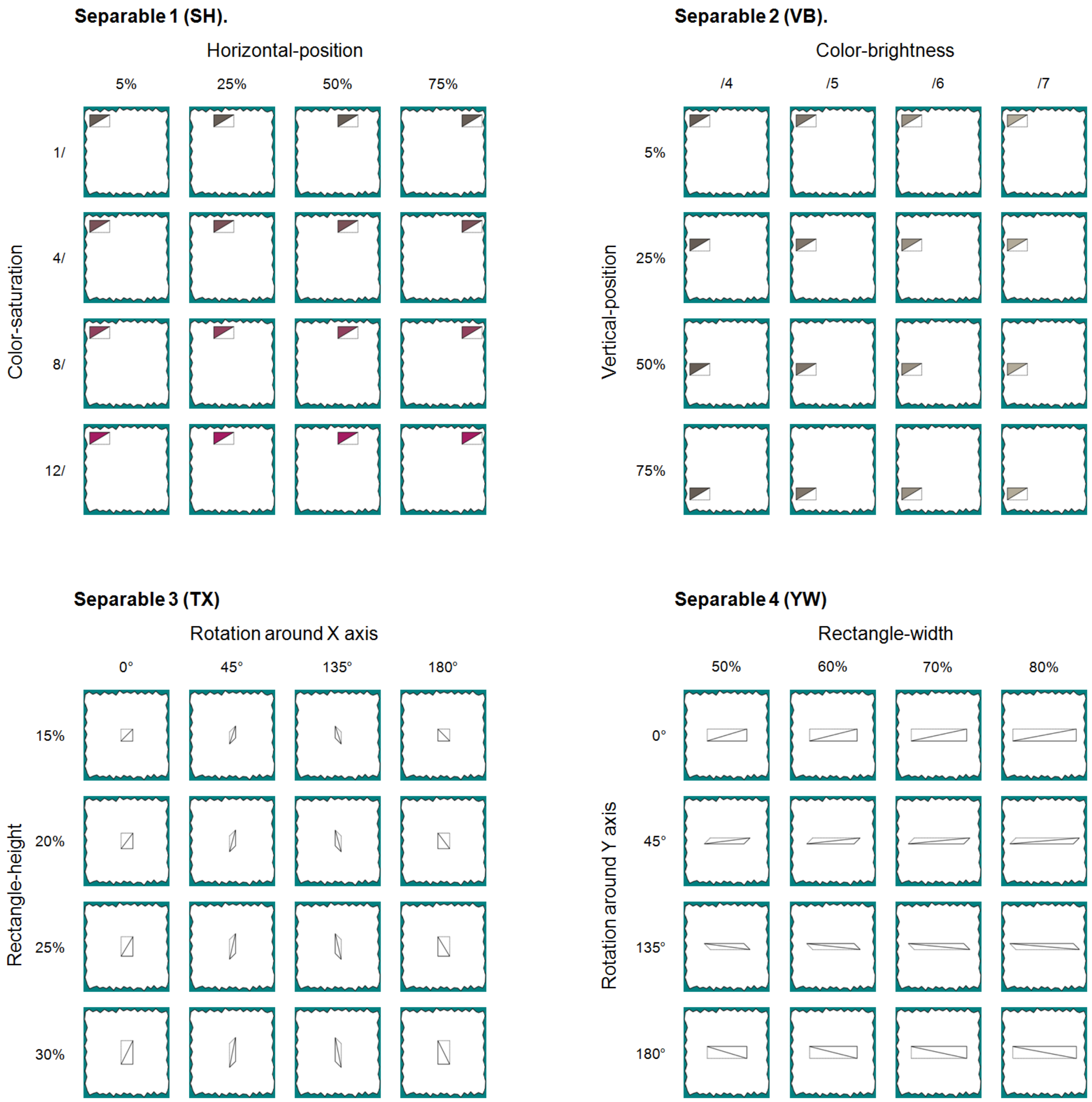
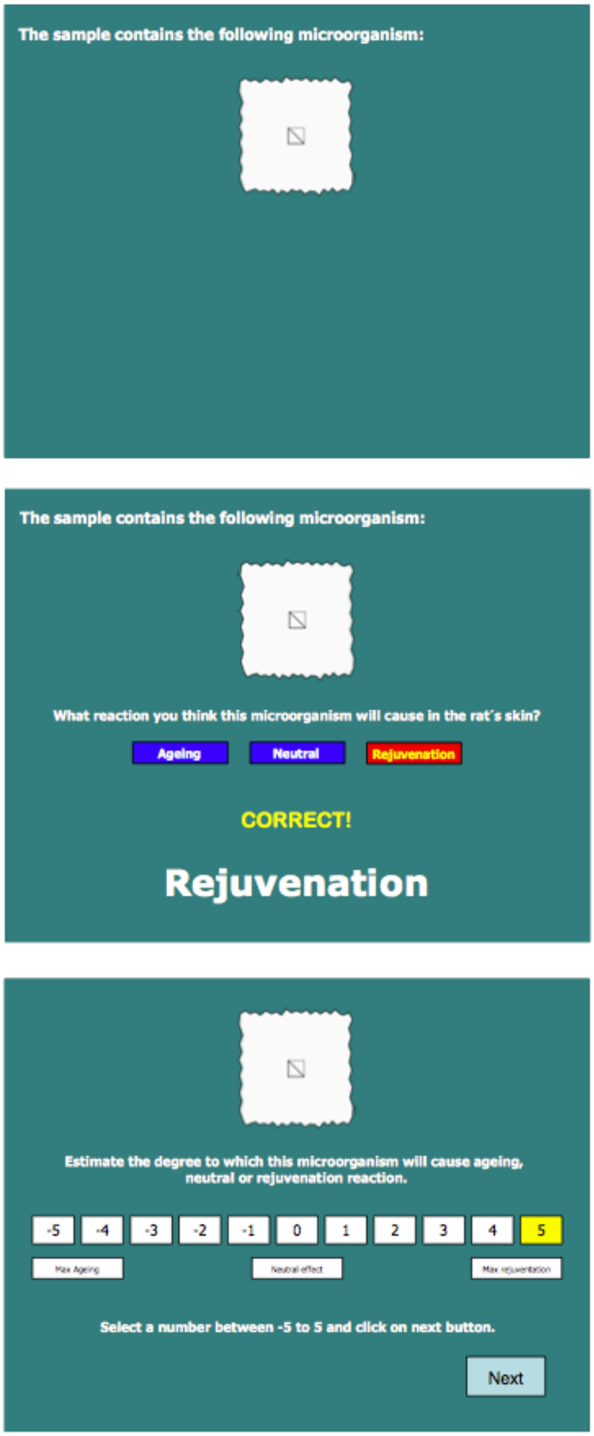


Figure C2. Stimuli used in the separable condition of our experiments.



The sample contains the following microorganism:

The sample contains the following microorganism:

What reaction you think this microorganism will cause in the rat´s skin?

CORRECT!

Rejuvenation

Estimate the degree to which this microorganism will cause ageing, neutral or rejuvenation reaction.

Select a number between -5 to 5 and click on next button.

Figure C3. Example of the screens presented to the participants during training (top and middle plots) and testing (bottom plot) in Experiment 1.

D. Experiment 3

As indicated in the main article, the set of stimuli used in this study (see Figures C1 and C2) offers the advantage that the same dimensions were used in both integral and separable conditions. This feature of our study allowed us to focus specifically on the effect of different types of dimensional interaction on multidimensional generalization, controlling for the possible differential effect of unidimensional generalization across conditions. Although this is an advantage of our study compared with most previous research that has compared integral and separable dimensions (e.g., Dunn, 1983; Goldstone, 1994; Lachnit, 1988; Soto & Wasserman, 2010), it also forced us to use completely new stimuli and combinations of dimensions. We selected several pairs of dimensions that have been reported to be integral (e.g., Dunn, 1983; Garner & Felfoldy, 1970; Lachnit, 1988; Monahan & Lockhead, 1977; Soto & Wasserman, 2010) and assumed that cross combinations of some dimensions belonging to those pairs would produce separable pairs. However, we have offered no independent evidence that the specific pairs of dimensions used in our studies did indeed differ in their level of separability and integrality. The goal of Experiment 3 was to provide an independent assessment of the integrality and separability of the stimuli used in our experimental setting.

The traditional way to determine whether a pair of dimensions is integral or separable is through a series of converging operations (Garner, 1974). Among those operations, one of the most commonly used is determining the metric of a multidimensional spatial model that describes generalization or similarity data best (e.g., Dunn, 1983; Hyman & Well, 1967; Soto & Wasserman, 2010). Unlike other operational definitions, this metric is directly related to generalization phenomena, which is the focus of the present study.

Thus, in the present experiment we focus on determining whether the metric of a spatial model used to describe human similarity judgments for our stimuli (Figures C1 and C2) differed between integral and separable sets of dimensions. Finding such a difference in metric would validate the stimulus manipulations used in Experiments 1 and 2.

In a spatial model, each stimulus is represented by its coordinates in a space with *K* dimensions. This means that stimulus *i* is represented by *K* values arranged in the vector $\mathbf{x}_i$. To compute the distance between stimuli *i* and *j*, d_{ij}, from their coordinates $\mathbf{x}_i$ and $\mathbf{x}_j$, one can use the generalized Minkowski formula (e.g., Shepard, 1987; Melara, Marks & Lesko, 1992), which states that:

$$d_{ij} = \left(\sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^{r} \right)^{1/r}, \qquad (1)$$

where *r* is the Minkowski exponent determining what metric is used in the spatial representation. The city-block metric linked to separable dimensions is obtained when *r* is equal to one. One can see from Equation (1) that in this case the distance between two stimuli is the sum of their distances along each dimension. The Euclidean metric linked to integral dimensions is obtained when *r* is equal to two. In this case, the distance between two stimuli is computed using the Pythagorean formula.

To examine the metric of a spatial model that best describes the perceived similarity of our stimuli, we asked different groups of participants to judge the dissimilarity among all possible pairs of stimuli from each set shown in Figures C1 and C2. Each participant provided dissimilarity scores for a single pair of dimensions (16 stimuli, 120 stimulus pairs), which were analyzed using Multidimensional Scaling (MDS) techniques. We expected that the metric that best describes dissimilarity data obtained from integral dimensions (brightness/saturation, vertical/horizontal position, rectangle width/height, and rotation around the X-axis/ rotation around the Y-axis) would be closer to the Euclidean metric than the metric that best describes dissimilarity data obtained from separable dimensions (saturation/horizontal position, vertical position/ brightness, rectangle height/ rotation around the Y-axis, and rectangle width/ rotation around the X-axis), which in turn would be closer to the city-block metric.

Method

Participants. A total of 64 undergraduate psychology students at the University of Talca, Chile participated in the experiment for course credit. They had a mean age of 18.63 years (*S* = 0.24). They were tested individually and had no previous experience in similar research.

Materials and procedure. The materials and strategy for stimulus construction of Experiment 3 were identical to those of Experiment 1. In the present experiment, all 16 stimuli from each of the four integral and 4 separable stimulus sub-sets were used (see Figures C1 and C2). Each participant was randomly assigned to one of these conditions (n=8) and asked to rate the dissimilarity of all possible pairs obtained from the 16 stimulus sub-set.

The instructions were as follows: *"In this task, you will be shown two images side-by-side in the computer screen. Please, look at these images and judge how different they are"*. At the beginning of each trial, two stimuli appeared in the top part of the computer screen, one to the left and one to the right of the middle line. Participants were asked to rate how different were the two images displayed on the screen by choosing a number from a scale displayed below the stimulus pair. The scale was composed of eleven points, with numeric labels going from 0 (minimal difference) to 10 (maximal difference). Choices were recorded through a mouse-click on the selected dissimilarity rating. To confirm their choice and move to the next stimulus pair, participants had to click on a button labeled “next.”

In order to familiarize participants with the task, the experimental session began with 20 practice trials involving randomly selected pairs. Next, the participants were required to rate the dissimilarity of all pairs of different stimuli twice, once with each left-right positioning of the stimuli in the screen, for a total of 240 trials.

Data analysis. Dissimilarity scores for the same stimulus pairs presented in different spatial order were averaged before analysis, resulting in a total of 120 dissimilarity scores per participant. To find the best-fitting value of the Minkowski exponent *r*, these dissimilarity data were analyzed trough a constrained MDS model (see Borg & Groenen, 2005; Heiser & Meulman, 1983). This type of analysis allowed both to include assumptions in the model that were common to our previous experiments and to make the results easier to interpret[1]. Specifically, we assumed correspondence (see Beals, Krantz, & Tversky, 1968; Dunn, 1983), meaning that the pairs of dimensions experimentally manipulated to build stimuli were represented as orthogonal axes in the spatial model, corresponding to psychological dimensions perceived by people. This is necessary to understand the interaction between the dimensions explicitly manipulated in our experiments 1 and 2. We also assumed intra-dimensional homogeneity (Dunn, 1983;

[1] Additional analyses carried out using unconstrained non-metric MDS led to many degenerate solutions, making the interpretation of the estimated Minkowski exponents difficult.

Heiser & Meulman, 1983), meaning that objects with the same level on an experimenter-defined dimension shared the same level in the corresponding psychological dimension, with the MDS solution forming a regular grid. This ensures that the final solution respects the geometry of our simulations of experiments 1 and 2, and avoids the problem that in two-dimensions the city-block metric becomes identical to the dominance metric, except for a rotation and scaling (Shepard, 1991). Finally, we assumed that the function relating distances in psychological space (d_{ij}) and perceived dissimilarity (δ_{ij}) was exponential (Shepard, 1987) and included parameters for translation (α) and scale (β) of dissimilarities (i.e., $\delta_{ij} = \alpha + \beta e^{-d_{ij}}$). We want to emphasize that these assumptions are not unique to our study; each of them has been previously used by researchers to investigate the spatial metric of stimulus dimensions (e.g., Hyman & Well, 1967; Ronacher & Bautz, 1985; Soto & Wasserman, 2010).

Previous research has shown that measures of badness of fit (e.g., stress) rarely have a single minimum at a particular value of the Minkowski exponent. For separable dimensions (Shepard, 1991) and for artificial data generated from a city-block metric (Caporossi, 2008), the cost function has an inverted-V shape, with maximal stress around $r = 2$, but descending both when $r < 2$ and $r > 2$. Likewise, for integral dimensions (Shepard, 1991) and for artificial data generated from a Euclidean metric (Caporossi, 2008), the cost function can show at least two minima (W shape), one at each side of the true value of $r = 2$, but none of them at the actual true value. This means that, depending on the starting value of r, the optimization algorithm could find optimal solutions for r that do not correspond to the actual underlying metric. To mitigate this problem, we followed the strategy proposed by others (e.g., Lee, 2001; Okada & Shigemasu, 2010) of limiting *r* to a maximal value of 2.5. Finally, to avoid local minima, the optimization was run 100 times, each time with a different set of starting coordinates and varying the initial value of the Minkowski metric (randomly chosen from 0 to 2.5).

Most previous research trying to determine the best-fitting metric in an MDS model has used group averages of the dissimilarity ratings as input to the analysis. However, this approach is known to be problematic (see Ashby, Maddox, & Lee, 1994), which led us to perform separate MDS analyses on the data provided by each participant. The optimal exponents obtained from these analyses were used as input to a 2 (dimension type: integral versus separable) x 4 (dimension pair) mixed-effects ANOVA with dimension pair nested inside dimension type. This ANOVA allowed us to determine whether there were significant differences in the Minkowski exponent estimated for integral and separable dimensions.

Results and discussion

Figure D1 presents the mean best fitting values of *r* obtained with the constrained MDS model for the integral (BS, VH, TW, and YX) and separable groups (SH, VB, TX, and YW). It is clear that all integral pairs exhibited greater mean values of *r* than the separable pairs. The ANOVA showed a reliable effect of dimension type, *F* (1, 56) = 9.680, $p < 0.01$, indicating greater overall *r* values in the integral than in separable condition, and no reliable interaction effect between dimension type and dimension pair, F (6, 56) < 1, indicating that the main effect of dimension type did not depend on specific pairs of dimensions within each condition.

The data depicted in Figure D1 indicate that the average estimated value for the Minkowski exponent *r* was around 1.0 for condition Separable, and around 1.5 for condition Integral. Thus, although the data from separable dimensions fit relatively well with a metric close to city-block, the data from integral dimensions approximated a metric intermediate between city-block and Euclidean. This finding is not too surprising, for three reasons: (i) the relation between type of dimension and metric is known to be only approximate and similar results have been reported earlier (e.g., Dunn, 1983; Shepard, 1991; Ronacher & Bautz, 1985), (ii) simulated data generated from an Euclidean metric is fit better by values of *r* lower than 2 than by a value of exactly 2 (Caporossi, 2008), and (iii) this is a common pattern of results that has been interpreted as evidence that integral dimensions are true psychological dimensions that are usually, but not always, processed holistically (Kemler-Nelson, 1993; see general discussion).

In conclusion, the MDS analyses suggest an overall difference in the metric used by participants to judge dissimilarity among stimuli belonging to the integral and separable conditions of our experimental setting. Since this difference is in the expected direction of greater Minkowski exponents for the integral than for the separable conditions, there is reason to believe that the experimental conditions of experiments 1 and 2 adequately manipulated the integrality and separability of the dimensions involved, as it was intended.

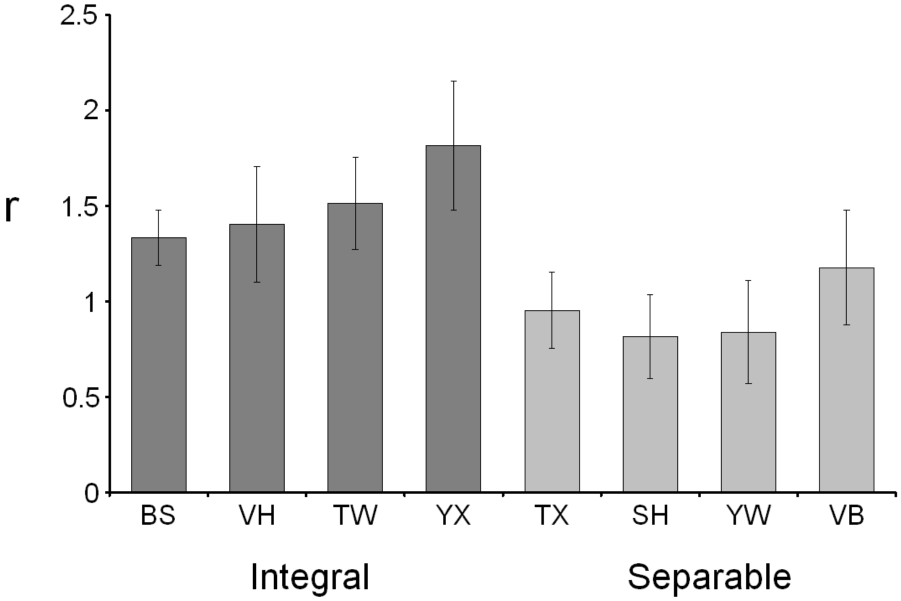


Figure D1. Mean estimated Minkoski exponent (r) for each of the dimension pairs included in Experiment 3. An exponent of 2.0 represents an Euclidean metric and an exponent of 1.0 represents a city-block metric. Error bars are standard errors of the mean.

References

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. Psychological Science, 5(3), 144-151.

Austerweil, J. L., & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 73-78).

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. Psychological Review, 75(2), 127-142. doi: 10.1037.h0025470

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer: New York.

Caporossi, G. (2008). Identification of the Minkowski parameter for multidimensional scaling. Les Cahiers du GERAD G–2008–61.

Courville, A. C. (2006). A latent cause theory of classical conditioning (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Dunn, J. C. (1983). Spatial metrics of integral and separable dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(2), 242-257.

Garner, W. R. (1974). *The processing of information and structure*. New York: Lawrence Erlbaum Associates.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*(3), 225–241. doi:10.1016/0010-0285(70)90016-2

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Griffiths, T. L., & Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185-1224.

Heiser, W. J., & Meulman, J. (1983). Constrained multidimensional scaling, including confirmation. Applied Psychological Measurement, 7(4), 381-404. doi:10.1177/014662168300700402

Hyman, R., & Well, A. (1967). Judgments of similarity and spatial models. *Perception and Psychophysics, 2*(6)*,* 233–248.

Kemler-Nelson, D. G. (1993). Processing integral dimensions: The whole view. Journal of Experimental Psychology: Human Perception and Performance, 19(5), 1105-1113.

Lachnit, H. (1988). Convergent validation of information processing constructs with Pavlovian methodology. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 143–152. doi:10.1037/0096-1523.14.1.143

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. Journal of Mathematical Psychology, 45, 149-166. doi: 10.1006/jmps.1999.1300

Melara, R. D., Marks, L. E., & Lesko. K. E. (1992). Optional processes in similarity judgements. *Perception & Psychophysics*, *51*(2), 123-133.

Monahan, J. S., & Lockhead, G. R. (1977). Identification of integral stimuli. Journal of Experimental Psychology: General, 106(1), 94-110. doi:10.1037/0096-3445.106.1.94

Okada, K., & Shigemasu, K. (2010). Bayesian multidimensional scaling for the estimation of a Minkowski exponent. Behavior Research Methods, 42(4), 899-905. doi:10.3758/BRM.42.4.899

Ronacher, B., & Bautz, W. (1985). Human pattern recognition: Individually different strategies in analyzing complex stimuli. Biological Cybernetics, 51(4), 249-261.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323. doi:10.1126/science.3629243

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In J. Pomerantz & G. Lockhead (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53–71). Washington, DC: American Psychological Association.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review, 121*(3), 526-558. doi:10.1037/a0037018

Soto, F. A., & Wasserman, E. A. (2010). Integrality/separability of stimulus dimensions and multidimensional generalization in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(2), 194–205. doi:10.1037/a0016560

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. doi:10.1017/S0140525X01000061