CONCEPT FORMATION IN DIMENSIONAL SPACES

Peter Gärdenfors and Kenneth Holmqvist

Lund University Cognitive Science
Kungshuset, Lundagård
S–223 50 Lund, Sweden
E-mail: Peter.Gardenfors@fil.lu.se, Kenneth.Holmqvist@fil.lu.se

1. INTRODUCTION

This article presents a series of experiments on concept formation. These experiments compared four different categorisation rules. The rules presume that the stimuli to be classified can be modelled psychologically as points in a dimensional space (Shepard 1987, Gärdenfors 1990a, 1990b). Two of the rules are based on the assumption that there are prototypical representatives of a concept, a third rule is a "nearest neighbour" model, and the fourth is based on average distances (Reed 1972), where distances are measured in the dimensional space. Before giving a more precise description of the categorisation rules and the experiments, we will present the general theoretical approach that is adopted.

1.1 Theoretical background

In this article, we take *categorisation* to be a rule for classifying objects. The result of categorisation will be a number of *concepts*. The concepts will generate classifications of stimuli. In the model presented here, where stimuli are represented as points in dimensional spaces, a categorisation will generate a partitioning of the space and a concept will correspond to a region of the space.

One can find many proposals for categorisation rules in the philosophical and psychological literature. A classical example is Aristotle's theory of *necessary and sufficient conditions* (see Smith and Medin (1981) for a presentation of this and other theories of concept formation). His view on how concepts are determined has had an enormous influence throughout the history of philosophy.

As is shown by numerous experiments (e.g. Rosch 1975, 1978, Labov 1973, Mervis and Rosch 1981, Smith and Medin 1981), the Aristotelian theory is not realistic as a cognitive or psychological account of how people form and use concepts. As a result of a growing dissatisfaction with the classical theory of concepts, several alternative theories have been developed within cognitive psychology. The most well known is *prototype theory* (e.g. Rosch 1975, 1978, Mervis and Rosch 1981, Lakoff 1987). The main idea of this theory is that within the class of objects falling under a concept certain members are judged to be more representative of the concept than others. The most representative exemplars of a concept are called *prototypical* members. A given set of prototypes for a class of concepts can be used for generating a categorisation by the rule that a stimulus is classified according to which prototype it is most similar to.

However, even if prototype theory fares much better than the Aristotelian theory in explaining how people use concepts, the theory does not explain *how* such prototype effects can arise as a result of learning to use our concepts. The theory can neither account for how new concepts can be created from relevant exemplars, nor explain how the extensions of concepts are changed as new concepts in the same category are learned.

The purpose of this article is to present four models based on a "dimensional" or "geometric" approach to categorisation, together with the experiments that we have performed in order to test the models. The key notion is that of a *conceptual space* consisting of a number of dimensions which are used as the framework for various categorisation rules. A conceptual space can be seen as a geometric structure for which several categorisation rules can be formulated and tested. We want to show that a model based on conceptual spaces together with a classification rule based on so called Voronoi

Lund University Cognitive Studies – LUCS 26 1994. ISSN 1101–8453.

tessellations can provide us with an explanation of how concepts are formed and develop, at least for certain classes of objects. The space used in our experiments is a representation of *shell shapes.*

There are several related studies in the psychological literature. Reed's (1972) investigation of classifications of Brunswick faces uses an approach that is similar to ours, even if the "space" of faces has a very limited structure in comparison to the shell space presented below. Similarly, Ashby and Gott's (1988) investigations of decision rules for categorising multidimensional stimuli are based on a methodology that shares many features with ours. Some further examples of studies of concept formation that utilise dimensional notions are Pittenger and Shaw (1975) on faces, Labov (1973) on cups, Nosofsky (1986) on semicircles, and Nosofsky (1988) on colours.

1.2 Conceptual spaces and the geometric categorisation models

In this subsection, we introduce the notion of a *conceptual space* which serves as a theoretical framework for the different categorisation models. A conceptual space consists of a number of *quality dimensions.* The dimensions that will be considered in this article are assumed to be generated by our perceptual mechanisms, but in a general theoretical investigation one may consider quality dimensions that are of a more abstract non-sensory character. As examples of quality dimensions one can mention colour, pitch, temperature, weight, and the three ordinary spatial dimensions.

The notion of a *dimension* should be understood literally. It is assumed that each of the quality dimensions is endowed with a certain topological or metric structure. This structure can be determined by psycho-physical investigations. For example, perception of *weight* is one-dimensional with a zero point, isomorphic to the half-line of non-negative numbers; and the *hue* of a colour can be represented by a circular dimension (see e.g. Gärdenfors 1990a, 1992).

We cannot provide a complete list of the quality dimensions generated by our perceptual mechanisms. Some of the dimensions seem to be *innate* and to some extent hardwired in our nervous system, as for example colour, pitch, and probably also ordinary space. Other dimensions are presumably *learned.* Learning new concepts often involves expanding one's conceptual space with new quality dimensions. Quine (1969:123) notes that something like a conceptual space is needed to make learning possible:

> Without some such prior spacing of qualities, we could never acquire a habit; all stimuli would be equally alike and equally different. These spacings of qualities, on the part of men and other animals, can be explored and mapped in the laboratory by experiments in conditioning and extinction. Needed as they are for all learning, these distinctive spacings cannot themselves all be learned; some must be innate.

Let us now turn to an outline of how conceptual spaces may be used as a basis for a theory of categorisation. A first rough idea is to describe a concept as determined by a *region* of a conceptual space S, where "region" should be understood as a spatial notion determined by the topology and metric of S. For example, the point in the time dimension representing "now" divides this dimension, and thus the space of vectors, into two regions corresponding to the concepts "past" and "future".

Shepard (1987, p. 1319) gives an evolutionary argument that supports this proposal:

> An object that is significant for an individual's survival and reproduction is never sui generis; it is always a member of a particular class – what philosophers term a "natural kind." Such a class corresponds to some region in the individual's psychological space, which I call a consequential region. I suggest that the psycho-physical function that maps physical parameter space into a species' psychological space has been shaped over evolutionary history so that consequential regions for that species, although variously shaped, are not consistently elongated or flattened in particular directions.

One way of giving Shepard's idea a mathematical formulation is the following criterion where the topological characteristics of the quality dimensions are utilised to introduce a spatial structure on categories (cf. Gärdenfors 1990a, 1990b):

> *Criterion P*: A *natural category* is a convex region of a conceptual space.

A *convex* region is characterised by the criterion that for every pair of points v_1 and v_2 in the region all points *between* v_1 and v_2 are also in the region. The motivation for the criterion is that if some objects are located at v_1 and v_2 in relation to some quality dimension (or several dimensions) and both are examples of the category *P*, then any object located between v_1 and v_2 on the quality dimension(s) will also be an example of *P.* Criterion P presumes that the notion of "betweenness" is meaningful for the relevant quality dimensions. This is, however, a rather weak assumption that demands very little of the underlying topological structure. However, in what follows we shall work with the stronger assumption that the dimensions we consider have a *metric* so that we can also talk about *distances* between points in the space.

In support of Criterion *P* it can be shown that if prototype theory is combined with the idea of a metric conceptual space as a framework for categorisation, then the representation of categories as convex regions is to be expected. To see this, assume that some quality dimensions of a conceptual space are given, for example the dimensions of colour space, and that we want to partition it into a number of categories, for example colour categories. If we start from a set of prototypes p_1, ..., p_n of the categories, for example the focal colours, then these should be the central points in the categories they represent. One way of using this information is to assume that for every point p in the space one can measure the *distance* from p to each of the p_i's. If we now stipulate that p belongs to the same category as the *closest* prototype p_i, it can be shown that this rule will generate a partitioning of the space that *consists of convex areas* (convexity is here defined in terms of the assumed distance measure). This is the so called *Voronoi tessellation*, a two-dimensional example of which is illustrated in Figure 1.

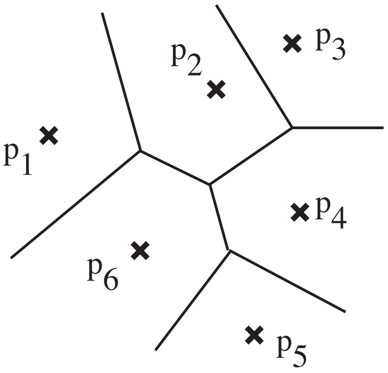


Figure 1. An example of a Voronoi tessellation of the plane into convex sets determined by a set of prototypical points.

Thus, assuming that a metric is defined on the subspace that is subject to categorisation, a set of prototypes will, by this method, generate a unique partitioning of the subspace into convex regions. Hence there is an intimate link between prototype theory and the analysis of this article where categories are described as convex regions in a conceptual space.

In the experimental investigations, we assume that the prototype for a category is determined from the set of exemplars of the category that a subject has seen. The rule we employ for calculating the prototype from a class of exemplars is that the *i*th coordinate $\mathrm{p_i}$ for the vector representing the prototype is the *mean* of the *i*th co-ordinate for all the exemplars. Applying this rule means that a prototype is not assumed to be given *a priori* in any way, but is completely determined by the experience of the subject. Figure 2 shows how a set of 9 exemplars, represented as differently filled circles, grouped into three categories generate three prototypical points, represented as black crosses, in the space. These prototypes then generate a Voronoi tessellation of the space.

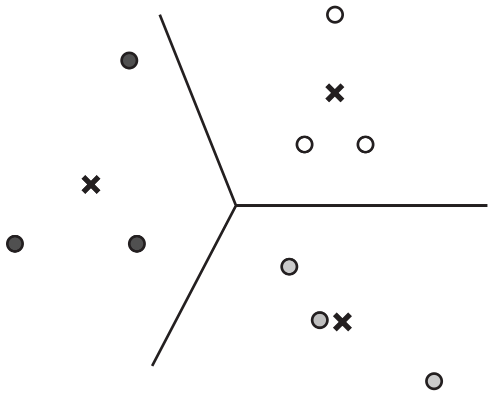


Figure 2. Voronoi tessellation generated by three classes of exemplars.

Furthermore, when the subject sees a new item in a category, the prototype for that category will, in general, change somewhat, since the mean of the class of examples will normally change. Figure 3 shows how this categorisation is changed by the addition of one new exemplar, marked by an arrow, to one of the categories. This addition shifts the prototype of that category, which is defined as the mean of the exemplars, and consequently the Voronoi tessellation is changed. The old tessellation is marked by hatched lines.

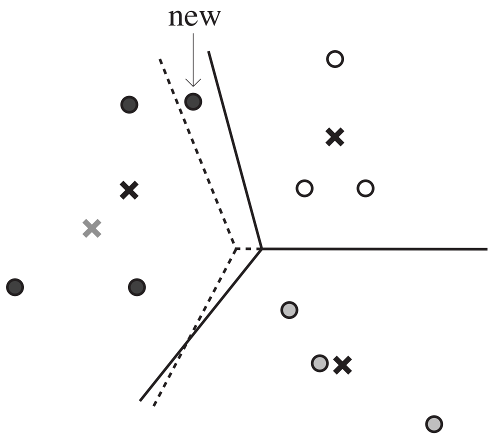


Figure 3. Change of Voronoi tessellation in Figure 2 after adding a new exemplar.

The Voronoi tessellation generated from a set p_1, ..., p_n of prototypes yields the following decision rule for categorisation:[1]

Prototype Voronoi categorisation (PV): An object represented as a vector x_i in a conceptual space belongs to the category for which the corresponding prototype is the closest, i.e., the p_j that minimises the distance between x_i and p_j.

A drawback of the standard Voronoi tessellation is that

[1] This rule is the same as the "Prototype model" in Reed (1972).

it is *only* the prototype that determines the partitioning of the conceptual space. However, it is quite clear that for many natural categorising systems some concepts correspond to "larger" regions than others. For example, the concept "duck" covers a much larger variety of birds than "ostrich," even though both concepts are *basic level* concepts in the terminology of Rosch (1978), i.e., they are both members of the same categorisation of birds.

So the question arises whether there is some way of generalising the Voronoi tessellation that can account for varying sizes of concepts in a categorisation, but which will still result in a convex partitioning of the underlying conceptual space. It can be shown (see Holmqvist and Gärdenfors, in preparation) that such a generalisation is possible. The standard Voronoi tessellation is based on the ordinary Euclidean metric, so that in order to determine the lines that form the tessellation one solves the equation

$$\Sigma_i (v_i - x_i)^2 = \Sigma_i (w_i - x_i)^2$$

where $v = (v_1,...,v_n)$ and $w = (w_1,...,w_n)$ are the vectors of two prototypical points in the conceptual space. However, instead of saying that there is only a prototypical *point* for a particular concept one can introduce the notion of a prototypical *area* and then determine a *generalised* Voronoi tessellation by computing distances from such areas instead. In relation to the earlier example, the prototypical area for ducks could then be taken to be larger than the corresponding area for ostriches. We model this idea by assuming that the prototypical area for a concept can be described by a circle with centre $(v_1,...,v_n)$ and radius c_v. By varying the radius, one can change the size of the prototypical area (for example, the c_v for ducks would be larger than that for ostriches).

In order to determine the generalised Voronoi tessellation one then solves equations of the form

$$\Sigma_i (v_i - x_i)^2 - c_v = \Sigma_i (w_i - x_i)^2 - c_w$$

It can be proven (Holmqvist and Gärdenfors, in preparation) that for all choices of prototypical circles, this equation generates a set of straight lines that will partition the space into convex subsets. A simple illustration of a generalised Voronoi tessellation is given in Figure 4. The prototype Voronoi tessellation generated from the centres of the circles, corresponding to the prototypes, is indicated by hatched lines.

The metric generated by this kind of equation will not be Euclidean. All points on the prototypical circle will have distance zero from the prototype, and it turns out that points within the circle will have imaginary numbers as distances.

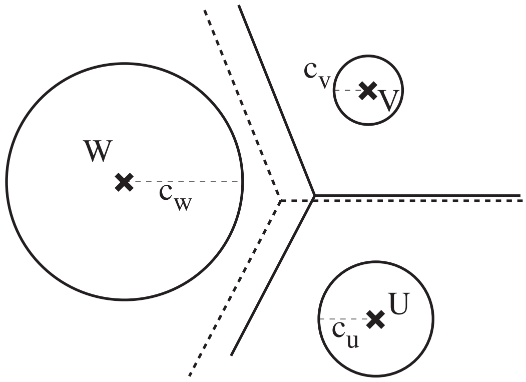


Figure 4. An example of a generalised Voronoi tessellation determined by the circles. The generalised Voronoi tessellation is represented by continuous lines and the prototype Voronoi tessellation by dashed lines.

Before the generalised Voronoi tessellation can be applied, the values c_v must be determined. A natural choice, that we will use throughout this article, is to define c_v as the magnitude of the *standard deviation* of the exemplars from the prototype. This choice entails that the generalised Voronoi tessellation can be completely determined from the co-ordinates of the exemplars of the different categories.

The generalised Voronoi tessellation corresponds to the following rule for categorisation:

Generalised Voronoi categorisation (GV): An object represented as a vector x_i in a conceptual space belongs to the category for which the corresponding prototypical circle is the closest.

Following Reed (1972), we will compare the results of the two Voronoi rules PV and GV to two other categorisation rules:

Nearest neighbour categorisation (NN): An object represented as a vector x_i in a conceptual space belongs to the category to which the exemplar that is closest to x_i is included.[2]

Average distance categorisation (AD): An object represented as a vector x_i in a conceptual space belongs to the category to which x_i has the smallest average distance to the examples for the category.

The four rules that have been introduced here often result in very similar categorisations. As a consequence, it will become difficult to distinguish between them in empirical tests of which rule best describes the behavior of subjects in classification tasks. Figure 5 illustrates an extreme case where the four rules generate clearly separate partitionings.

[2] Reed (1972), pp. 385-86, calls this rule the "proximity algorithm".

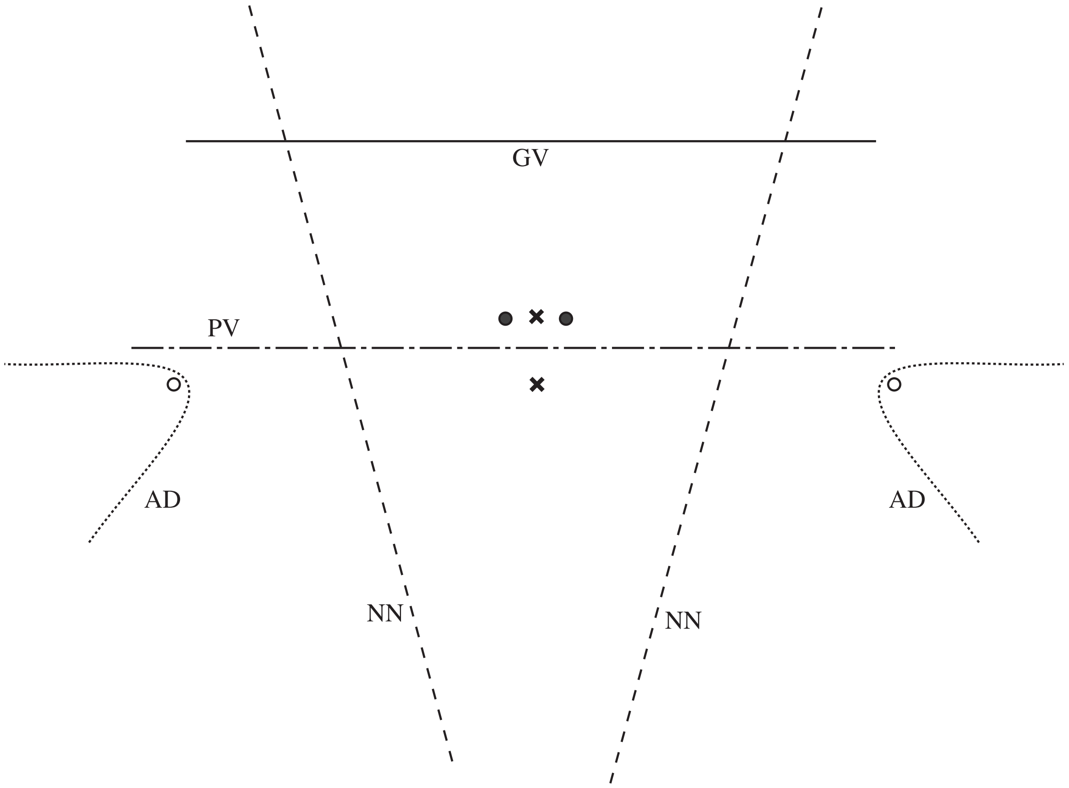


Figure 5. An example of how the four categorisation rules result in extremely different categorisations of the space. The two filled dots are exemplars of one category and the cross between them its prototype. The unfilled dots are exemplars of another category, again with its prototype marked as a cross between them. PV, GV, NN and AD denote partitionings generated by the four categorisation rules.

1.3 The shell space

In order to evaluate the four models presented above we have performed a series of experiments concerning the categorisation of *shell* shapes. There are several reasons why shells constitute a useful domain for empirical investigations:

(1) It is possible to generate a great number of fairly realistic shell shapes in a conceptual space that is built up from three dimensions. (This space will be described below.)

(2) The shells can easily be drawn by a graphic program where the only inputs are three co-ordinates in the shell space.

(3) The pictures generated by our program are identified by the subjects as pictures of realistic 3D shells. They are thus much more natural than most of the stimuli used in classification tasks in current cognitive psychology, like e.g. the dot-patterns (Posner and Keele 1968, Shin and Nosofsky 1992) or the semi-circles with an additional radius (Nosofsky 1986, Ashby and Lee 1991). Not even the schematic faces used by Reed (1972) have a very high degree of "ecological validity" (Gibson 1979).

(4) Even though test subjects recognise the object on the pictures as shells, they normally have no "prejudices" concerning how shells are actually classified in biology. This means that we can "create" new categories for the subjects by showing appropriate shells, i.e., more or less prototypical examples, in a desired region of the shell space.

2. PILOT STUDIES

In order to obtain an estimate of the metric space underlying perceptions of shell forms, we performed three pilot studies. These studies can be seen as calibration experiments determining the scaling solution of the underlying dimensions. The results of the experiments strongly confirm the hypothesis that the psychophysical shell space is indeed three-dimensional with an identifiable metric. The perceptually grounded metric that was estimated during the pilot studies is the one that was used in the two main experiments where the four categorisation rules presented above are evaluated. It is also the metric used below in Figures 8 and 11 where we illustrate the various classification tasks.[3]

[3] The metric is not identical to the one used in the graphics program but a simple transformation of it.

2.1 Stimuli

Throughout all experiments, we used depictions of shell shapes as stimuli. A shell normally grows in a spiralling way. The shape of the shell is, according to Raup (1966), largely determined by three factors (see Figure 6):

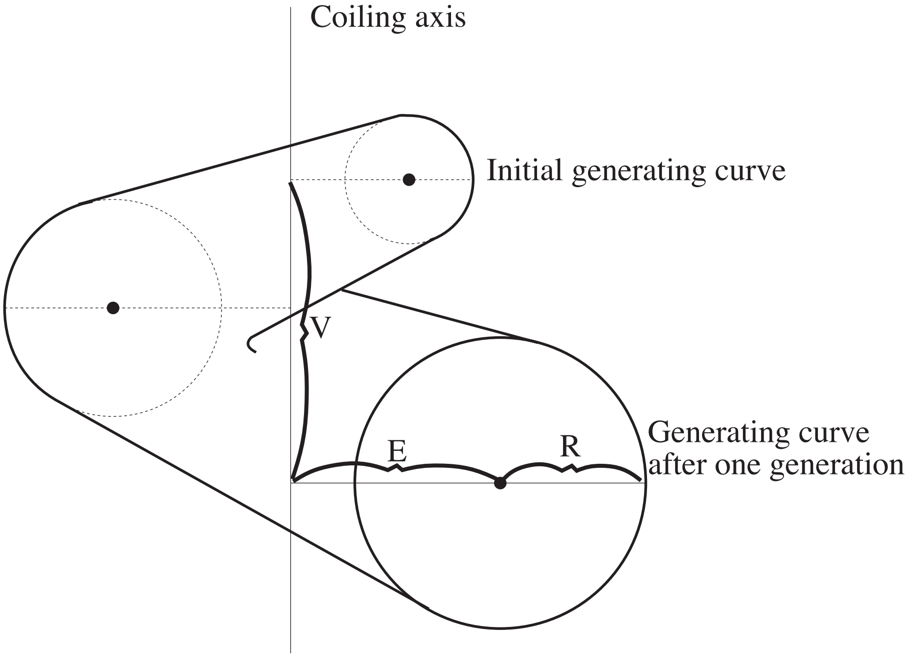


Figure 6. The three dimensions V, E, and R, generating a shell shape.

(1) The rate E of whorl expansion which determines the curvature of the shell. The rate E is assumed to be a constant defined as the quotient e_{n+1}/e_n between the distance e_{n+1} of the central point to the generating axis and the distance en one revolution earlier. Small curvature results in densely spiralled shells, while a high curvature produces openly spiralled shapes.

(2) The rate V of vertical translation along the coiling axis. The rate V is also assumed to be a constant. It is defined as the quotient v_{n+1}/v_n between the vertical distance v_{n+1} of the central point to the initial level on the generating axis and the corresponding distance v_n one revolution earlier. Having no vertical translation yields flat shells, while rapid growth results in elongated shapes.

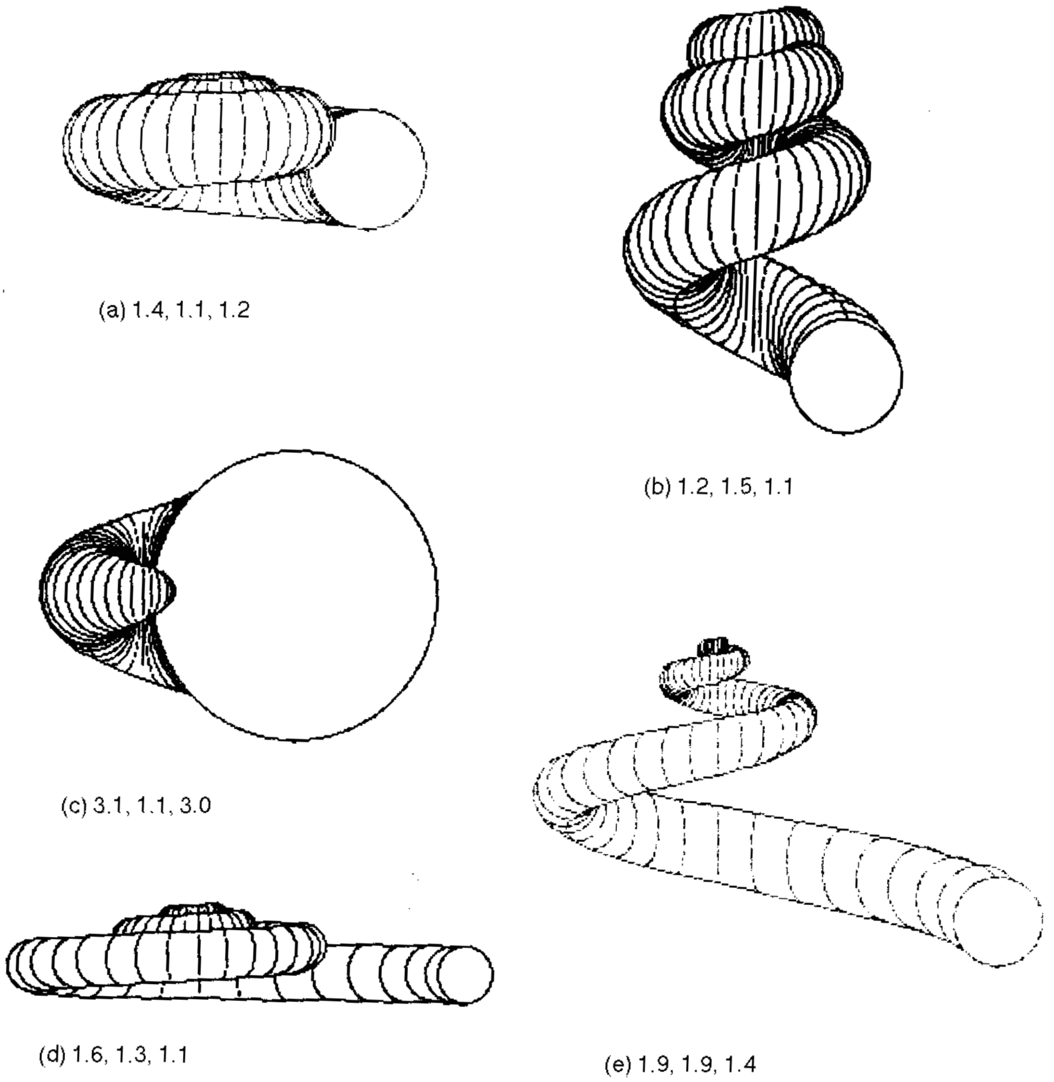
(3) The expansion rate R of the generating curve (the aperture) of the shell. Our graphic program only operates with a circular generating curve, but one finds in nature a large variation of the outline of the apertures (which is largely determined by the shapes of the soft body of the molluscs). The growth rate R, which is assumed to be constant, is defined as the quotient r_{n+1}/r_n between the radius r_{n+1} and the radius r_n one revolution earlier. Slow growth results in tube formed shells. Very rapid growth produces shells that look like in Figure 7c below.

Figure 7 shows some examples of shapes that are produced by different combinations of values for the co-ordinates. All shell pictures here and in the following are generated by our graphics program. The only input to the program are the three co-ordinates.

The three dimensions V, E, and R, span a space of possible shell shapes that is suitable for testing the four models of categorisation described above. However, the dimensional space is defined with the aid of three mathematical dimensions. As a preliminary step, it is necessary to test the psychological validity of the hypothesis that our perceptions of shells also form a three-dimensional space. But even if the perceptual shell space is three dimensional, it does not at all follow that the metric of the space is the same as the mathematical co-ordinates used by the graphic program.[4] Before we can apply the four categorisation

[4] It seems to us that this point is sometimes missed in the psychological literature. For example Ashby and Gott (1988) work with stimuli composed of two lines, one horizontal and one vertical, joined at the upper left corner (cf. their figure 2 on p. 35). The stimuli are described by a two dimensional vector (x,y) where x denotes the length of the horizontal component and y the length of the vertical component. Throughout the article, Ashby and Gott discuss the space generated by these axes and define their decision rules using this metric. However, from a number of observations in their article which are problematic for them, it seems that these co-ordinates do not produce an appropriate metric for the perceptual space producing the categorisations. It seems to us that if one makes a co-ordinate shift by defining $x' = x/y$ and $y' = x \cdot y$, one obtains a metric that is much better suited to explain the observed phenomena. The dimension x' measures the proportions of the lengths of the two line segments, while y' is a measure of the relative size of the stimulus.

models it is necessary to establish the relevant psychological metric, i.e., the scaling solution, of the shell space.



(a) 1.4, 1.1, 1.2

(b) 1.2, 1.5, 1.1

(c) 3.1, 1.1, 3.0

(d) 1.6, 1.3, 1.1

(e) 1.9, 1.9, 1.4

Figure 7. Some examples of shell shapes drawn by the graphic program together with their generating co-ordinates (rate of whorl expansion, vertical growth, growth of radius of aperture).

This methodology is completely in line with Shepard's (1987, p. 1318) recommendations for how psychological laws should be obtained:

> Analogously in psychology, a law that is invariant across perceptual dimensions, modalities, individuals, and species may be attained only by formulating the law with respect to the appropriate abstract psychological space. The previously troublesome variations in the gradient of generalisation might then be attributable to variations in the psycho-physical function that, for each individual, maps physical parameter space (the space whose co-ordinates include the physical intensity, frequency, and orientation of each stimulus) into that individual's psychological space. If so, a purely psychological function relating generalisations to distance in such a psychological space might attain invariance.

2.2 Calibration of Stimuli

The three dimensions given by Raup's (1966) model are not the most natural ones from a perceptual point of view. When estimating the vertical and horizontal growth rates of shells, we don't look at the centre of the generating circle. Instead we focus on its extreme points in the vertical and horizontal direction, loosely speaking the height and width of the shell. This means that the vertical and horizontal expansion rates should instead be described by the following values:

$$V' = V + R - 1$$

$$E' = E + R - 1$$

This shift of co-ordinates will not change the distances between the points in the space. Using the space generated by the dimensions V', E', and R, we wanted to check whether further transformations of the dimensions were necessary to obtain a satisfactory description of subjects' perceptions of shell forms. The transformations we tested for each of the dimensions were instances of Stevens' power law $d'(x) = k(d(x))^b$.[5] Since there were three dimensions (V', E' and R), we should estimate the values of six parameters $k_{V'}$, $k_{E'}$, k_R and $b_{V'}$, $b_{E'}$, b_R.

2.3 Pilot study 1

Method, design and procedure

In the pilot studies, subjects were given two types of questions :

(i) Similarity judgements

Subjects were presented with three shell drawings (produced by our program) on a sheet of A4 paper. Each shell was presented from two views, one side and one top view. Subjects were asked to judge whether the shell in the middle was more similar to the shell to the left than to the one to the right.

Subjects answered these questions by putting an X on a 20 cm long line along the bottom of the paper. The endpoints of the line represented the two outer shells, while the line indicated the perceived perceptual distance between those two shells. The X drawn by a subject was to mark the perceptual position of the middle shell in relation to the two outer shells.

(ii) Classification judgements

In addition to the similarity questions, subjects were also asked to provide classification judgements. Subjects first saw two groups of pictures of exemplars of shells (three or four shells in each group). They were then shown a picture of a new shell and asked to classify this shell into one of the two groups. In pilot study 1, each classification question involved two groups of exemplars, drawn from a total of four groups of exemplars.

The classification judgements serve as tests of the predictions of the different categorisation rules discussed above. These tests will be the focus of the two main experiments described in Sections 3 and 4.

All subjects received 25 different similarity questions and 14 classification questions. All shell forms that we used were drawn from a large number of points that we picked out from all over the shell space. In both the similarity and the classification questions, subjects were asked to indicate the similarities on a continuous scale.

Subjects were tested singly. Before starting, the subject was given instructions about the tasks. The shells were then presented on ordinary sheets of paper. First we presented the 25 similarity questions and then the 14 classification questions, both in a random order. We did not put any time limits on the subjects, nor did we measure the time it took them to complete the test.

Subjects

Thirteen subjects (colleagues and friends) participated in this first pilot study. They were not paid for participation.

Response coding

Answers to similarity questions were interpreted in the following way: Assume that a subject placed the X representing the middle shell 3 cm from the endpoint representing the left shell. We then let the middle of the scale be 0, with endpoints –10 and 10, and this distance answer be coded as –7. With this encoding, we calculated the average for each similarity question of test sheet distances over test subjects.

The hypothesis was that this average of the distances on the test sheet should correlate with the corresponding distances in the shell space. We therefore calculated the distance d_{left} in the shell space between the middle and left shells and the distance d_{right} between the middle shell and the right shell. The distance value in the shell space was then calculated as $d_{left} - d_{right}$. This averaged test sheet distances for the similarity questions was then correlated against corresponding shell space distances[6].

Answers to classification questions were analysed in the same way, with one exception. In classification questions, we did not have single left and right shells but instead a group of exemplar shells for the left and right categories. Therefore, when we used the four models PV, GV, NN and AD to calculate the distances from the middle shell to each of the two categories in the shell space. These two distances could then be used to calculate the proportion value that we needed.

It should be noticed that the distances depend on the metric chosen. In our case the metric varied with the values of the vector ($k_{V'}$, $k_{E'}$, k_R, $b_{V'}$, $b_{E'}$, b_R) of constants for the three instances of Stevens' power law that were used. As we correlated the proportion values we also varied the vector so as to maximise the correlation

[5] An alternative method would have been to use multi-dimensional scaling (Shephard 1962a,b, 1987). However, this would have involved asking the subjects to make comparisons of the relative similarity of *pairs* of shell figures, instead of the question concerning triples of shells that was used in our pilot studies.

[6] Throughout all experiments, we used the standard Pearson r-correlation.

coefficient. This was our way of estimating the calibration of the stimuli (as we noted above), which was necessary for the correct calculation of distances as we compared the four categorisation models.

Results

Using the space generated by the dimensions V', E' and R, with the non-transforming power vector (1,1,1,1,1,1), we obtained a correlation coefficient of 0.70. However, by replacing these dimensions with the transformations generated by varying the b-coefficients only, the correlation could be improved considerably so that the vector (1, 1, 1, 0.88, 1, 0.58) yielded a correlation of 0.86. The three b-coefficients confirmed our intuitions about the relative prominence of the three dimensions.

At this early stage, we performed only a preliminary analysis of the answers to classification questions made by the subjects. The results suggest that Nearest neighbour categorisation (NN) and Generalised Voronoi categorisation (GV) produced better predictions than Prototype Voronoi categorisation (PV) and Average distance categorisation (AD): Out of the fourteen classification questions, NN predicted eight of the modal answers among subjects and GV predicted seven clear cases and three borderline cases (where there was no clear modal response).

Discussion

We learned from this pilot study that the top view of the shells did not contribute much information about the shell shapes. Since they also made the interpretation of shapes more difficult because the subjects had to integrate two pictures of a shell to make a judgement, we decided to only use the side view of the shell in the following studies.

Furthermore, letting subjects mark their answers on a continuos scale gave less information than we had hoped for. The subjects tended to cluster their answers around three or four points along the scale, thus making it tantamount to a discrete set of answer options. We therefore replaced the continuos line with discrete alternatives in the following studies.

2.4 Pilot study 2

Method, design and procedure

The motivation for the second study was to make a more precise determination of the vector of constants in the three power laws. To this end, we aimed at choosing examples of similarity judgements that, on the basis of the results of the first study, would be as informative as possible. The different pairs of shells were distributed over the entire shell space and along all three dimensions, but also diagonally through the space. We used 14 different similarity questions. No classification questions were asked.

The shells were again presented on ordinary A4 sheets of paper, but this time each shell was presented only from the side view as in Figure 8. The other difference from pilot study 1 was that in this study, subjects were asked to indicate the similarities in a 3-alternative forced choice response as in Figure 8.

Subjects

Ten people, consisting of a mixture of colleagues and computer science students, mostly male, served as subjects. None was paid for their participation.

Response coding

In calculating the results, we basically used the same correlation between averaged test sheet distances and shell space distances as in pilot study 1. The difference was that in this study, answers were discrete. We therefore had to replace the average test sheet distances with an average of the subjects' 3-alternative answers, each of which was coded as –1, 0 and 1.

Results

As a consequence of aiming at making the similarity judgements as informative as possible about the psychophysical constants, we only obtained a correlation of 0.30 using the space generated by the dimensions V', E' and R, without any transformation. However, the transformations of this space generated by the optimal vector (1, 1.55, 1, 1.72, 1, 0.99) gave a correlation as high as 0.92.

Discussion

The result strongly confirms our hypothesis that it is possible to identify an underlying perceptual shell space of shell forms with sufficient accuracy. Since small changes of the optimal vector resulted in clearly smaller correlation coefficients, we decided to use the calibrated shell space generated by this vector in the main experiments.

Pilot study 2 revealed the calibration of the three dimensions V', E' and R, but it was still not certain that the dimensions are orthogonal. This would be tested in pilot study 3.

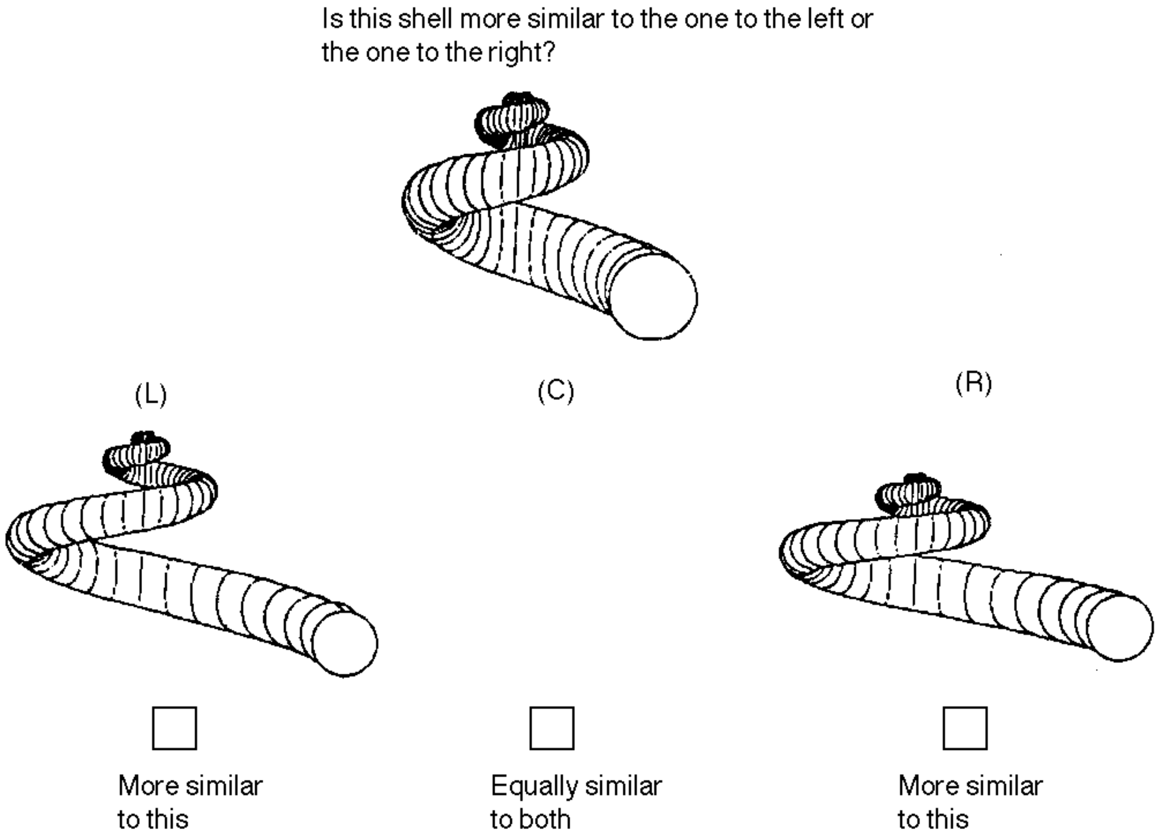


Figure 8. Example of judgements of similarities. Subjects were asked to check one of the squares.

2.5 Pilot study 3

In order to further establish the validity of the perceptual shell space that has been identified, we wanted to check whether the dimensions V', E' and R are orthogonal to one another. If the dimensions are orthogonal, they give linearly independent contributions to the perceptions of shell forms. This means that each of the dimensions can be modified independently of the others. The operationalisation of the orthogonality test can best be explained with the aid of Figure 9.

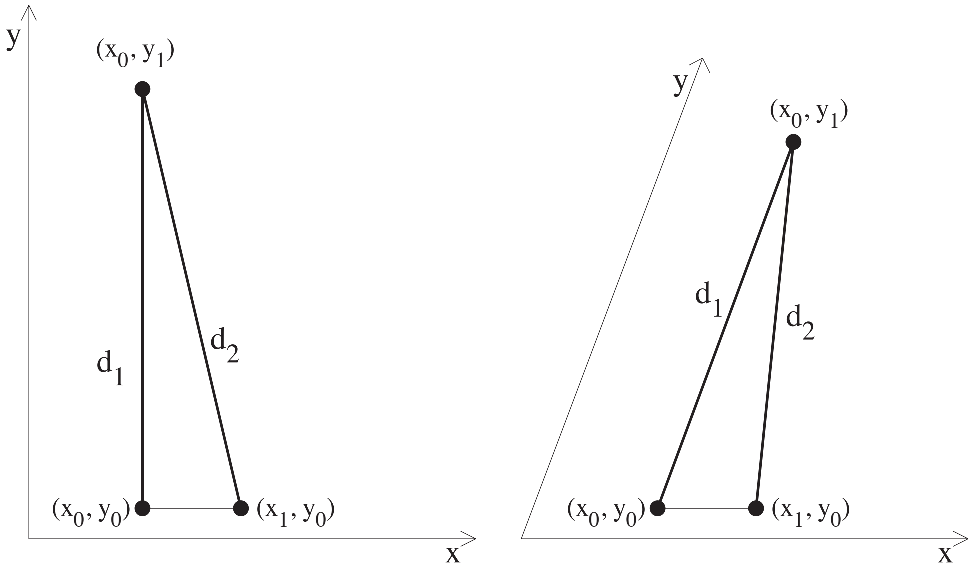


Figure 9. Operational test of orthogonality.

To show that the axes x and y (with a given metric) are orthogonal to one another, subjects are asked to compare the similarity of an object with co-ordinates (x_0, y_1) to objects with co-ordinates (x_0, y_0) and (x_1, y_0) where x_1 is relatively close to x_0. If the axes are orthogonal as in the left part of Figure 9, then (x_0, y_1) should be judged to be more similar to (x_0, y_0) than to (x_1, y_0) since d_1 is smaller than d_2. However, if the axes are not orthogonal as in the right part of the figure, this may not hold since the distance d_2 between (x_0, y_1) and (x_1, y_0) may then be smaller than the distance d_1 between (x_0, y_1) and (x_0, y_0).

Method, design and procedure

In this pilot study, we used 24 different similarity judgement questions but no classification questions. The similarity questions were all of the form indicated in Figure 9, but the metric of the two axes were given by the dimensions V', E' and R.

We used the same procedure as in pilot study 2. The shells were presented on ordinary A4 sheets of paper, and each shell was presented only from the side view, as in Figure 8. Also in this study, subjects were asked to indicate the similarities using a 3-alternative forced choice.

Subjects

The subjects were 10 computer science students, all male, between ages 20 and 28. Again, no one was paid for participation, nor did we put any time limits or measure the time subjects used.

Response coding

The analysis in pilot study 3 was identical to the analysis made in pilot study 2.

Results

Using the uncalibrated space generated from V', E' and R, we obtained a correlation coefficient of 0.72 between the subjects answers and the distances in that space. However, we had not calibrated the similarity questions in the test with respect to the vector that was estimated in pilot study 2. This defect may explain the comparatively low correlation. By investigating the power law transformations of this space, we found that the vector (1, 1.35, 1, 1.5, 1, 1) gave a maximum correlation of 0.80.

Discussion

Even if this result is not identical with the results in pilot study 2, the general pattern of parameters is the same. This result suggests that the dimensions, with the metrics estimated in pilot study 2, have a sufficient degree of orthogonality to be treated as independent perceptual dimensions.

2.6 General Discussion

The data obtained from the three pilot studies suggest that we have identified a psychological space for similarity judgements of shell shapes. The orthogonality of the dimensions in this space is reasonably supported and the metric of the space is sufficiently well identified. On the basis of the results obtained, we felt confident in proceeding to the main classification experiments.

One complication should be mentioned, though. During the pilot tests (as well as during the main experiments) we took notes of the spontaneous comments from the subjects after they had finished the similarity judgements. One frequent comment was that when judging the similarity of different shells, or when classifying them, they looked for whether the shell was "closed" or "open" in the sense that a shell was "closed" if the aperture overlaps with the previous whorl of the shell (cf. the A-shells in Figure 12) and "open" if it was separated from it (cf. the B-shells in Figure 12). This gestalt feature[7] of the shells was added as a fourth dimension in the power law analysis in pilot study 1. The resulting vector is (1, 1, 1, 0.90, 1, 0.60, 0.015, 0.68), where the two last numbers are the k- and b-parameters of the "openness" dimension. This resulted in a correlation coefficient of 0.862 which is almost identical with the correlation obtained without the openness dimension. Furthermore, the values for the other dimensions were very close to the previous values where the openness dimension was not included. Consequently, we did not include this factor in our subsequent analysis.

3. EXPERIMENT 1

The goal of the two main experiments was to evaluate the four classification rules that were presented in section 1.2, i.e., Nearest Neighbour categorisation (NN), Generalised Voronoi categorisation (GV), Prototype Voronoi categorisation (PV), and Average Distance categorisation (AD). The preliminary results from pilot study 1 suggest that the best predictors are NN and GV.

The two main categorisation experiments were based on different set-ups. In the first experiment (to be described in this section), all the exemplars that were used to induce a category for a subject were visible on sheets of papers during the classifications of new test shells, so that exemplars and test shell could be directly compared. We later assumed that such a set-up would favour the NN model, which functions by comparing test shells to the exemplars. In contrast, the second experi-

[7] We call it a gestalt feature, because the values for the three generating dimensions determine whether the shell is "closed" or "open".

ment (described in Section 4), required subjects to learn the category from the exemplars which were presented on a computer screen. These exemplars were not shown when the new shells to be classified were presented.

We expected that the closer a shell lies to a border between two categories in the psychological space the more diversity we will find in the classification answers from the subjects. In order to model this intuition, we defined, for each of the four categorisation models, a *predicted response frequency* for a given category x. Each of the four models provides a distance measure that determines the distance from the stimulus shell x to any category n. For PV, $d_{PV}(x,n)$ is the distance between x and the prototype for category n. For GV, $d_{GV}(x,n)$ is the distance between x and the prototypical circle for category n.[8] For NN $d_{NN}(x,n)$ is the distance between x and the nearest shell in category n. For AD, finally, $d_{AD}(x,n)$ is the average distance between x and the instances of category n.

If $d_M(x,n)$ is the distance function for one of the four models, we then define the predicted response frequency $p_M(x,n)$, i.e., the predicted number of times that subjects will answer that the shell represented by x will belong to category n, by the following equation:[9]

$$p_M(x,n) = \frac{d_M(x,n)^{-1}}{\sum_i d_M(x,i)^{-1}}$$

The main method of evaluating the predictive power of one of the four models is then to compute the correlation between the predicted response frequency and the actual frequency.

Our goal is not to evaluate the overall absolute predictive power of the four models, but to find instances of shells where the four models would predict *different* responses. Since the models in many cases predict the same classification, we had to be careful in selecting the shells to use as stimuli in the classification tasks. The borderlines between categories generated by the different models are in general very close to one another. Consequently, the difference between the predicted response frequencies are not very large. Furthermore, the points in the psychological space where the models yield different predictions sometimes correspond to comparatively unnatural shells. A consequence of this is that the stimuli are difficult to classify for the test subjects. Hence we did not expect any accuracy in the predictions of the models, only that the relative differences would be so large that we could distinguish between the models.

For example, Figure 10 illustrates a two-dimensional cross-section of the psychological space used to generate the shells for trial T15 (and T25 in experiment 2). The co-ordinates for shells of two categories A and B that were shown to subjects are indicated by a's and b's with indices and the shells that the subjects were asked to classify are indicated by Greek letters.[10] The boundaries between the categories induced by the four different models on the basis of the A- and B-shells are shown by lines. As can be seen, none of the test shells is a "clear case" for any of the models since they all lie close to the border lines.

Subjects

A total of 40 subjects, mainly students from the technical university in Lund, took part in the experiment, being paid 50 SwCr (about US$ 7) for their participation. We only noted the age and sex of the subjects. They performed their task without any time constraints. The experiment took between 30 minutes and one hour.

Method, design and procedure

The experiments consisted of ten trials, each containing four or five classification questions. The sets were presented in random order. In the trials where only two categories were involved, one sheet of paper (size A4) with four exemplars of one category of shells, the "A-shells" was put on one side (randomised left or right) of a central sheet, and one sheet with four pictures of another category, the "B-shells", was put on the other side. These two sheets remained in the same place throughout a trial. The sizes of the pictures of the shells generated by the computer program were adjusted so that all illustrations were of roughly the same size.

For each of the ten trials, between four and five pictures of test shells to be classified were then presented in a random order to the subjects. Each test shell was printed on a separate sheet which was put, one by one, between the A- and the B-shells. The subjects were given a 2-alternative forced choice to classify each test shell as an A- or a B-shell.

[8] As noted in Section 1.2, this distance is not real-valued if x is located inside the circle.

[9] This equation was chosen because it has the following properties:
(i) $p_M(x,n) \geq 1$ when $d_M(x,n) \geq 0$, i.e., if x is the prototype for a category it should be classified as such by 100% of the subjects.
(ii) If $d_M(x,n) = d_M(y,n)$, then $p_M(x,n) = p_M(y,n)$, i.e., if two objects are equally far from a prototype, they are equally likely to be classified with that prototype.
(iii) If $d_M(x,n) = d_M(x,m)$, then $p_M(x,n) = p_M(x,m)$, i.e., if an object is equally far from two prototypes, it is equally likely to be classified with either prototype.
(iv) $p_M(x,n) \geq 0$.
(v) $\Sigma_x p_M(x,n) = 1$.
Where prototypes are used in PV and GV, exemplars are used in NN and AD.

[10] The actual shapes of the A- and B-shells are illustrated as the classes "GRE" and "FLA" in Figure 13 below.

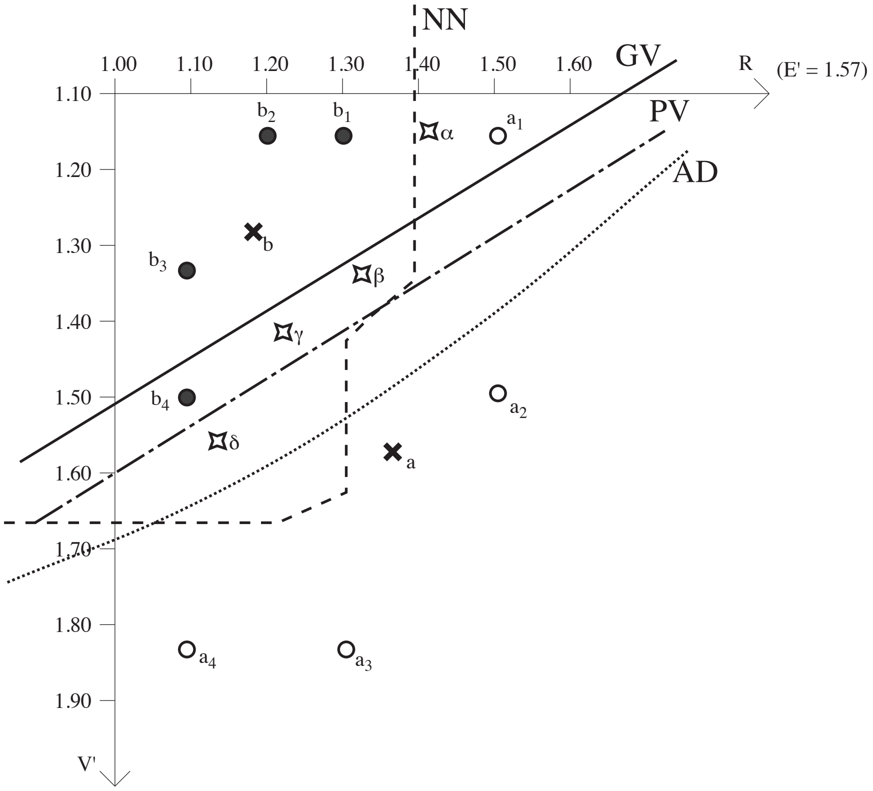


Figure 10. An example of the positions in the shell space of the shapes used in one of the classification trials. The two axes give the values of the dimensions V' and R. The value of E was kept fixed at 1.57. The crosses mark the prototypes of the A- and B-shells respectively. The diamonds mark the test shells, which are named by Greek letters.

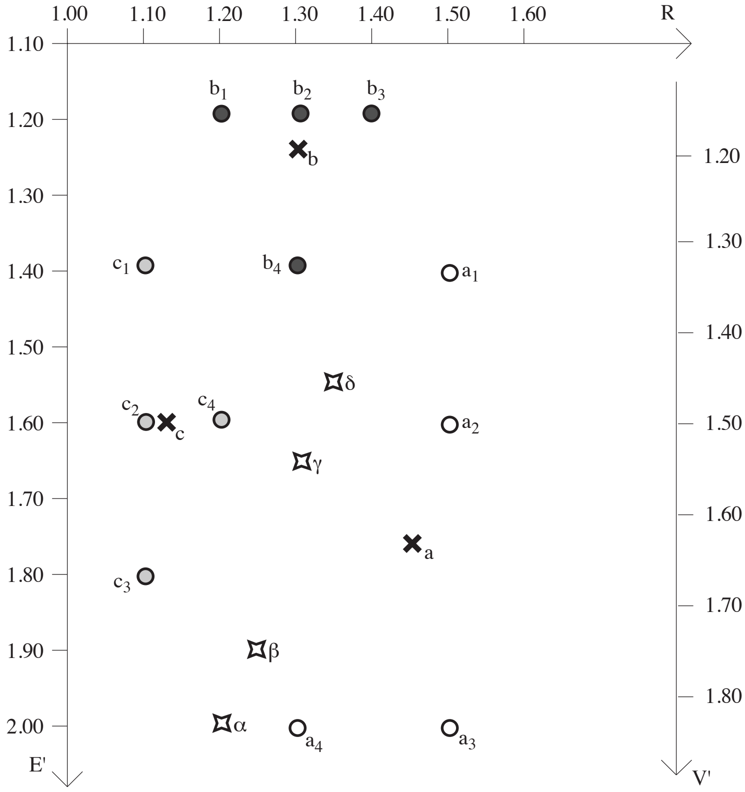


Figure 11. The locations in the psychological space of the exemplar shells in trial T17.3 involving exemplars from three categories and their prototypes. The diamonds mark the test shells which are named by Greek letters. The shells are located on a slanting plane in the shell space. The three co-ordinates can be read from the values on the three axes.

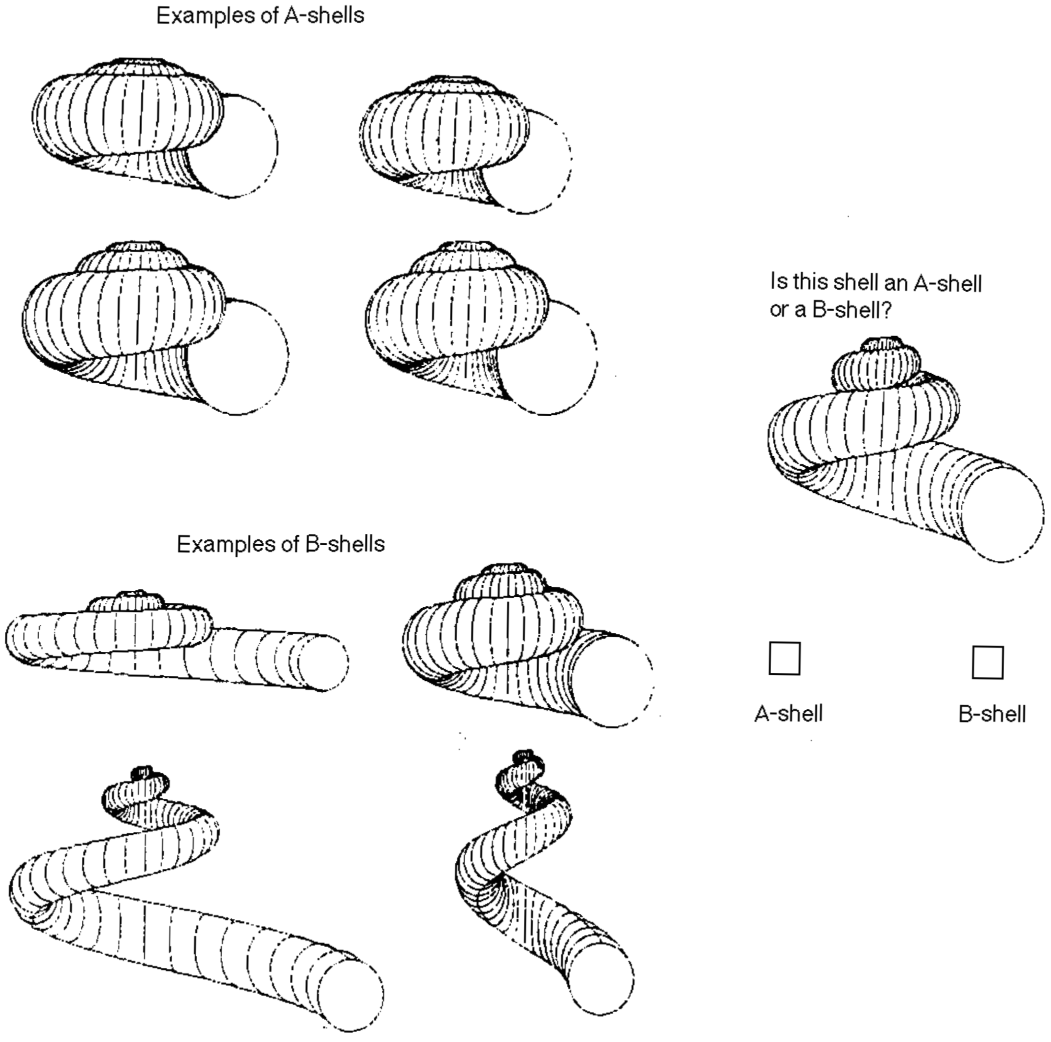


Figure 12. Example of classification of shells (trial T13). Subjects were asked to check one of the squares.

In three of the ten trials, there were examples of three categories, "A-shells", "B-shells" and "C-shells".[11] As before, there were four exemplars of each category. In these trials, the sheet with A-shells was put on one side, the sheet with B-shells was put on the other side, and the C-shells above the middle sheet which contained a test shell to be classified as an A-, B-, or C-shell.

The purpose of these trials was to study the effects of adding a new category to a classification trial. In order to investigate this, two of the trials (T17 and T18) were designed so that half of the subjects were presented with two groups of shells (T17.2 and T18.2), A-shells and B-shell, together with a set of test shells, while the other half of the subjects were presented with the same test shells, but in the context of *three* sets of exemplars, which consisted of the same A- and B-shells as for the other group, but also of a new category of C-shells (T17.3 and T18.3). They were then asked to classify the test shells by a 3-alternative forced choice. The layout of the exemplars and the test shells in T17.3 is presented in Figure 11.

All in all, there were 32 test shells in trials with two categories (T11 – T18.2), 8 test shells in trials T17.3 and T18.3 and 10 test shells in the pure three-category trials T19 and T1A, i.e. 50 different tests.

For each test shell in all of the trials, the subjects were required to make a 2-alternative (or 3-alternative) forced decision, indicating whether the shell on the middle sheet is an A-shell or a B-shell (or a C-shell). They were given free time to make their decision, which consisted of checking one of the relevant boxes on the middle sheet. The experimental set-up is illustrated in Figure 12.

Before the experiment, the subjects were given a practise trial, using categories that were different from those used in the classifications of the main experiment. They were also allowed to ask questions after this exercise trial.

[11] As will be described below, the trials involving three categories differed for different groups of subjects.

Results

For each test shell x, let f(x,A) be the actual frequency of subjects that classified x as an A-shell. For all trials, these frequencies, both for experiment 1 and experiment 2, are presented in the appendix together with a figure describing the location of exemplar shells and test shells in the shell space. For each model M among the four we studied, this number was compared to the predicted answer frequency $p_M(x,A)$. For each model, we computed the average squared sum of errors SSE = $\Sigma_x(p_M(x,A) - f(x,A))^2/n$, where n is the number of test shells involved. Since we were studying trials with both two and three categories, we present in table 2 below, apart from the value SSE for all test shells, also the value SSE2 which gives the average value for classifications involving two categories and SSE3 which covers classifications with three categories. The column "binary" in Table 1 accounts for the number of cases out of the 50 different tests when the category predicted by a model was different from what the majority of the subjects answered.[12] For all columns in Table 1, the smallest values indicate the best predictive value of a model.

Model	*SSE2*	*SSE3*	*SSE*	*Binary*
PV	0.30	0.22	0.27	20
GV	0.42	0.24	0.36	20
N N	0.21	0.22	0.21	8,5
AD	0.35	0.26	0.32	15

Table 1. SSE values for the four models with respect to classifications involving two categories (SSE2), classifications involving three categories (SSE3), all classifications (SSE), and classifications different from that of the majority of subjects (binary).

These results suggest that NN has the best predictive power, while GV has the worst. However, the differences are not significant at the 5% level (t-test). Also, as will be seen in Table 2 below, different classification trials resulted in very different degrees of fit for the different models.

It should be remembered that we selected test shells that were close to the border lines between the categories that were generated by the four different models. One hypothesis that we formed at this stage was that when classification is "difficult," i.e., when the test shell is perceived to be as similar to the A-shells as to the B-shells, subjects tend to focus on the A- or B-shells that are most similar to the test shell, which means that they employ the NN model so that the other models would not be applicable.

In this experiment the subjects could *see* the different examples of A- and B-shells during all trials. This is an advantage for the NN model since it benefits from visible examples. An easily accessible strategy (which may be unconsciously used by the subjects) is that in order to classify a test shell as an A- or a B-shell, a subject identifies the example among the *displayed* A- and B-shells that is *most similar* to the test shell and then classifies the test shell according to this example. In experiment 2, subjects could not see the examples of A- and B- shells when they were asked to classify test shells. Hence, the similarity judgements necessary for the NN model could not be based on direct visual information.

In order to operationalise the degree of difficulty involved in the classifications, we compared the degree of unanimity among the subjects. In the eight (out of twelve) classification tests where only two categories (A- and B-shells) are considered, the unanimity can be measured as the deviation from 50% B-responses to the performances of the models. The *polarity value* for a test shell x is thus defined as $|0.5 - f(x,B)|$. This measure is only applicable for the trials that contained two categories. It seems difficult to define a polarity measure for the remaining four classification trials where three categories are involved.

Restricting ourselves to the eight classification trials that contained two categories, we can then calculate the correlation between the polarity values and the SSE values for the four models. The results are presented in the following table:

Trial	*PV*	*GV*	*NN*	*AD*	*Polarity*
T11	0.18	0.04	0.63	0.51	0.33
T12	0.56	0.95	0.09	0.25	0.20
T13	0.14	0.01	0.10	0.30	0.21
T14	0.40	0.12	0.27	0.59	0.28
T15	0.34	1.01	0.09	0.24	0.26
T16	0.22	0.74	0.07	0.13	0.12
T17.2	0.38	0.44	0.21	0.36	0.20
T18.2	0.20	0.08	0.23	0.43	0.44
Corr.	–0.28	–0.53	0.54	0.66	

Table 2. SSE values of the four models for the eight trials in Experiment 1 and the polarity values for the trials. Correlations between SSE values for the four models and the polarity values are also shown. For each trial, the best value, i.e., the lowest SSE value, has been underlined.

As can be seen GV is best in four cases while NN is best in the other four. In the remaining four classification trials, GV gives the best predictions for one, NN for one, and in two cases there is no clear winner. Hence, PV and AD failed to provide the best predictions for

[12] This means treating the trial as a binary classification and ignoring the magnitude of the values $p_M(x,A)$ and f(x,A) respectively. The advantage of the binary classification is that it is independent of the assumptions connected with $p_M(x,A)$.

any of the classification trials.[13]

The correlation values represent the correlation between the polarity column and the corresponding columns for each of the four models. Low polarity is an indication that the subjects have some form of difficulty in deciding to which category a test shell belongs. One source of difficulty is that some classes of example shells formed "unnatural" categories in the sense that the examples within a such a class showed a great variation in shapes (cf. the B-shells in Figure 11[14]). Another source of difficulty was that the test shells that were presented were located close to the classification borderlines generated by the different models. If polarisation is high, this means that neither of these difficulties should have occurred.

Regarding the trials having variants with three categories, i.e., trials T17.3 and T18.3, the results strongly support that the relations between the A- and B-category remained stable after the addition of the C-shells. If the relative proportions of subjects that classify the test shells as A- and B-shells respectively are correlated between the two and three category versions of T17 and T18, the correlations are 1.00 for T17 and 0.91 for T18.

Discussion

Since GV (and PV) presume category formation, while NN and AD only presume individual examples, one should therefore expect that high polarisation should be correlated with good predictions from GV (and PV). As can be seen from table 2, the correlation values strongly support this hypothesis.

In concluding this section, we note that experiment 1 gave us valuable information concerning the predictive success of the four categorisation models. The relative accuracy of the prediction gave further support for our assumption that the shell space that we had identified in the pilot studies has great explanatory power when the classification behaviour of the subjects is analysed. The experiment also helped us form new hypotheses for further testing.

One cognitively important limitation of the experimental layout used in experiment 1 is that the subjects always had *visual* access to the different examples of A- and B-shells that we wanted them to use to form the concepts "A-shell" and "B-shell." Since the examples are in front of the subjects during the whole test phase, there was nothing that forced them to remember or in any other way internalise the examples. As noted above, this may explain why the NN model could predict a comparatively high proportion of the test shells, since the availability of externally given exemplars made it easy to employ the model.

Furthermore, we wanted to further test the hypothesis that when classification is "difficult" in the sense that the test shell is perceived to be as similar to the A-shells as to the B-shells, subjects tend to focus on the A- or B-shells that are most similar to the test shell. Thus, NN should have a comparative advantage over GV in the "difficult" classification trials, while we expected GV to perform well in "normal" trials.

4. EXPERIMENT 2

Subjects

Forty people from Lund, mainly students from the technical university, were recruited to the experiment. They were paid 50 SwCr (about US$ 7) for their participation, which took between half an hour and an hour. The participation was anonymous – we only noted the age and sex of the subject.

Method, design and procedure

This time the shells were presented on a computer screen and not on paper sheets. We partly reused the trials of experiment 1 and partly rebuilt them. The experiment was run in two phases. First, examples of A- and B-shells (and sometimes C-shells) were presented to the subject and he/she was trained to remember the exemplars so that they could be correctly classified. The purpose of this identification learning was to induce the subjects to create some form of internal representation of the two categories.[15]

Only after this phase was completed, the subject would see and classify new test shells in the same way as in experiment 1, but now the test shells were presented alone on the computer screen, without the presence of the exemplars from the learning phase. Our hypothesis was that since this experimental set-up blocked the visual availability of the different examples, we would obtain a better evaluation of the predictive power of the four categorisation models, in particular the results of GV and NN.

The program was run a Macintosh IIci with a 21" black and white screen. Before the experiment, the subjects were given instructions on the screen and a practice trial involving both phases in order to familiarise

[13] In trials T11, T13, T15, and T18.2, the differences between the best and the second best model are significant at the 5% level. However, in all trials the differences between the best and the worst model are significant at the 5% level.

[14] Our speculation as to why subjects had difficulties with this trial was that the distinction between the categories was determined by the diameter dimension, which seems to be difficult to perceive.

[15] Our methodology is similar to that of Nosofsky (1988), except that he did not show any new stimuli in the test phase (which he calls the transfer phase).

them with the program. Then a total of six concept formation trials were presented in random order. Five of the trials contained two concepts and the remaining one (T26 below) contained three concepts. In each trial five or six test shells were presented. The total number of test shells was 34. Some of the trials from experiment 1 were repeated in order to check whether the two experimental methods resulted in significant differences in the subjects' responses.

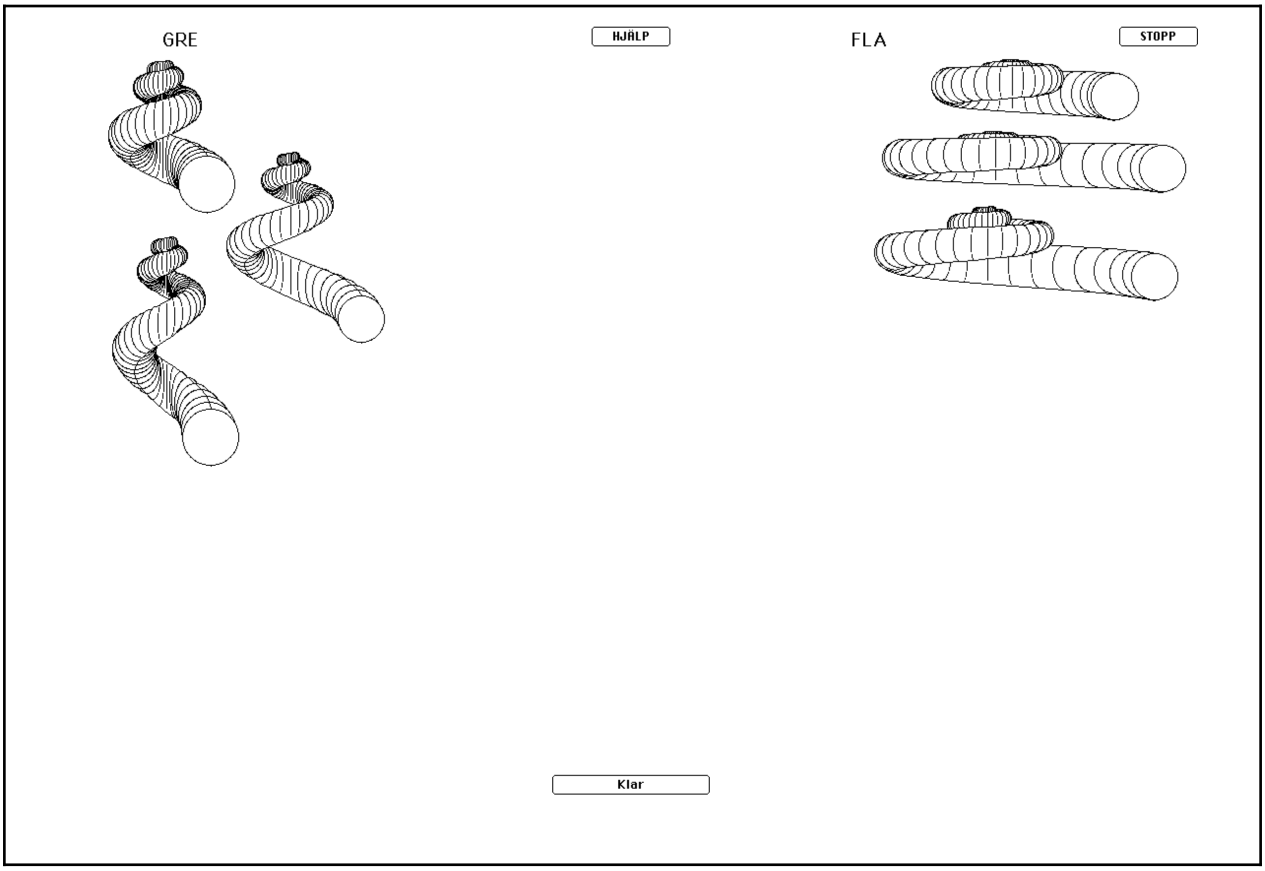


Figure 13. Picture of a screen during the presentation of the two classes of exemplars.

Each trial was divided into two phases. In the first learning phase, a subject was presented with four exemplars each of two (or sometimes three) categories of shells and s/he was asked to try to remember the examples so that they can be correctly classified. Instead of calling the categories "A" and "B" as in the previous experiment, the groups of examples were now labelled with randomly generated Swedish nonsense syllables (cf. Figure 13).[16] When the subject judged that he/she could recognise the exemplar shells, he/she clicked on a "ready" button on the screen. The program automatically measured the time between the appearance of the shells and the pressing of the button. (A "help" button was also present on the screen.)

Then the examples that the subject had just studied were presented one by one, in randomly ordered blocks consisting of one picture each of the example shells. The subject's task in this phase is forced choice recognition with feedback. The subject was instructed to determine what kind of shell was shown (without seeing the original pictures). This was done by clicking on a response button that was named with the same nonsense syllables as before, and which were placed on the same side of the screen as the corresponding examples (cf. Figure 14). A "HELP" button was also present here. The response time of the subject was measured by the program. The subject was informed by the program whether or not the response was correct. This was done, with a one second delay, by displaying the words "not correct"[17] in big letters on the screen. The identification learning procedure continued until the subject showed an error rate of less than 10%. However, all subjects were required to go through at least two blocks each containing all example shells. After passing this error rate, a subject was considered to have learned the example shells sufficiently well.

[16] The syllables were generated by Sigurd's (1983) phonotax algorithm for Swedish which resulted in strings that are clearly pronounceable, but usually have no meaning, like "hunk", "skuns", "smesk".

[17] In Swedish: "INTE RÄTT".

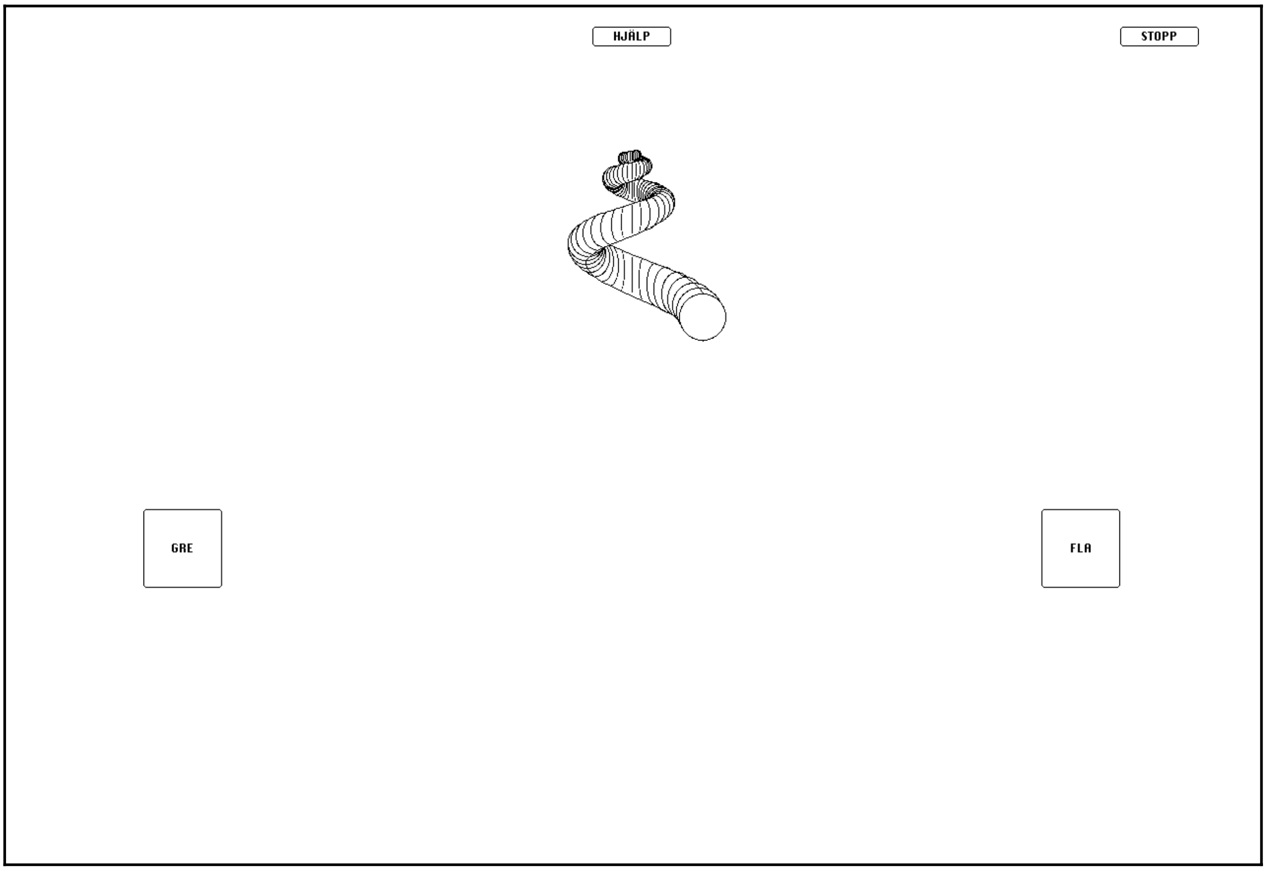


Figure 14. Picture of the computer screen during the learning phase

In the second phase of a trial, the test phase, the subjects were presented with a series of test shells that were always different from the example shells. The order of the test shells were randomised. The subject was asked to classify the test shell as one of the two (or three) categories that had been studied in the learning phase. The set-up of the screen was the same as in Figure 14. Again the response time was measured.

Performance measures

As in experiment 1, many of the classification trials were "difficult" in the sense that some classes of example shells (the A-, B- and C-examples) formed unnatural categories. Similarly, another source of difficulty was that the test shells that were presented were located close to the classification borderlines generated by the different models.

The use of a computer for the presentation and classification trials enabled us to measure several factors that we judged to be correlated to the difficulty of the trials. First of all, we hypothesised that the unnaturalness of a class of examples would be reflected in the difficulty of learning to correctly classify these examples in the first phase of a trial. In order to operationalise this we measured two variables: (1) *Learning rounds*, which was defined as the number of presentations of the example shells that had to be made before a subject passed the 90% correctness level. (2) *Learning time*, which is the time a subject spent studying the screen with the two or three classes of examples before pressing the "ready" button. (3) *Learning errors*, which was defined as the total number of classification errors during the learning phase.

Furthermore, we supposed that both the unnaturalness of a class of examples as well as the closeness to the category borderlines would have an effect on the subjects' performance during the test phase of a trial. This was measured by the following two variables: (4) *Classification time*, i.e., the time it took for a subject to classify a test shell as an A- or B-shell (or C- shell) after it was presented on the screen. (5) *Polarity*, which is the same measure as was used in experiment 1. All responses and reaction times were measured automatically by the program and stored on a log file which was then used for statistical analysis.

Results and discussion

Again, the answer frequencies are presented in the appendix together with a figure describing the location of exemplar shells and test shells in the shell space.

	PV		*GV*		*NN*		*AD*	
Trial	*SSE*	*Bin.*	*SSE*	*Bin.*	*SSE*	*Bin.*	*SSE*	*Bin.*
T21	0.33	1	0.33	1	0.40	1	0.33	1
T22	0.07	1	0.09	1	0.31	2	0.15	2
T23	0.44	1	0.81	2	0.12	0	0.29	2
T24	1.20	2	0.64	0	0.85	0	1.54	3
T25	0.58	2	0.58	2	0.47	1	0.68	2
T26	1.66	2	1.62	2	1.50	0	1.23	2
Average:	0.70		0.67		0.59		0.70	

Table 3 SSE values and binary choice error (Bin.) with respect to the four models for the six categorisation trials of experiment 2.

SSE

First of all, when evaluating the results the predictive accuracy of the models should be compared. This was done by calculating the average SSE values for each trial. The results are given in table 3. The columns "Bin" refer to the number of binary choice errors, i.e., the number of cases when the category predicted by the model was different from what the majority of the subjects answered. As in Table 1, low numbers mean accurate predictions.[18]

It should be noted again that the absolute values of the errors are not so important since we have not striven to make a "representative" selection of categorisation trials and test shells in order to "confirm" some of the models. On the contrary, our choices have been made according to falsificationist principles so that we have tried to select trials that would test the limits of the models. Hence it is the relative results of the models that are interesting. Nevertheless, the predictive values of the models were quite good. Out of the 34 test examples PV made the same classification prediction as the majority of the subjects in 25 cases, GV in 26 cases, NN in 30 and AD in 22 cases.

Performance analysis

Our general hypotheses was that NN performed better than GV when the categorisations were "difficult". Thus the results of the models should be compared to the performance measures that were introduced above. Table 4 contains the results. Note that polarity is not defined for T26 which contained three categories.

Trial	*L. Rounds*[19]	*L. Time*	*L. Errors*	*Cl. time*	*Polarity*
T21	2.00	26	1	3.2	0.09
T22	2.75	83	31	4.0	0.30
T23	2.60	108	39	4.6	0.33
T24	2.35	41	8	2.2	0
T25	2.50	68	31	2.8	0.14
T26	3.30	143	156	4.8	*

Table 4. Performance measures and polarity for the six categorisation trials of experiment 2.

Already a quick glance at the table shows that the measures are strongly correlated. The correlations between the first four measures are given in the following matrix:

	L. Round	*L. Time*	*L. Error*	*Cl. time*
L. Rounds	1	0.94	0.92	0.70
L. Time		1	0.88	0.85
L. Error			1	0.69

Table 5. Correlations between the four performance measures of experiment 2.

As can be seen, the correlation between the three learning variables is very high. Each of these will thus function as a reliable indicator of the difficulty for the subjects in forming the categorisations from the learning examples presented to them.

Polarity values could not be calculated for the sixth trial so the correlation values between this variable and the others are based on the first five trials only, as presented in Table 6.

[18] In trials T23 and T24, the differences between the best and the second best model are significant at the 5% level. In T22 and T25 the differences between the best and the worst model are significant at the 5% level. In T21 no differences are significant.

[19] As mentioned above, all subjects were required to go through at least two rounds, before starting on the test phase.

	L. Rounds	*L. Time*	*L. Error*	*Cl. time*
Polarity	0.71	0.89	0.83	0.90

Table 6. Correlations between polarity and the four performance measures for the five first trials of experiment 2.

Also these correlations are high. which strengthens the value of the five measures as indicators of the difficulties of the classification of the test shells.

Now a central result for our investigation is the correlations between, on the one hand, the SSE of GV and NN, which can be seen as measures of the predictive powers of the two models, and, on the other hand, the five performance values. Table 7 provides the results.[20]

Measure	*PV*	*GV*	*NN*	*AD*
L. Rounds	–0.26	0.29	–0.01	–0.24
L. Time	–0.37	0.36	–0.48	–0.45
L. Error	–0.38	0.37	–0.36	–0.44
Cl. Time	–0.73	–0.10	–0.90	–0.82
Polarity	–0.73	0.07	–0.67	–0.79

Table 7. Correlations between the SSE of the four models and the performance measures together with the polarity value.

These figures suggest that there is a general negative correlation between the difficulty of the classifications trials and the degree of misfit of PV, NN, and AD. In other words, these models have a smaller misfit when the trials are difficult. The converse is true of GV, although the correlations are not quite as strong in this direction. As can be seen from the classification time and the other measures in Table 4, the trials T21, T24, and T25 are clearly less difficult than T22, T23, and T26. If we call the first group the "easy" trials and the second the "difficult," then the average SSE values for the four models are as in Table 8.

Trials	*PV*	*GV*	*NN*	*AD*
Easy	0.70	0.52	0.57	0.85
Difficult	0.70	0.84	0.61	0.54

Table 8. SSE values for the four models with respect to two groups of trials.

The general conclusion is that GV performs well as a model of the subjects' behaviour when the categories are comparatively natural while NN and AD seem to be the best models when the examples presented to the subject are difficult to weld into a category.[21]

Error analysis

We conclude the analysis of experiment 2 by some speculations about the difficulty to internalise a category based on the distribution of its exemplars. In the test we logged the errors that the subjects made when tested on their internalisation of a category. From these data it is clear that not all exemplars are equally easy to learn. Some of the test shells are placed on a position in the shell space that for the subjects must feel less natural for that category than other positions. A clear example of this is given by T25 (cf. Figures 10 and 13) where the learning errors made by test subjects for different exemplars are as follows:

T25:	A:	a_1: 8,	a_2: 1,	a_3: 0,	a_4: 4
	B:	b_1: 6,	b_2: 1,	b_3: 1,	b_4: 11

The figures strongly suggest that the exemplars 1 and 4 in both categories are more difficult to incorporate in the category, while exemplars 2 and 3 are more clearly core exemplars of the two categories. In Figure 10, we saw that GV also misclassified two exemplars, viz. a_1 and b_4, exactly the same exemplars which were most often misclassified by test subjects in the learning phase. In contrast, NN, by definition misclassifies no learning exemplars, which is clearly not the way test subjects behaved. This speaks against NN as being the classification method that best describes the behaviour of the subjects.

Test subjects did not make many misclassifications in the "easy" trial T24:

T24:	A:	a_1: 0,	a_2: 0,	a_3: 2,	a_4: 0
	B:	b_1: 0,	b_2: 0,	b_3: 4,	b_4: 2

Still, the problematic positions are the "fronts" where the two exemplar extensions meet (cf. Figure 15), while the concept cores are on the "backsides".

In the difficult trial T23, the errors made by test subjects show no pattern of difficult versus core areas in the exemplar extension, which suggests that they have general problems with picking out the characteristics of the categories.

T23:	A:	a_1: 6,	a_2: 8,	a_3: 5
	B:	b_1: 9,	b_2: 8,	b_3: 3

[20] This table is based on the results from trials T21 - T25 only. The three-category trial T26 was excluded since the SSE values are not comparable to those of the two-category trials.

[21] However, none of the differences in Table 8 are significant at the 5% level.

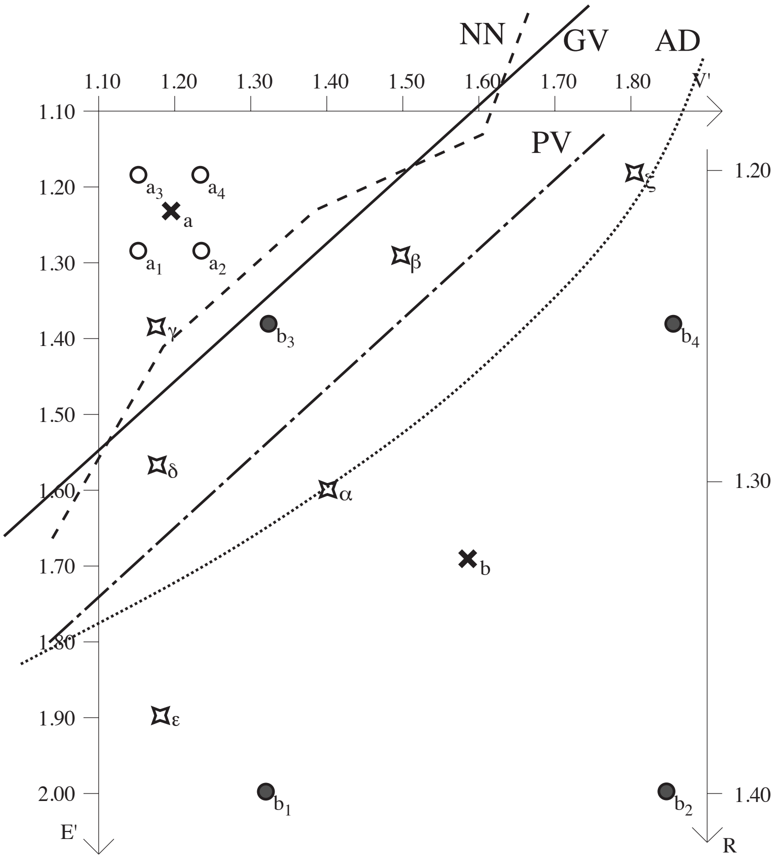


Figure 15. The positions in the psychological space of the shells in trial T24. The diamonds mark the test shells which are named by Greek letters. The shells are located on a slanting plane in the shell space. The three co-ordinates can be read from the values on the three axes.

In trial T26, with its three categories, test subjects made the largest number of misclassifications during the learning phase.

T26:	A:	a_1: 5,	a_2: 5,	a_3: 6,	a_4: 5
	B:	b_1: 47,	b_2: 10,	b_3: 10,	b_4: 9
	C:	c_1: 14,	c_2: 24,	c_3: 11,	c_4: 10

In T26, the A category seems the easiest to internalise. The B category is torn between categories A and C; 34 of the misclassifications of b_1 placed the exemplar in the A category, all of b_3 and b_4 placed them in the C category. The C category is difficult do distinguish from the B category; all C misclassifications but two placed the exemplars in the B category. It is interesting to notice once again that GV misclassifies b_1, the exemplar which was most often misclassified by test subjects.

This analysis of the learning errors made by the subjects suggests that GV is the model that best predicts their behaviour, even though the model does not always give the best predictions of their classifications of the test shell.

5. GENERAL DISCUSSION

The two main experiments were based on different modes of presenting the exemplars of the categories and they used different methods for testing the categorisations. In experiment 1, the subjects could see the exemplars of the different categories while classifying the test shells, but in experiment 2, they were trained to correctly classify the category exemplars before classifying the test shells which were then presented without the presence of the exemplars.

Since some of the classes of exemplars were the same in the two experiments (T11 was identical to T22, T12 the same as T23, and T13 the same as T24) it is of some interest to compare the results of the subjects' classifications of the test shells in the two experiments. In total there are 12 test shells (out of 32 test shells of two-category trials in experiment 1 and 29 in experiment 2) that were used in both experiments where we can directly compare the classification frequencies.

One might perhaps expect that since it was more difficult to classify test shells in experiment 2 than in experiment 1 where the exemplars were visible, the SSE values should have been larger in experiment 2. However, an investigation of the responses reveals somewhat conflicting differences between the two experiments. On the one hand, the SEE values did increase for all models (by 37% for PV, 13% for GV, 21% for NN, and 24% for AD) when we look at the average over the 12 shells that were classified in both experiments. On the other hand, the SSE values decreased for *all* models for T11/T22 and T12/T23. It was only the large increase in T13/T24 that accounted for the increase in the overall average of the SSE values.

Furthermore, we expected, for the same reasons, that the polarity of the responses would decrease in experiment 2, i.e., that subjects would be less unanimous in their classifications. This was indeed true for T11/T22 and T12/T23, but for T24 the unanimity was drastically greater than for the identical T13.[22] We have no explanation of this fact.

The general conclusion to be drawn from the two experiments is that none of the four models is superior to the others in explaining the subjects' classifications in all situations. However, in the process of testing the models, we have discovered that the difficulty in forming new categories and remembering them can vary enormously. We have used a number of performance measures (learning rounds, learning time, learning errors, and classification time) that serve as good indicators of the difficulty of forming a category from some exemplars. The strong correlation between these measures support their validity.

If the level of difficulty of a categorisation trial is taken into account a clearer pattern in the performance of the four models can be discerned. If the exemplars used to generate a category form a "natural kind" so that categorisation is "easy", then the Generalised Voronoi model is the best predictor of the test results. On the other hand, if categorisation is "difficult", as measured by the indicators used above, then the Nearest Neighbour model seems to perform best.

The main purpose of this article has been to present and test some models of concept formation in dimensional spaces. We have tried to establish that the shell space is an example of such a dimensional space which is particularly interesting for investigations of human concept formation since the stimuli that are generated by our graphic program have a high degree of ecological validity in comparison to other kinds of stimuli that have been used in categorisation experiments. Perhaps the most important finding of our experiments is that the level of "difficulty" of categorisation, as measured by the indicators we have identified, has been shown to play a crucial role for the performance of the models. This notion merits further attention in future studies of categorisation.

Leaving the shell shapes used as stimuli in this article, a central question is whether the methods and models used here can be applied also to other kinds of stimuli. We believe that they have a great potential for further applications. One argument in favour of this is that Shepard (1987) and others have convincingly demonstrated that for many different cognitive domains it is possible to identify, with high reliability, an underlying psychological space. And as regards many of the stimuli used in categorisation studies, e.g. faces in Reed (1972) and Pittenger and Shaw (1975), cornered lines in Ashby and Gott (1988), cups in Labov (1973), semicircles in Nosofsky (1986), and colours in Nosofsky (1988), the models studied here should be readily applicable.

ACKNOWLEDGEMENTS

Research for this paper has been supported by the Swedish Council for Research in the Humanities and Social Sciences and by the Erik Philip-Sörensen Foundation. We wish to thank Ulf Gärdenfors for shell expertise, Karl Radeborg for valuable advice on the experimental set-up, and David Stansvik and Simon Winter for their work during the experiments. We are also grateful to Paul Hemeren and Måns Holgersson for their detailed comments on an earlier draft.

REFERENCES

Ashby, F. G. and Gott, R. E. (1988): "Decision rules in the perception and categorization of multidimensional stimuli," *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 33–53.

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

Gärdenfors, P. (1990a): "Induction, conceptual spaces and AI," *Philosophy of Science, 57*, 78–95.

Gärdenfors, P. (1990b): "Frameworks for properties: Possible worlds vs. conceptual spaces," in *Language, Knowledge and Intentionality,* L. Haaparanta, M. Kusch, and I. Niiniluoto, eds. (*Acta Philosophica Fennica, 49*), 383–407.

Gärdenfors, P. (1992): "A geometric model of concept formation," in *Information Modelling and Knowledge Bases III,* S. Ohsuga et al., eds., IOS Press, Amsterdam, 1–16.

Gibson, J. J. (1979): *The Ecological Approcch to Visual Perception*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Holmqvist, K. (1988): "Aspects of parameterizing concept representations," manuscript, Department of Philosophy, Lund University.

Holmqvist, K. and Gärdenfors, P. (in preparation): "Convexity in generalized Voronoi tessellations", Department of Cognitive Science, Lund University.

[22] The great polarity in T24 is the main source of the large increase in the SSE values.

Labov, W. (1973): "The boundaries of words and their meanings," pp. 340–373 in J. Fishman (ed.), *New Ways of Analyzing Variation in English*, Georgetown University Press, Washington, DC.

Lakoff, G. (1987): *Women, Fire, and Dangerous Things*, University of Chicago Press, Chicago.

Quine, W. V. O. (1969): "Natural kinds," in *Ontological Relativity and Other Essays*, Columbia University Press, New York, NY, 114–138.

Mervis, C. and Rosch, E. (1981): "Categorization of natural objects," *Annual Review of Psychology,* 32, 89–115.

Nosofsky, R. M. (1986): "Similarity, frequency, and category representations," *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 54–65.

Nosofsky, R. M. (1988): "Attention, similarity, and the identification–categorization relationship," *Journal of Experimental Psychology: General, 115*, 39–57.

Pittenger, J. B. and Shaw, R. E. (1975): "Aging faces as viscal–elastic events: Implications for a theory of nonrigid shape perception," *Journal of Experimental Psychology: Human Perception and Performance, 1*, 374–382.

Posner, M. I. and Keele, S. W. (1968): "On the genesis of abstract ideas," *Journal of Experimental Psychology, 77*, 353–363.

Raup, D. M. (1966): "Geometric analysis of shell coiling: General problems," *Journal of Paleontology, 40*, 1178–1190.

Reed, S. K. (1972): "Pattern recognition and categorization," *Cognitive Psychology, 3*, 382–407.

Rosch, E. (1975): "Cognitive representations of semantic categories," *Journal of Experimental Psychology: General, 104*, 192–233.

Rosch, E. (1978): "Prototype classification and logical classification: The two systems," *New Trends in Cognitive Representation: Challenges to Piaget's Theory*, ed. E. Scholnik, Lawrence Erlbaum Associates, Hillsdale, NJ, 73–86.

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

Shepard, R. N. (1962b): "The analysis of proximities: Multidimensional scaling with an unknown distance function. II.," *Psychometrika*, *27*, 219–246.

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

Shin, H. J. and Nosofsky, R. M (1992): "Similarity-scaling studies of dot-pattern classification and recognition," *Journal of Experimental Psychology: General, 121*, 278–304.

Sigurd, B. (1983): *Möt språkforskningen*, Liber förlag, Lund.

Smith, E. och Medin, D. L. (1981): *Categories and Concepts*, Harvard University Press, Cambridge, MA.

APPENDIX

In this appendix we show the psychological spaces of trials that were used in the two main experiments. Each space includes:

1) The dimensions V', E' and R of the space. Sometimes one dimension was kept fixed on one value.

2) The positions of exemplar shells which are always dots. The filling of the dot as well as the indices show what category the exemplar was assigned to.

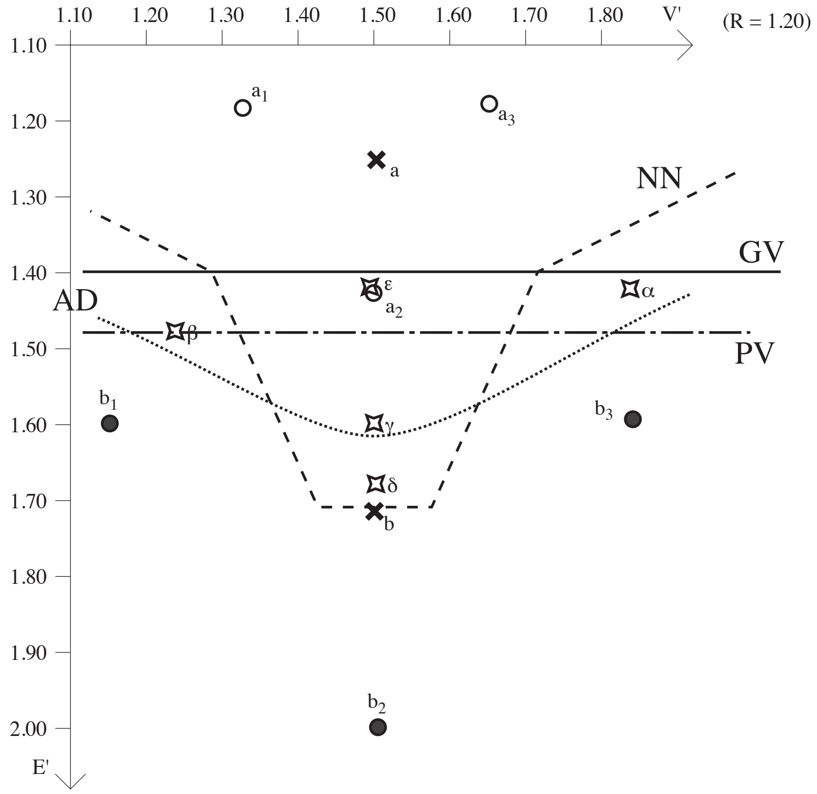
3) The positions of the calculated prototypes (for the PV and GV models) which are marked as crosses. The index tells what category the prototype belongs to.

4) The positions of test shells are marked as diamonds with Greek letter indices.

5) The border lines between those regions in the psychological space that the four models PV, GV, NN and AD assign to each of the two or three categories in the trial.

6) A table of the recorded answers, where for each test shell we note the percentage of subjects who assigned it to the categories involved in the trial.

Observe that in several cases the psychological spaces were similar for the two main experiments, with only one or two test shells differing.



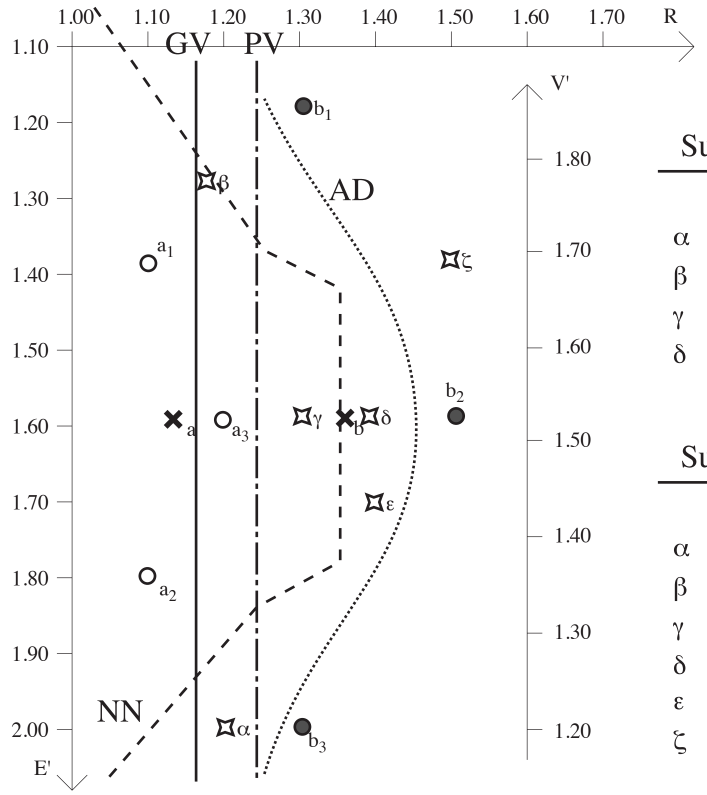
Subjects (in %), Exp. 1

	A	B
α	40.0	60.0
β	20.0	80.0
γ	5.0	95.0
δ	2.5	97.5

Subjects (in %), Exp. 2

	A	B
α	55.0	45.0
β	35.0	65.0
γ	35.0	65.0
δ	15.0	85.0
ε	80.0	20.0

Figure A1. Trial T11 of experiment 1 and trial T22 of experiment 2.



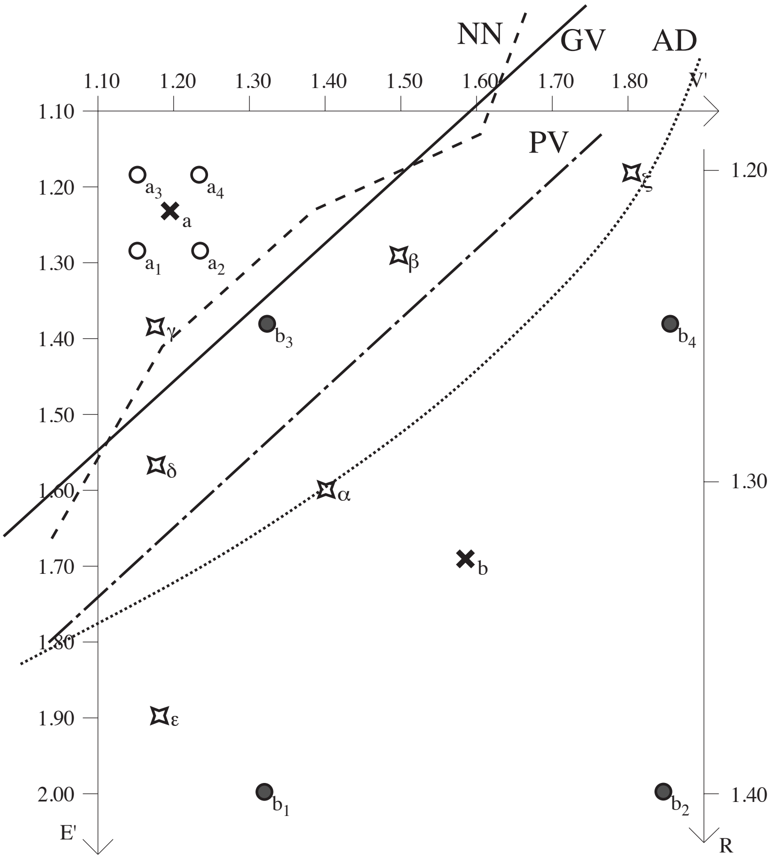
Subjects (in %), Exp. 1

	A	B
α	2.5	97.5
β	57.5	42.5
γ	65.0	35.0
δ	60.0	40.0

Subjects (in %), Exp. 2

	A	B
α	20.0	80.0
β	55.0	45.0
γ	55.0	45.0
δ	50.0	50.0
ε	35.0	65.0
ζ	5.0	95.0

Figure A2. Trial T12 of experiment 1 and trial T23 of experiment 2.



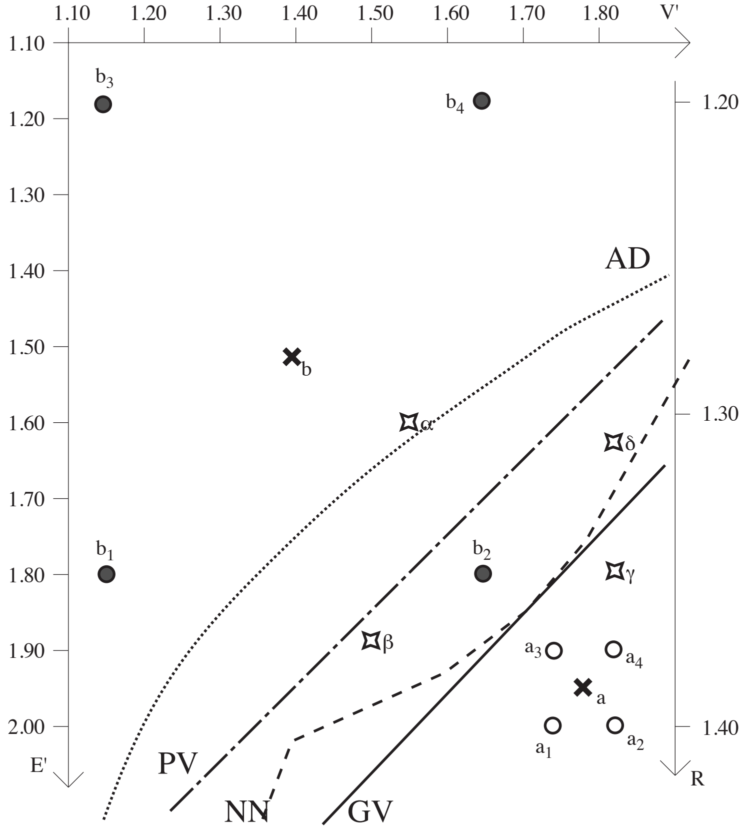
Subjects (in %), Exp. 1

	A	B
α	10.0	90.0
β	40.0	60.0
γ	67.5	32.5
δ	32.5	67.5

Subjects (in %), Exp. 2

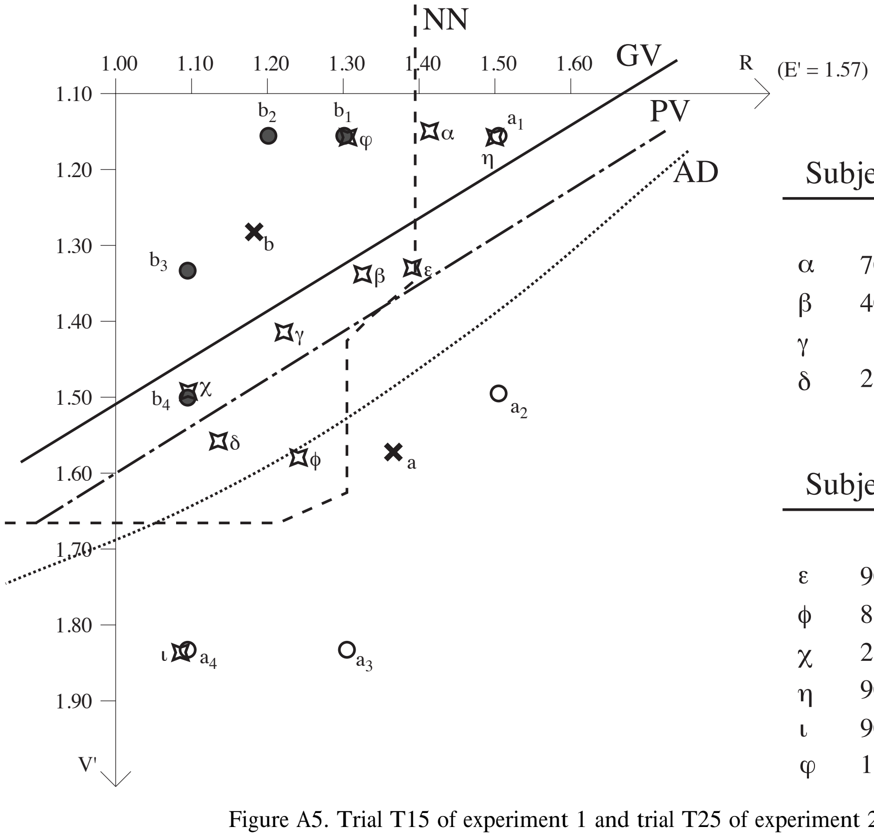
	A	B
α	0.0	100.0
β	0.0	100.0
γ	100.0	0.0
δ	0.0	100.0
ε	0.0	100.0
ζ	0.0	100.0

Figure A3. Trial T13 of experiment 1 and trial T24 of experiment 2.



Subjects (in %), Exp. 1	A	B
α	7.5	92.5
β	2.5	97.5
γ	62.5	37.5
δ	42.5	57.5

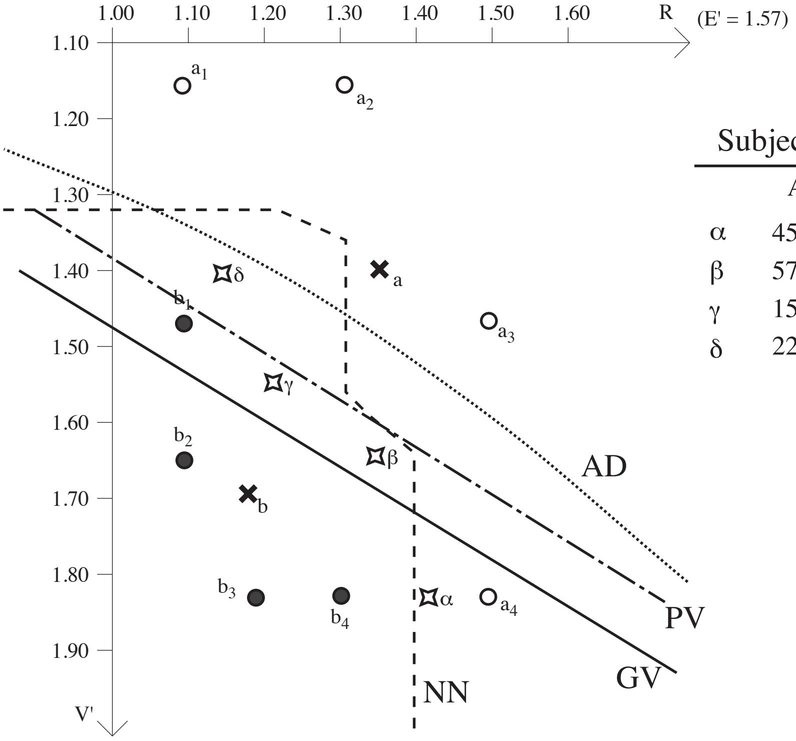
Figure A4. Trial T14 of experiment 1.



Subjects (in %), Exp. 1	A	B
α	70.0	30.0
β	40.0	60.0
γ	5.0	95.0
δ	22.5	87.5

Subjects (in %), Exp. 2	A	B
ε	90.0	10.0
ϕ	85.0	15.0
χ	25.0	75.0
η	90.0	10.0
ι	90.0	10.0
φ	15.0	85.0

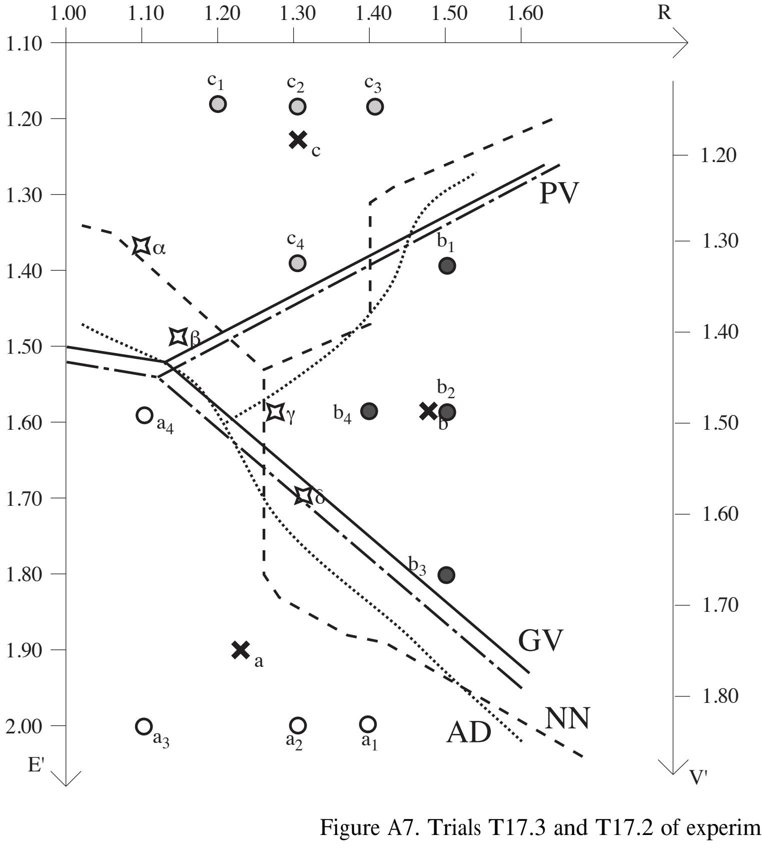
Figure A5. Trial T15 of experiment 1 and trial T25 of experiment 2.



Subjects (in %), Exp. 1

	A	B
α	45.0	55.0
β	57.5	42.5
γ	15.0	85.0
δ	22.5	77.5

Figure A6. Trial T16 of experiment 1.



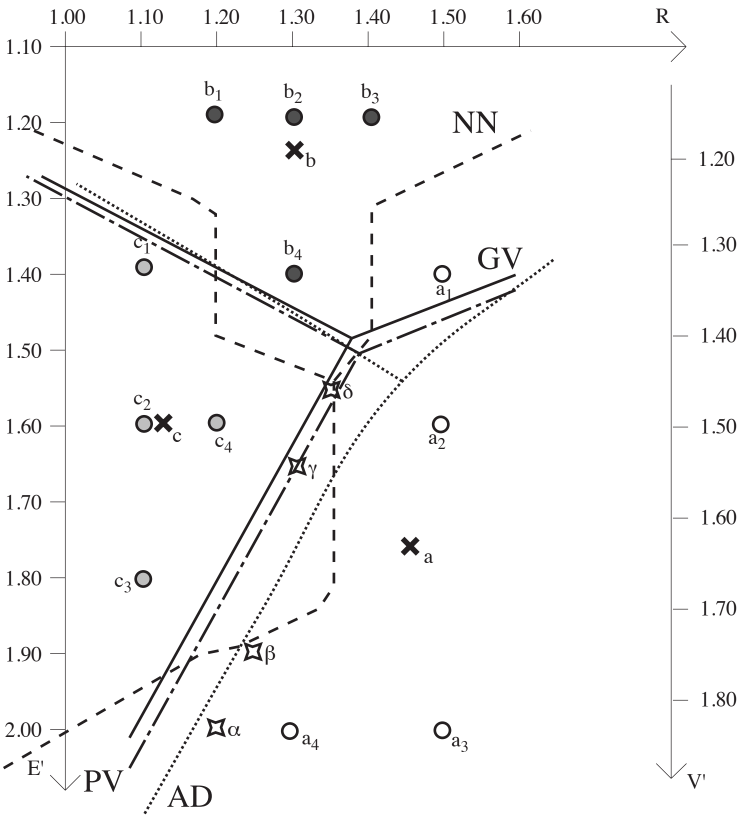
Subjects (in %), Exp. 1

	A	B	C
α	10.0	0.0	90.0
β	35.0	5.0	60.0
γ	5.0	95.0	0.0
δ	20.0	80.0	0.0

Subjects (in %), Exp. 1

	A	B
α	75.0	20.0
β	70.0	25.0
γ	20.0	80.0
δ	5.0	95.0

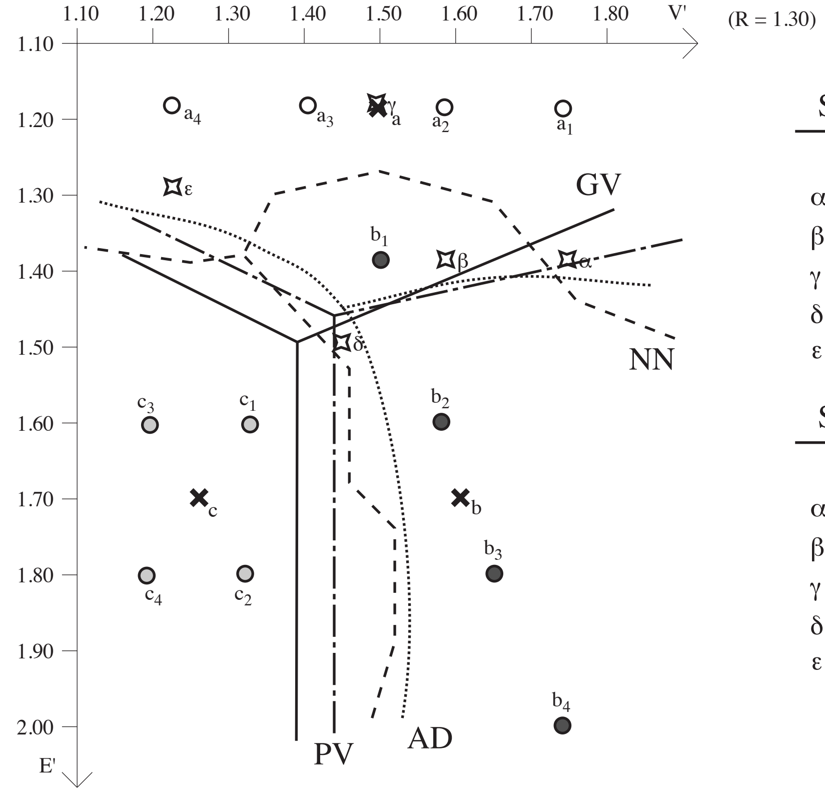
Figure A7. Trials T17.3 and T17.2 of experiment 1.



	A	B	C
Subjects (in %), Exp. 1			
α	90.0	0.0	10.0
β	85.0	0.0	15.0
γ	35.0	0.0	65.0
δ	30.0	10.0	60.0

	A	B
Subjects (in %), Exp. 1		
α	100.0	0.0
β	100.0	0.0
γ	100.0	0.0
δ	75.0	25.0

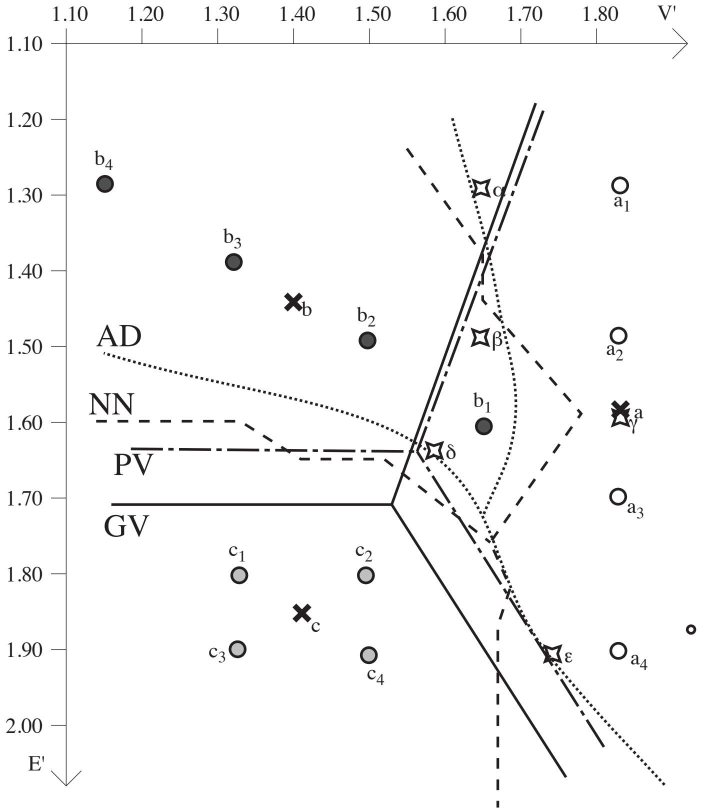
Figure A8. Trials T18.3 and T18.2 of experiment 1.



	A	B	C
Subjects (in %), Exp. 1			
α	65.0	35.0	0.0
β	55.0	45.0	0.0
γ	90.0	7.5	2.5
δ	7.5	37.5	55.0
ε	55.0	10.0	35.0

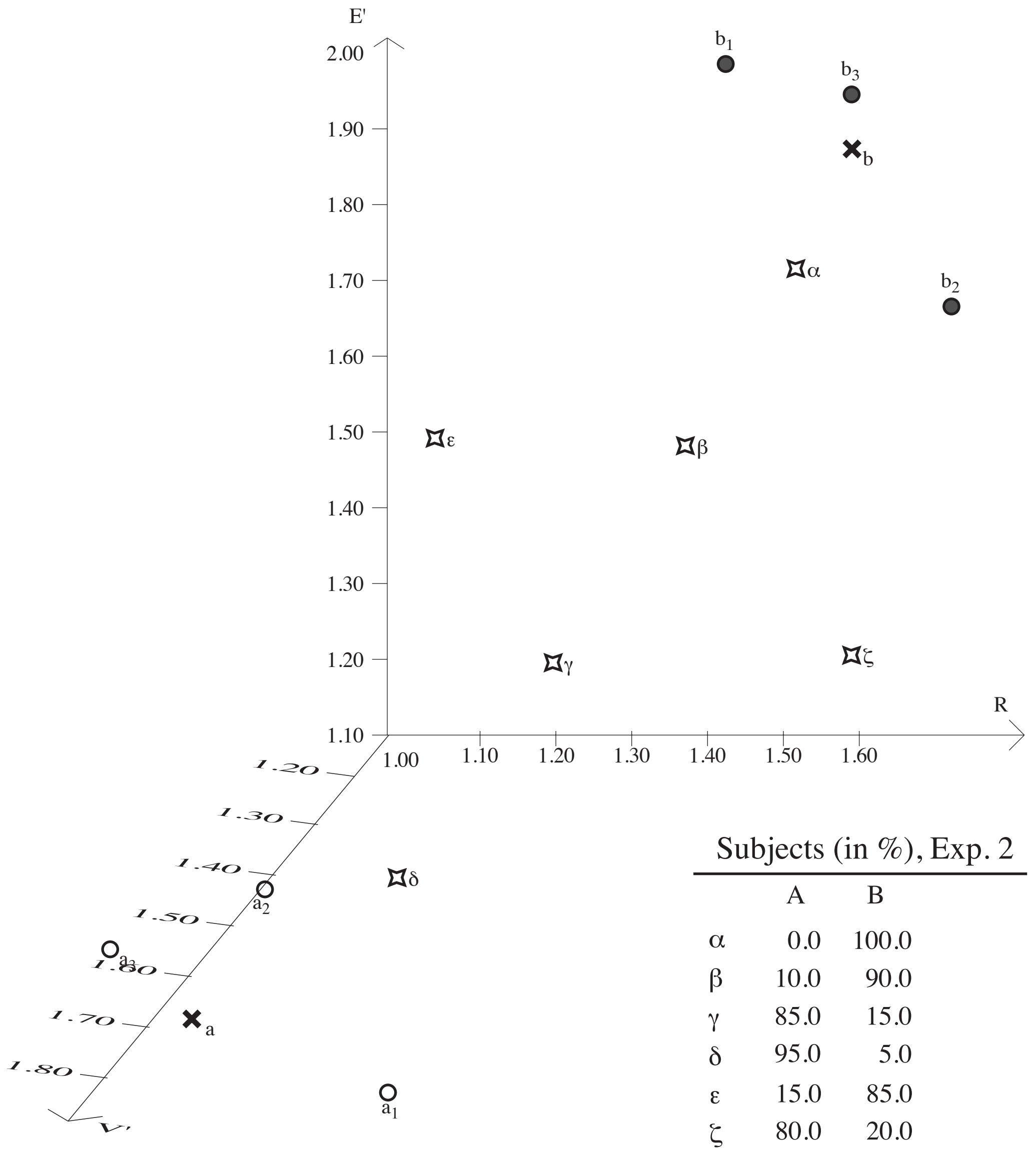
	A	B	C
Subjects (in %), Exp. 2			
α	55.0	40.0	5.0
β	15.0	80.0	5.0
γ	95.0	0.0	5.0
δ	0.0	55.0	45.0
ε	45.0	20.0	35.0

Figure A9. Trial T19 of experiment 1 and trial T26 of experiment 2.



Subjects (in %), Exp. 1	A	B	C
α	82.5	17.5	0.0
β	27.5	70.0	2.5
γ	82.5	15.0	2.5
δ	25.0	55.0	20.0
ε	22.5	0.0	77.5

Figure A10. Trial T1A of experiment 1.



Subjects (in %), Exp. 2

	A	B
α	0.0	100.0
β	10.0	90.0
γ	85.0	15.0
δ	95.0	5.0
ε	15.0	85.0
ζ	80.0	20.0

Figure A11. The trial T21 of experiment 2. This is a fully 3D test, not a subplane. We have therefore not indicated any borders in the depiction.