Perception & Psychophysics
1999, 61 (7), 1320-1335

Judgments of perceptual groups: Reliability and sensitivity to stimulus transformation

BRIAN J. COMPTON and GORDON D. LOGAN
University of Illinois, Champaign, Illinois

The reliability of subjects' judgments of the groups present in dot patterns and the sensitivity of those judgments to stimulus transformation were assessed. The subjects indicated the groups that they saw within random dot patterns, and each judgment was compared with those of other subjects and with their own judgments for related presentations. Within subjects, each pattern appeared in an initial presentation, an identical repetition, and a transformed state (a rotation or a change in scale). Within-subjects judgments were more reliable than between-subjects judgments. An interpretation of within-subjects results was made in relation to predictions made by a formal algorithm of grouping by proximity (the CODE algorithm), which assumes that grouping by proximity is invariant over transformations such as rotations or changes in scale. A slight cost to transforming the patterns was found. The implications for CODE and for using grouping judgments as data are discussed.

The purpose of this article is to examine the reliability of subjects' grouping judgments and the invariance of those judgments over transformation. The issue is important practically and theoretically. It is important practically because researchers use grouping judgments as a basis for selecting stimuli and testing hypotheses about theories of grouping. It is important theoretically because many theories of grouping have assumed, implicitly or explicitly (e.g., van Oeffelen & Vos, 1982), that grouping processes are invariant over certain types of transformation, such as rotation and changes in scale.

It is important to ask these questions within the context of a formal theory of grouping, so that the possibility that a single pattern can be grouped in different ways can be considered. For example, Figure 1 might be seen in a number of different ways—for example, as consisting of five groups of 3 or 4 dots each or of two groups containing 7 and 11 dots, respectively (see Palmer, 1977). These two organizations are related, because the two-group organization involves only joining (and never separating) dots that had belonged to different groups in the five-group organization. A psychologically relevant theory of grouping should allow these two different organizations to be defined as being more closely related to each other than to many other possible organizations. Our investigation of the reliability of grouping judgments will take into account these types of relations between possible organizations of a pattern.

The influence of Gestalt organizational principles, such as grouping by proximity, in the study of visual processing has been limited by their fundamentally phenomenological nature: They appear convincing when used as demonstrations but are difficult to state formally in a way that allows predictions about the organizations subjects will see in novel patterns. There have been steps toward establishing formal approaches to some Gestalt organizational principles (e.g., Compton & Logan, 1993; Kubovy, 1994; Prytulak, 1974; van Oeffelen & Vos, 1982, 1983). One reason the principle of grouping by proximity is an important principle to model is its broad scope: It deals with relative locations among elements in a pattern, which is an aspect of organization that is always present in patterns containing more than one element.

The CODE Algorithm

The contour detection (CODE) algorithm was proposed by van Oeffelen and Vos (1982) as an objective method for determining the proximity-based groups within dot patterns. Compton and Logan (1993) made an empirical test of CODE and offered a modification and extension of it on the basis of their results. It has since been used by Logan (1996) and Logan and Bundesen (1996) as part of a broader theory of visual attention.

CODE assumes that the way individuals will group patterns is invariant over changes in orientation (such as rotation and reflection) and scale. One goal of the present article is to test this invariance assumption of CODE, by determining whether subjects are as likely to group a pat-

This research was supported by Grants BNS 91-09856 and SBR 94-10406 to G.D.L. from the National Science Foundation. We thank Gail Heyman and three reviewers for helpful comments on an earlier version of the manuscript and Julie Delheimer for assistance with data collection. Correspondence concerning this article should be addressed to B. J. Compton, Department of Psychology, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0109 (e-mail: bcompton@psy.ucsd.edu), or to G. D. Logan, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign, IL 61820 (e-mail: glogan@s.psych.uiuc.edu).

—*Accepted by previous editor, Myron L. Braunstein*

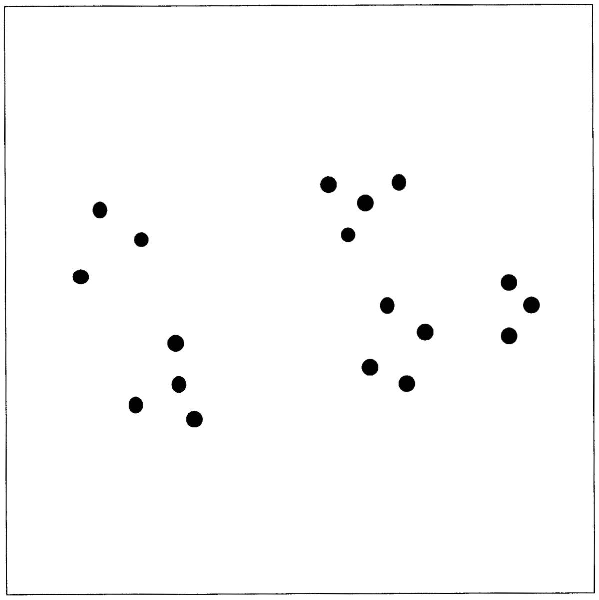

Figure 1. A dot pattern that can be grouped in multiple ways.

tern the same way when it is presented again as a rotation or a reflection or is changed in scale, as when it appears again in its original form. As such, the present experiments provide a falsification test for CODE.

It is the Compton and Logan (1993) version of CODE that was used in this study, and it will be described first. The Compton and Logan version of CODE contains both a data-driven component and a conceptually driven component.

Data-driven component. In the CODE algorithm, each dot exerts an influence on neighboring dots that declines monotonically with distance. This influence will be referred to as the *strength of grouping*. When the strength of grouping in a particular region of the pattern is high, dots that lie within that region are likely to be defined as belonging to a single group. Conversely, when the strength of grouping in a particular region is low, dots are likely to remain ungrouped. The strength of grouping associated with each dot in a pattern is graded, with strength maximized at the location of the dot and diminishing as distance from the dot increases. Strength of grouping is represented as a Laplace distribution, centered at the dot. Figure 2 shows the strength gradients associated with a one-dimensional dot pattern consisting of five dots (labeled A, B, C, D, and E) located along a line. In this example, the stimulus is indicated along the x dimension, with the strength gradients being shown in the y dimension. The strength gradient for each dot is represented by a curve (thin line) centered immediately above it. The areas under the strength gradient curves are equal, with the shapes of the curves being determined by the standard deviation of each strength gradient function, which is equal to one half the distance from the dot to its nearest neighbor.

Once the strength gradients of each dot are determined, they are summed at each location to form a curve, as represented by the thick line in Figure 2. For two-dimensional dot patterns, such as the one shown in Figure 3A, the strength gradients are symmetrical and located in the z dimension, above the stimulus plane. For two-dimensional patterns, the strength gradients, when summed, create a *CODE surface*. Figure 3B shows the CODE surface for the dot pattern shown in Figure 3A. The creation of the CODE surface is the final component of the data-driven part of the CODE algorithm.

Conceptually driven component. Once data-driven processes create the CODE surface, conceptually driven processes can generate different judgments by varying the height of a threshold. (These are referred to as conceptually driven processes because the threshold is thought to be controllable by the perceiver; see Compton & Logan, 1993.) Above-threshold regions of the CODE surface define the groups: All the dots within a single above-threshold region belong to the same group. Returning to the one-dimensional example in Figure 2, the five horizontal lines marked I, II, III, IV, and V represent five different thresholds that define the groups {ABCDE}, {AB}{CDE}, {AB}{DE}{C}, {DE}{A}{B}{C}, and {A}{B}{C}{D}{E}, respectively. Each of these five thresholds defines a unique way in which to segment this dot pattern, defining a *judgment*. (This term will be applied to how patterns are grouped by the CODE algorithm, as well as by subjects.) When the threshold is lowest, a single group is formed that includes all five dots. Then, as the threshold moves up, dots A and B break away from dots C, D, and E, and so on, until the threshold is at its

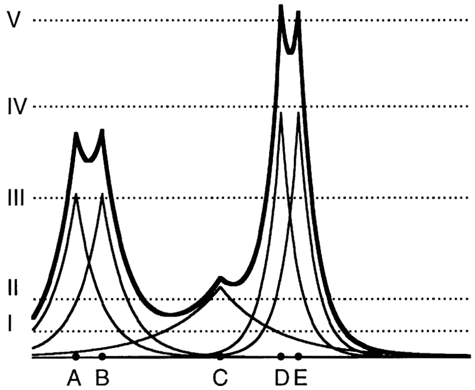

Figure 2. The CODE algorithm as applied to a one-dimensional dot pattern. Dots A, B, C, D, and E are located along the x-axis, and strength gradients associated with each dot (dotted lines) and with their sum (solid line) appear in the y dimension. Five thresholds (labeled I–V) are shown that define the judgments {ABCDE}, {AB}{CDE}, {AB}{DE}{C}, {DE}{A}{B}{C}, and {A}{B}{C}{D}{E}, respectively.

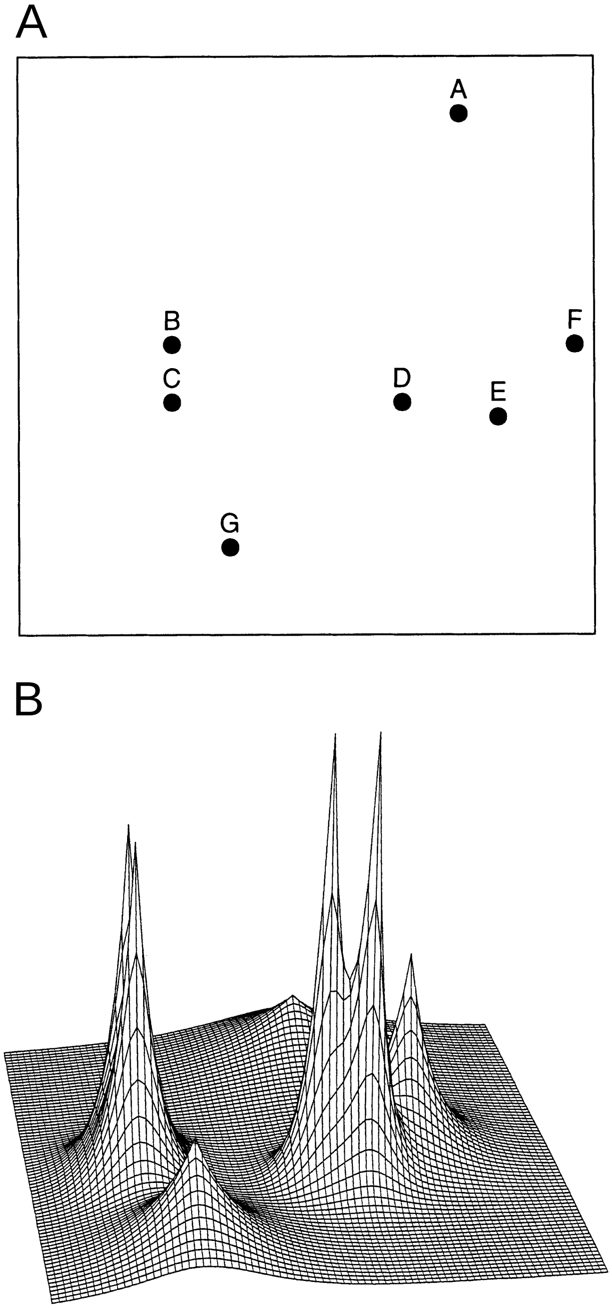

Figure 3. (A) A two-dimensional dot pattern, and (B) its CODE surface.

maximum value, at which point no dots are grouped together.

The number of different judgments that the CODE algorithm is capable of identifying within each pattern is equal to the number of elements in the pattern. If one considers the lowest threshold to be the starting point, then at each meaningful increment of the threshold, the number of potential judgments that CODE can identify before all of the dots are broken apart (and no groups remain) is reduced by one.

Figure 4 shows how thresholds are used in the case of two-dimensional patterns. Figure 4A shows a single threshold applied to the CODE surface that was shown in Figure 3B. This threshold specifies the judgment {BCG} {DEF}{A}. Figure 4B shows two of the seven judgments that CODE specifies for the same pattern: {BCG}{DEF} {A}, which is indicated by dotted lines, and {BC}{DE} {A}{F}{G}, indicated by solid lines.

The data-driven component of CODE generates the CODE surface, and the conceptually driven component applies different thresholds to generate multiple judgments for a single pattern (assuming that at least two dots are present).

It should be noted that CODE determines the groups within a pattern solely on the basis of the relative distances among the elements. As a result, CODE is invariant over changes in rotation, reflection, and scale.

Differences Between the Two Versions of CODE

The Compton and Logan (1993) version of CODE differs from the original formulation proposed by van Oeffelen and Vos (1982) in three ways. The first modification was to extend CODE by adding a top-down component that assumes that the subject selects from a small set of CODE-generated groupings for a given pattern, rather than generating a single grouping for a given pattern, which is a strictly bottom-up approach. The CODE algorithm, as originally formulated by van Oeffelen and Vos (1982), indicated only one way of grouping any given dot pattern.

Van Oeffelen and Vos (1982) represented strength gradients by using a normal distribution. As a second modification, Compton and Logan (1993) used Laplacian rather than Gaussian strength gradients, for three reasons. First, Compton and Logan found the Laplacian distribution to work as well as or better than the Gaussian in accounting for subjects' grouping judgments. Second, Shepard (1987) has argued that exponential functions (such as the Laplacian) best characterize generalization gradients in psychological space. Third, the Laplacian distribution is more tractable than the Gaussian analytically.

Van Oeffelen and Vos (1982) rescaled the strength gradient associated with each dot, so that all the strength gradients were equal in height, which eliminated the equivalence in volume among strength gradients having different spread functions. Compton and Logan (1993) found that this step significantly reduced the ability of CODE to account for subjects' grouping judgments, so the third modification to CODE entailed omitting this step.

In their evaluation of CODE, Compton and Logan (1993) found a set of configurations of parameters that most successfully described subjects' grouping judgments. No single configuration within this set was better than the others. The algorithm parameters selected for use in the present study were chosen from among this set

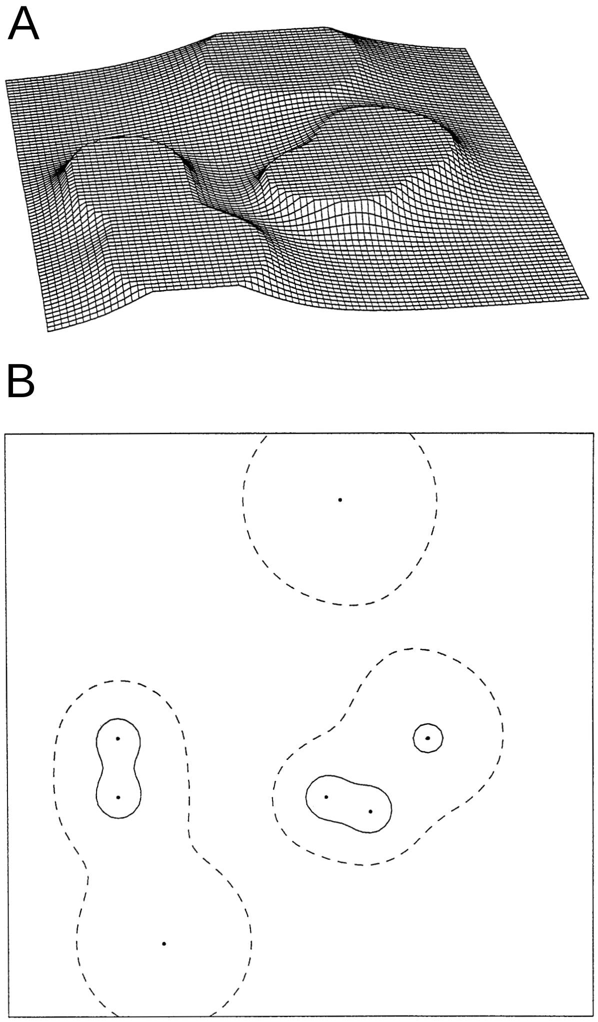

Figure 4. (A) The CODE surface for the dot pattern seen in Figure 3A, with a threshold applied that defines the judgment {BCG}{DEF}{A}. (B) An overhead view of two thresholds that define the judgments {BCG}{DEF}{A} and {BC}{DE}{A}{F}{G}, indicated with dotted lines and solid lines, respectively.

of configurations of algorithm parameters. The following configuration of parameters (presented in terms of the five parameters evaluated by Compton and Logan) was used in the present study. (1) Each dot's spread function was represented by a Laplace distribution. (2) The standard deviation was set separately for each dot to (3) one half of the distance from that dot to its nearest neighbor (Compton & Logan did not find this to be an especially important parameter; Logan, 1996; Logan & Bundesen, 1996). In building the CODE surface, (4) the sum, rather than the maximum value, of all spread functions was used. And (5) the spread functions were not rescaled so that their heights were equal. It should be noted that not all of these parameters had a major influence on the performance of CODE. For example, Logan and Logan and Bundesen have made use of an essentially equivalent implementation of CODE in which the same standard deviation is used for all elements.

Kubovy (1994) has provided a formal model to account for the perceived organizations of patterns composed of dots arranged in lattices. Kubovy has correctly pointed out that CODE is unable to produce the proper organization for patterns of this type. Kubovy refers to the grouping principle at work as *grouping by proximity*, but it might alternatively be considered to be good continuation. Patterns of this type, in which some elements are grouped not with their nearest neighbors, but with relatively more distant dots as part of a compelling linear organization, are outside the scope of the CODE approach, as presently formulated. CODE excels at detecting clusters, not lines.

The Present Study

In this study, the following questions about the nature of grouping judgments are addressed. What is the reliability of judgments among different subjects who group the same patterns? What is the reliability of judgments made by a single subject who groups the same patterns on two different occasions? Is the CODE algorithm correct in its characterization of grouping judgments as being invariant over reflections, rotations, and changes in scale?

We asked subjects to make grouping judgments for a series of dot patterns (the *standards*). In Experiments 1 and 2, we presented the dot patterns again, both in their original form (the *repetitions*). In Experiment 3, the subjects first grouped a set of standards, as they had in Experiments 1 and 2, and then received a surprise recognition memory test. This was done to help rule out the possibility that the subjects' judgments of repetitions and transformations were affected by their recognizing that the patterns were the same as or related to ones they had already seen.

In making their grouping judgments, the subjects were instructed to indicate as many or as few groups as they saw and were told that it was acceptable to leave dots ungrouped if they did not seem to belong to any group, but that dots could not belong to more than one group. This group will be referred to as the *selection constraint*. According to the selection constraint, there are multiple potential judgments for patterns having more than one element. For example, for any pattern with 3 elements (labeled A, B, and C), there are five possible judgments: {A} {B} {C}, {AB} {C}, {AC} {B}, {BC} {A}, and {ABC}. Table 1 shows, for dot patterns having up to 10 elements, the number of logically possible judgments under the selection constraint (see Compton & Logan, 1993, for the mathematical basis for these calculations).

EXPERIMENT 1

In Experiment 1, we investigated the reliability of grouping judgments and the sensitivity of the judgments to rotations of the stimulus. All the subjects grouped the same set of patterns, which allowed between-subjects

Table 1
Number of Logically Possible Judgments for Dot Patterns Having up to 10 Elements

Number of Elements	Possible Judgments
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4,140
9	21,147
10	115,975

agreement to be addressed. The subjects grouped each pattern in the first block of trials and then again later, allowing within-subjects reliability to be addressed. The subjects also grouped rotations of the patterns they had seen in the first block. In Experiment 1, we sought to determine whether the grouping processes underlying subjects' judgments would be invariant over rotation. Because invariance over rotation is a result of a basic operating assumption of CODE, a test of whether grouping judgments are invariant over rotation is a test of CODE itself.

Method

Subjects. The subjects were 30 introductory psychology students at the University of Illinois, who received course credit for their participation.

Apparatus and Stimuli. The stimuli were presented on a Dell UltraScan computer monitor controlled by a Dell 486P/33 computer. The subjects sat approximately 50 cm from the computer screen, but viewing distance was unconstrained. Responses were entered with a Dell two-button computer mouse.

The stimuli consisted of patterns containing 7, 8, 9, or 10 white dots placed on a black background. The patterns were based on an imaginary 40 × 40 square grid, with grid location 20, 20 corresponding to the center of the screen. The dots measured 0.3 cm in diameter (approximately 0.3º of visual angle). The entire grid measured 12 cm on a side (approximately 13.8º of visual angle), with the sides of each grid unit measuring 0.3 cm.

A single stimulus set containing 144 patterns was created for use by all the subjects. The stimulus set contained three types of patterns, which will be referred to as *standards*, *repetitions*, and *transformations*. The set of standard patterns consisted of 12 examples each of patterns containing 7, 8, 9, and 10 dots, for a total of 48 patterns. The 48 repetitions were identical to the standards. The 48 transformations were rotations (and, in some cases, rotations and reflections) of each of the 48 standard patterns.

Each standard pattern was created by placing the dots in random locations in the imaginary 40 × 40 grid, under the constraint that no dot be located within 4 grid units (1.2 cm; approximately 1.4º of visual angle) of another dot, as measured between centers of dots by a Euclidean metric. This constraint was applied on a dot-by-dot basis, as follows. For each dot that was to be added, a potential location was selected at random from the set of 1,600 potential grid locations. If the location was at least 4 grid units away from all the other dots, the dot was added to the pattern; otherwise, it was discarded, and the process was repeated.

As was noted above, the repetitions were identical to the standards. The transformations were created in the following manner. Of the 12 patterns at each numerosity level, 6 were first reflected about the vertical axis and then rotated. Two of these patterns were rotated 90º, 2 were rotated 180º, and 2 were rotated 270º. The other 6 patterns at each numerosity level were rotated in the same manner without being reflected. It should be noted that each of these transformations (reflected and rotated 90º, 180º, or 270º; not reflected and rotated 90º, 180º, or 270º) would produce a different dot pattern, assuming that the pattern to be transformed is not symmetrical about the vertical, horizontal, or diagonal axis.

Each subject completed three blocks of 48 trials each. The first block of trials consisted of the 48 standard patterns. The second and third blocks consisted of the 48 repetitions and the 48 transformations. Half of the patterns appeared as repetitions in Block 2 and as transformations in Block 3, and the other half of the patterns appeared as transformations in Block 2 and as repetitions in Block 3.

A single set of patterns was used for all the subjects. The assignment of patterns to reflection and rotation conditions in creating the transformations and the order of appearance as a repetition or a transformation in Blocks 2 and 3 for each pattern were the same for all the subjects. The order in which the 48 patterns within each block appeared was determined randomly, separately for each subject.

Procedure. The subjects completed a single experimental session consisting of three blocks of 48 trials each, for a total of 144 experimental trials. The first block consisted of the 48 standard patterns, and the second and third blocks consisted of an even mixture of the 48 repetitions and the 48 transformations.

The subjects were told that they would be making subjective judgments concerning the organization of a series of dot patterns. They were told that the task had no right or wrong answers and that they should report the impression they had formed of the groups present in the pattern at the time the pattern first appeared. They were instructed on the use of the mouse, including how they were to indicate the groups, clear the pattern to start over, and advance to the next trial. The experimenter told the subjects to spend only enough time on each pattern as would be needed to enter their grouping judgments.

Each experimental trial began with a plus sign fixation point, presented at the center of the screen for 200 msec. Then the fixation point was removed, and the dot pattern was presented. Simultaneous with the onset of the dot pattern, a mouse cursor appeared near one of the four corners of the imaginary 40 × 40 grid. The four screen locations at which the mouse cursor could initially appear were located outside of the grid, at 1 cm above and 1 cm to the left of the upper left corner, 1 cm above and 1 cm to the right of the upper right corner, 1 cm below and 1 cm to the left of the lower left corner, or 1 cm below and 1 cm to the right of the lower right corner. The starting location of the mouse cursor was determined randomly for each trial, with each of the four locations having an equal probability of being selected.

The mouse cursor consisted of a plus sign that appeared in light gray (IBM 7) when the mouse was moving and no buttons were being depressed. So that it could be easily located by the subjects, it alternated between light gray and dark gray (IBM 8) every 400 msec during periods in which the mouse was stationary and neither button was being depressed.

The subjects indicated the groups they saw in the pattern by holding down the left mouse button and moving the mouse to encircle the dots that made up each group that they chose to report. Whenever the subject held down the left button, the mouse cursor was removed, and the mouse began to draw a dark gray (IBM 8) line measuring 0.2 cm wide, starting at the location of the cursor at the moment the left button was depressed. This line remained on the screen until the subject created a closed region by intersecting the line with itself. Once a closed region was created, the entire length of the line disappeared immediately, regardless of the contents of the region. If the region contained a legal group according to the selection constraint, consisting of one or more (previously un-

Table 2
Between-Subjects Agreement as Percentages for Standard Patterns in Experiment 2, by Numerosity Level

No. of Subjects	Numerosity Level				
Agreeing	7	8	9	10	Mean
1	16	20	32	24	23
2	13	7	14	10	11
3	18	7	7	10	11
4	9	4	16	0	7
5	8	8	11	17	11
6	10	3	7	3	6
7	0	8	4	0	3
8	4	4	4	4	4
9	0	10	0	0	3
10	6	0	6	0	3
11	0	12	0	12	6
12	7	0	0	0	2
13	0	7	0	0	2
14	0	0	0	8	2
15	8	0	0	0	2
16	0	9	0	0	2
17	0	0	0	0	0
18	0	0	0	0	0
19	0	0	0	0	0
20	0	0	0	11	3
All	100	100	100	100	16

grouped) white dots and no (previously grouped) gray dots, the dots within the loop changed from white to gray, thus defining a group. If the region contained only one dot or an illegal group of dots (one that included one or more than one gray dots, which would violate the selection constraint), the dots within the region remained unchanged. After a loop was made, the subject could continue to move the cursor, but the cursor did not draw the gray line until the subject released the left mouse button and depressed it again.

The subjects were able to clear any groups that they had already entered for the current trial, so that they could erase any mistake they may have made in entering their grouping judgment. To do this, they depressed both mouse buttons simultaneously, which caused any line that was being drawn to disappear and all the gray dots to revert to white. After releasing both buttons, the subject was able to make a new grouping judgment for the pattern. To finish the trial, the subjects depressed the right button alone, at which point the screen was cleared and the next trial began, following a 1,000-msec intertrial interval.

Results and Discussion

The data consisted of the grouping judgments made on each trial. Interpreting the results involved making pairwise comparisons between judgments produced by subjects and those made by the CODE algorithm for the same stimulus patterns. The results were analyzed with respect to between-subjects agreement, within-subjects agreement, and agreement with the judgments generated by the CODE algorithm.

The data generated by the CODE algorithm consisted of all of the algorithm-constrained judgments for each of the 48 standard patterns. The implementation of CODE used here was sufficiently sensitive to locate all the logically possible judgments, generating a total of 408 judgments for the 48 patterns.

In comparing subject data with the judgments generated by CODE, a subject's grouping judgment was defined as *CODE legal* if it matched one of the judgments generated by CODE for the pattern and as *CODE illegal* if it failed to match any of the CODE-generated judgments. Thus, CODE was used to sort the subjects' judgments into mutually exclusive categories.

It should be pointed out that the proportion of selection-constrained judgments that are CODE legal decreases as the number of elements in the pattern increases. For example, patterns with 7 dots have 877 selection-constrained judgments (as is indicated in Table 1), of which 7 are CODE legal, a 125:1 ratio. In contrast, patterns with 10 dots have 115,975 selection-constrained judgments, of which 10 are CODE legal, an 11,597.5:1 ratio. Although CODE generates multiple judgments for each pattern, its ability to match the judgments that subjects produce is nontrivial. With the pattern numerosities used in the present study, only a very small proportion of selection-constrained judgments are CODE legal.

Agreement between subjects. The extent to which the subjects agreed on how patterns should be grouped was assessed. Because each subject received the same set of patterns, it was possible to measure, for each of the patterns, how many subjects made identical judgments. Results concerning agreement between subjects involve comparisons among the standard patterns only, and not among the repetitions or the transformations. Table 2 presents these results, aggregated over the 12 patterns within each numerosity level and broken down by numerosity (indicated in the columns). The percentage of judgments within each numerosity level that matched *n* subjects' judgments for the same pattern is shown, with the rows representing different values of *n*. For example, the row labeled 1 indicates that for patterns with seven dots, 16% of the judgments were unique, whereas the row labeled 2 indicates that 13% of the judgments matched only 1 other subject's judgment.

These results can be summarized by calculating an *agreement index*, which is defined as the percentage of other grouping judgments in a category that match a particular judgment. When the agreement index is used to measure between-subjects agreement, it specifies the percentage of subjects in the experiment who matched a judgment, as a percentage of the number of other subjects in the experiment. For example, if a particular judgment were matched by 6 of the 29 other subjects, its agreement index would be 21% (since 6/29 = .21). The overall agreement index was 16%, indicating that, on average, a particular judgment matched 16% of the other subjects' judgments for the same pattern. The agreement index was 23%, 16%, 16%, and 11% for numerosities 7–10, respectively. The agreement index indicates that between-subjects agreement decreased with numerosity, which might be expected, because the number of potential judgments increases dramatically with numerosity. This result replicates the findings of Compton and Logan (1993).

Table 3
Percent of Judgments That Were CODE Legal (C+) and CODE Illegal (C−) Between Standards and Repetitions and Between Standards and Transformations, by Numerosity and Overall, in Experiment 1

	Standard		CODE Status			
			Repetitions		Transformations	
Numerosity Level	Code Status	%	C+	C−	C+	C−
7	C+	58	44	13	42	15
	C−	42	16	26	16	268
8	C+	45	29	16	29	16
	C−	55	13	42	20	35
9	C+	45	29	16	30	15
	C−	55	17	39	20	36
10	C+	34	19	14	20	13
	C−	66	13	54	17	49
All	C+	45	30	15	30	15
	C−	55	15	40	18	37

How should these agreement indices be interpreted? There was substantial disagreement among subjects on how to group the patterns. Overall, 23% of the judgments that subjects made were unique—that is, they were not matched by any of the 29 other subjects. On the other hand, the pattern of between-subjects agreement is far from random. For example, for patterns with 7 dots, the subjects agreed on 23% of the judgments, as compared with the 0.1140% agreement index that would be expected if the subjects simply made their judgments at random from the set of logically possible judgments (877 judgments are possible with 7 dots; see Table 1). For patterns with 10 dots, the agreement index was 11%, as compared with the 0.0009% agreement index that would be expected if logically possible judgments were selected at random.

These results are intermediate between total agreement (an agreement index of 100%) and total randomness (agreement indices ranging from 0.1140% to 0.0009%, depending on pattern numerosity; see Table 1). Given this divergence from what would be expected if the subjects selected judgments at random, it is clear that the subjects' judgments were strongly constrained by grouping processes that operated similarly across subjects.

Agreement within subjects. The proportion of repetitions and transformations that matched their corresponding standard patterns was calculated. Overall, 34% of the repetitions matched their standard, and 29% of the transformations matched their standard. The overall between-subjects agreement index was 16%, which is considerably lower than the within-subjects agreement indices of 34% and 29% for repetitions and transformations, respectively. This greater level of within-subjects agreement, relative to between-subjects agreement, indicates the presence of individual differences in the grouping judgments that the subjects made. As was seen with the between-subjects results, the within-subjects agreement indices fell between the extremes of total agreement and random selection from the set of logically possible judgments.

The within-subjects results allow the issue of invariance over rotation to be addressed. If no difference had been found in the number of repetitions versus number of transformations that matched their standards, it would have suggested that the grouping processes involved were invariant over rotation. In fact, such a difference was found. A 2 × 4 analysis of variance (ANOVA) was conducted on the number of judgments matching the standard, with pattern type (repetition and transformation) and numerosity (7–10) as factors. There was a main effect of pattern type: A greater proportion of repetitions (34%) than of transformations (29%) matched their standards [$F(1,29) = 16.84$, $MS_e = 1.51$, $p < .001$]. There was also an effect of numerosity [$F(3,87) = 14.18$, $MS_e = 2.70$, $p < .001$] and a pattern type × numerosity interaction [$F(3,87) = 8.25$, $MS_e = 2.28$, $p < .05$], resulting from a greater decline in matching to standards for the transformations as numerosity increased. This is inconsistent with CODE.

CODE legality. The extent to which the different types of judgments were CODE legal was fairly consistent: Overall, 46% of the standards, 45% of the repetitions, and 48% of the transformations were CODE legal. This level of agreement with CODE is in the same range as that found for similar numerosities by Compton and Logan (1993), using a similar data collection procedure.

A 3 × 4 ANOVA was performed on the number of judgments that were CODE legal, with pattern type (standard, repetition, and transformation) and numerosity (7–10) as factors. There was no main effect of pattern type, and no pattern type × numerosity interaction. There was an effect of numerosity: The proportion of CODE-legal judgments declined significantly as numerosity increased [$F(3,87) = 58.14$, $MS_e = 2.25$, $p < .001$], indicating that CODE was more successful in predicting the subjects' judgments at lower numerosity levels (as was found by Compton & Logan, 1993). Such a result might be expected, since as pattern numerosity increases, the number of judgments CODE generates increases linearly, whereas the number of possible judgments increases exponentially (see Table 1).

Another question concerns the persistence of CODE legality from standards to repetitions and transformations. Did patterns that were CODE legal as standards remain CODE legal when presented as repetitions or transformations, and did CODE-illegal standards also remain CODE illegal as repetitions or transformations? This question is important because it suggests a way to test for the possibility that CODE is better able to predict the organizations that subjects will see for some patterns than for others. If this were the case, one would expect that patterns would be likely to maintain their CODE legality from standards to repetitions or transformations.

Table 3 shows the proportion of judgments of each trial type that were CODE legal and CODE illegal, at each numerosity level and overall, as percentages. These proportions are agreement indices,[1] so they can be compared with the between-subjects results described above. The

Table 4
Percent of Judgments That Were Matches (=) and CODE-Related Mismatches (≠) Between Standards and Repetitions and Between Standards and Transformations, by Numerosity Level and Overall, in Experiment 1

	Pattern Type			
Numerosity	Repetitions		Transformations	
Level	=	≠	=	≠
7	39	17	41	14
8	35	11	31	9
9	32	9	27	13
10	31	8	17	13
All	34	11	29	12

results for the standards are shown, and those for repetitions and transformations are broken down by the CODE legality of their standards. Overall, 70% of the repetitions maintained the CODE legality of their standards (30% were CODE legal both as standards and as repetitions, and 40% were CODE illegal both as standards and repetitions), and 67% of the transformations maintained the CODE legality of their standards (30% CODE legal as standards and transformations; 37% CODE illegal as standards and transformations).

A series of chi-square tests was performed to determine whether the proportions of judgments that maintained their CODE legality from standard to repetition and from standard to transformation were significantly greater than would be expected by chance, given the overall proportion of CODE-legal judgments at each trial type. Two 2 × 2 chi-square tests were performed separately for each subject (Hintzman, 1980), which compared standards to repetitions and to transformations. The tests compared the number of judgments at each combination of standard CODE legality × repetition or transformation (depending on the test) CODE legality. (The values on which the chi-square tests were based are shown, averaged over subjects, in the two bottom rows of Table 3, with the comparison of standards with repetitions in the leftmost columns and the comparison of standards with transformations in the rightmost columns).

For the comparison of standards with repetitions, $\chi^2(1)$ ranged from 0.42 to 21.00, with a mean of 7.07, and was greater than the $p = .05$ criterion of 3.84 for 22 of the 30 subjects. For the comparison of standards with transformations, $\chi^2(1)$ ranged from 0.06 to 17.06, with a mean of 4.93, and was greater than the $p = .05$ criterion for 15 of the 30 subjects. Patterns tended to maintain their CODE legality from standards to repetitions and to transformations, indicating that the ability of CODE to predict subjects' judgments varies with pattern identity.

CODE-related mismatches. Thus far, the approach to determining the level of agreement has employed sets of pairwise comparisons between patterns, to see whether they match or mismatch: A binary distinction is made in which judgments are defined as either identical or completely different. The CODE algorithm allows a new pairwise relation to be defined, in which two patterns that mismatch but are both CODE legal are viewed as being related. In terms of the processing assumptions of CODE, these patterns have the same data-driven component (the CODE surface) but differ in the conceptually driven component (the threshold setting). From this perspective, reliability can be defined as including judgments that mismatch but are related by virtue of being CODE legal, as well as including judgments that match.

Finding the CODE-related mismatches involves determining what proportion of mismatching repetitions and transformations were CODE legal and also had CODE legal standards. Table 4 shows the CODE legality of repetitions and transformations that mismatched their standards. When the standard was CODE legal and mismatched the repetition (26% of the time), 42% of the mismatching repetitions were also CODE legal and, therefore, CODE related. Eleven percent of all repetitions (.26 × .42 = .11) mismatched the standard but were CODE related (27 of the 30 subjects produced at least one mismatching CODE-related repetition). When CODE-related mismatches, as well as matches, are included, the proportion of repetitions that are defined as being related to the standard increases from 34% to 45%.

The same analysis was applied to transformations. When the standard was CODE legal and mismatched the transformation (30% of the time), 44% of the mismatching transformations were also CODE legal, so that 13% of the transformations (.30 × .44 = .13) mismatched but were CODE related (28 out of the 30 subjects produced at least one mismatching CODE-related transformation). When the mismatching CODE-related judgments are included, the proportion of transformations defined as being related increases from 29% to 41%.

For both repetitions and transformations, CODE was better at predicting the subjects' judgments when the CODE-related mismatches, as well as the matches, are counted as successes. By counting CODE-related mismatches as successes, the top-down component of CODE, the threshold, is treated as a free parameter.

CODE and invariance over rotation. Experiment 1 provided a test of the assumption, made by CODE, that grouping judgments should be invariant over changes in reflection and rotation. As is shown at the bottom of Table 4, when all the patterns are considered, 34% of the repetitions matched their standard, but only 29% of the transformations matched their standard. This suggests that the grouping judgments were, in fact, sensitive to the changes in reflection and rotation and that CODE is incorrect to assume that these transformations do not affect grouping judgments.

What might account for this difference? One possibility is that, on some patterns, other grouping principles besides proximity were in effect and that these grouping principles are sensitive to changes in orientation. To assess this possibility, we compared the ability of repetitions and transformations to match their standard, but excluded patterns for which the standard was not CODE legal. When the analysis was restricted in this way, the

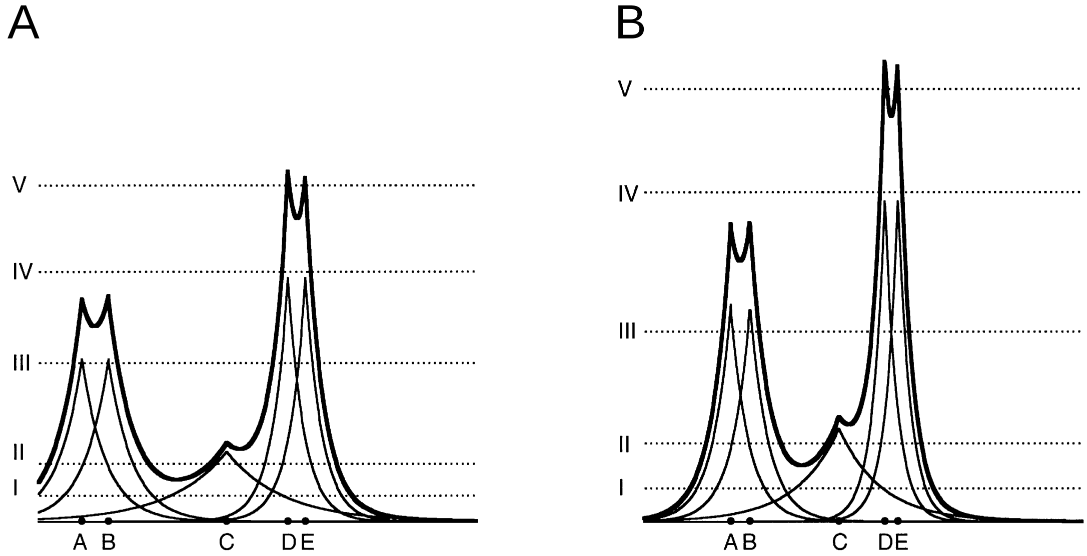

Figure 5. The CODE algorithm applied to (A) the one-dimensional pattern seen in Figure 2 and (B) the same pattern reduced in scale by 25%. Thresholds defining different judgments are indicated by Roman numerals.

transformation had little effect: 43% of the repetitions and 42% of the transformations matched their standards. Consequently, it appears that most of the cost of reflecting and rotating the patterns was due to patterns that CODE was not able to account for.

EXPERIMENT 2

The goal of Experiment 2 was to examine the effects of a different type of transformation: changes in scale. This was investigated by using a method analogous to that used in Experiment 1. Again, 48 standards were presented in the first block. Half of these standards (the *full-scale* standards) were created according to the same procedure used to create the standards in Experiment 1. The other half of the standards (the *reduced-scale* standards) were created in a similar fashion and then reduced in scale by a factor of 25%. Again, the repetitions were identical to the standards. For the transformations, full-scale patterns were presented as reduced-scale patterns, and vice-versa.

The 25% reduction in scale was chosen as a compromise between two competing goals. One goal was to have a relatively large difference in scale, so that any influence of the transformation on the subjects' judgments could be detected. A competing goal was to limit the differences in scale, to avoid drawing too much attention to the transformation (note that, for the reduced-scale patterns, the minimum interdot distance was reduced by 25%).

According to CODE, only relative distance matters in grouping judgments. Consequently, grouping should be independent of scale, and any effect of scale on subjects' grouping judgments would falsify CODE. Experiment 2 provided the opportunity to test for systematic influences of scale on subjects' grouping judgments. Specifically, one possibility was that subjects might not fully adjust their grouping judgments to compensate for the transformation in scale. This can be thought of as being analogous to inertia of the CODE threshold, so that decreases in scale would correspond to CODE thresholds that are relatively lower (producing fewer and larger groups), and increases in scale would correspond to CODE thresholds that are higher (producing more and smaller groups). Experiment 2 provided the opportunity to test for this possibility.

It should be noted that this discussion of CODE thresholds as being relatively higher or lower refers to different judgments generated by CODE for an individual pattern and does not imply any comparison of CODE thresholds between different patterns (including those that differ in scale). This point is demonstrated by Figure 5, which shows (panel A) the one-dimensional dot pattern that was presented in Figure 2 and (panel B) the same pattern reduced in scale by 25%. These two panels are plotted on the same *y*-axis (which corresponds to the height of the strength gradient). It is the discrete, categorical differences in the threshold height (e.g., moving from Threshold II up to Threshold III, or down to Threshold I), and not the real-number strength gradient values, that will be considered here.

Method

Subjects. The subjects were 30 introductory psychology students at the University of Illinois, who received course credit for their participation.

Apparatus and Stimuli. The stimuli were presented on IBM PS/2 computers, and the subjects entered their responses with a two-button IBM mouse. The stimulus set contained 48 *full-scale* patterns, which were generated by the same procedures that were used to create the 48 standards in Experiment 1. (Half of these full-scale patterns were ultimately designated as standards, and half

Table 5
Between-Subjects Agreement for Standard Patterns in Experiment 2, by Scale and Numerosity Level, as Percentages

No. of Subjects	Numerosity Level for Full-Scale Patterns				Numerosity Level for Reduced-Scale Patterns				
Agreeing	7	8	9	10	7	8	9	10	Mean
1	16	20	32	24	10	18	17	32	21
2	13	7	14	10	6	8	10	12	10
3	18	7	7	10	10	7	7	20	11
4	9	4	16	0	0	9	2	11	6
5	8	8	11	17	3	8	3	3	8
6	10	3	7	3	7	3	3	0	5
7	0	8	4	0	4	12	4	4	4
8	4	4	4	4	4	0	0	0	3
9	0	10	0	0	10	10	0	5	4
10	6	0	6	0	6	6	0	6	3
11	0	12	0	12	12	12	0	0	6
12	7	0	0	0	0	0	0	0	1
13	0	7	0	0	0	7	0	0	2
14	0	0	0	8	0	0	8	8	3
15	8	0	0	0	17	0	0	0	3
16	0	9	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	11	0	1
20	0	0	0	11	0	0	11	0	3
21	0	0	0	0	0	0	0	0	0
22	0	0	0	0	12	0	24	0	5
All	100	100	100	100	100	100	100	100	20

were designated as transformations.) The full-scale patterns were then reduced in size by a factor of 25% to create a set of 48 *reduced-scale* patterns. This reduction involved decreasing the distance between each dot and the center of the imaginary grid (grid position 20, 20) by a factor of one fourth.

For both full-scale and reduced-scale patterns, the size of the dots was the same as that in Experiment 1. For full-scale patterns, the size of the 40 × 40 imaginary grid and the minimum nearest-neighbor distance between dots were also the same. For reduced-scale patterns, the imaginary grid measured 9 cm on a side (approximately 10.3º of visual angle), and the minimum nearest-neighbor distance was 0.9 cm (approximately 1º of visual angle).

The standard set of patterns that was presented in Block 1 contained 24 full-scale standards and 24 reduced-scale standards. At each numerosity level, 6 of the 12 standards were full scale, and 6 were reduced scale. The 48 transformations appeared at the scale opposite to that of their standards: Full-scale standards became reduced-scale transformations, and vice-versa.

Procedure. The procedure was identical to that of Experiment 1.

Results and Discussion

As in Experiment 1, between-subjects and within-subjects agreement was assessed. In addition, the possibility that the direction of the change in scale from standard to transformations had a systematic effect on the nature of mismatching transformations or on CODE thresholds was investigated.

Agreement between subjects. Table 5 shows between-subjects agreement separately for full-scale and reduced-scale standard patterns and is similar in format to Table 2. The overall agreement index was 20%, which was slightly lower than the overall agreement index of 24% seen in Experiment 1. The agreement index was 22%, 19%, 23%, and 15% for numerosities 7–10, respectively. The results were again intermediate between total agreement and what would be expected if the subjects made randomly selected judgments.

Agreement within subjects. Overall, 40% of the repetitions matched their standards, and 35% of the transformations matched their standards. This result indicates greater within-subjects agreement than was seen in Experiment 1. Again, the overall between-subjects agreement index, which was 20%, was considerably lower than the within-subjects agreement indices for repetitions and transformations, indicating the presence of individual differences in subjects' grouping judgments. The within-subjects agreement indices fell between the extremes of total agreement and random selection from the set of logically possible judgments.

As in Experiment 1, the repetitions matched their standards significantly more often than did the transformations. A 2 × 4 ANOVA was conducted on the number of judgments matching the standard, with pattern type (repetition and transformation) and numerosity (7–10) as factors. The effect of pattern type was significant [$F(1,29) = 19.03$, $MS_e = 1.37$, $p < .001$], as was that of numerosity [$F(3,87) = 3.45$, $MS_e = 3.72$, $p < .05$]. There was no pattern type × numerosity interaction.

CODE legality. The extent to which the different types of judgments were CODE legal was somewhat lower than in Experiment 1: 57% of the standards, 52% of the repetitions, and 51% of the transformations were CODE legal. A 3 × 4 ANOVA was performed on the number of judgments that were CODE legal, with pat-

Table 6
Percent of Judgments That Were CODE Legal (C+) and CODE Illegal (C−) Between Standards and Repetitions and Between Standards and Transformations, by Numerosity and Overall, in Experiment 2

	Standard		CODE Status			
			Repetitions		Transformations	
Numerosity Level	Code Status	%	C+	C−	C+	C−
7	C+	65	47	19	49	16
	C−	35	12	22	15	20
8	C+	55	41	14	38	17
	C−	45	13	32	14	31
9	C+	57	40	17	37	19
	C−	43	10	34	9	35
10	C+	49	35	14	33	16
	C−	51	11	40	11	40
All	C+	57	41	16	39	17
	C−	43	11	32	12	32

tern type (standard, repetition, and transformation) and numerosity (7–10) as factors. There was a significant effect of pattern type [F(2,58) = 3.97, MS_e = 3.10, $p < .05$] and numerosity [F(2,58) = 21.00, MS_e = 2.94, $p < .001$], but no pattern type × numerosity interaction. The effect of pattern type was largely due to the greater level of CODE legality for standards, as was shown by a second ANOVA with repetitions and transformations only as pattern types: Pattern type was not significant. However, there was an effect of numerosity [F(3,87) = 16.05, MS_e = 2.82, $p < .001$], which was slightly larger for transformations, as is indicated by a pattern type × numerosity interaction [F(87,3) = 3.16, MS_e = 1.13, $p < .05$].

Table 6 follows the format of Table 3 in showing the proportion of judgments of each trial type that were CODE legal and CODE illegal, at each numerosity level and overall, as percentages. Overall, 73% of the repetitions and 71% of the transformations maintained the CODE legality of their standards. A series of chi-square tests was performed to determine whether the proportion of judgments that maintained their CODE legality from standard to repetitions and to transformations was significantly greater than would be expected by chance, given the overall proportion of CODE-legal judgments at each trial type. Two 2 × 2 chi-square tests were performed, separately for each subject, that compared the number of judgments at each combination of standard CODE legality × repetition or transformation CODE legality. For the comparison of standards with repetitions, $\chi^2(1)$ ranged from 0.05 to 17.57, with a mean of 9.16, and was greater than the $p = .05$ criterion of 3.84 for 25 of the 30 subjects. For the comparison of standards with transformations, $\chi^2(1)$ ranged from 0.62 to 31.11, with a mean of 8.06, and was greater than the $p = .05$ criterion for 23 of the 30 subjects. As in Experiment 1, patterns tended to maintain their CODE legality from standards to repetitions and transformations, indicating that the ability of CODE to predict subjects' judgments varies with pattern identity.

CODE-related mismatches. Fourteen percent of the repetitions mismatched the standard but were CODE related (27 of the 30 subjects produced at least one mismatching CODE-related repetition). When CODE-related mismatches, as well as matches, are included, the proportion of repetitions that are defined as being related to the standard increases from 40% to 54%. Sixteen percent of the transformations mismatched the standard but were CODE related (27 of the 30 subjects produced at least one mismatching CODE-related transformation). When the mismatching CODE-related judgments are included, the proportion of repetitions defined as being related to the standard increases from 35% to 51% (see Table 7 for complete results.)

CODE legality and scale. One prediction that follows from the notion that the grouping processes underlying subjects' judgments are invariant over changes in scale is that the level of CODE legality should be the same for full-scale and reduced-scale patterns. Table 8 shows the CODE legality of full-scale and reduced-scale patterns, aggregated over the three pattern types. A 2 × 4 ANOVA was conducted, with scale (full or reduced) and numerosity as factors, on the number of judgments that were CODE legal, aggregated over standards, repetitions, and transformations. There was no significant effect of scale, but there was an effect of numerosity, as was reported previously, and also a scale × numerosity interaction [F(3,87) = 3.82, MS_e = 2.54, $p < .05$] that showed no systematic pattern.

CODE and invariance over changes in scale. As is shown at the bottom of Table 7, when all patterns are considered, 40% of the repetitions matched their standard, but only 35% of the transformations matched their standard. When only patterns for which the standard was CODE legal are considered, 48% of the repetitions and 41% of the transformations matched the standard. This result contrasts with that of Experiment 1, which showed only a 1% difference between repetitions and transformations when only patterns with CODE-legal standards were considered.

Relation of changes in scale to CODE thresholds. The scale transformations used in Experiment 2, unlike the rotations and reflections in Experiment 1, altered the

Table 7
Percent of Judgments That Were Matches (=) and CODE-Related Mismatches (≠) Between Standards and Repetitions and Between Standards and Transformations, by Numerosity Level and Overall, in Experiment 2

	Pattern Type			
	Repetitions		Transformations	
Numerosity Level	=	≠	=	≠
7	44	15	38	20
8	43	14	39	14
9	40	13	33	16
10	35	14	30	14
All	40	14	35	16

Table 8
Percent of Full-Scale and Reduced-Scale Judgments That Were CODE Legal (C+) and CODE Illegal (C−), Aggregated Over Pattern Type, by Numerosity Level and Overall, in Experiment 2

Numerosity Level	Scale			
	Full		Reduced	
	C+	C−	C+	C −
7	62	38	63	37
8	51	49	56	44
9	47	53	55	45
10	48	53	45	55
All	52	48	55	45

absolute interdot distances. This allows an investigation of the nature of any effect that changes in scale might have on the CODE thresholds (for judgments that are CODE legal). If the subjects' judgments showed some inertia, so that they did not completely compensate for changes in scale, CODE thresholds should be lower (producing fewer and larger groups) when the transformation is to reduce the pattern in scale and higher (creating more, smaller groups) when the transformation is to increase the pattern in scale.

An analysis was conducted to investigate this possibility. The analysis involved only those judgments that were CODE legal both as standards and as transformations, but at different CODE thresholds. Of the patterns for which the transformation was a *reduction* in scale, 34% were CODE legal both as standards and as transformations. Of this 34%, 47% (16% of the total) involved a change in CODE threshold from standard to transformation. Of the patterns for which the transformation was an *increase* in scale, 45% were CODE legal both as standards and as transformations, and of these, 36% (16% of the total) involved a change in CODE threshold from standard to transformation.

The results showed that the subjects' judgments were indeed influenced by changes in scale in a systematic way. When the transformation was a reduction in scale, 87% of the changes in CODE thresholds involved a decrease, toward fewer and larger groups. In contrast, when the transformation was an increase in scale, 62% of the changes in CODE thresholds involved an increase, toward more and smaller groups. Twenty-three subjects showed this pattern when the transformation was a decrease in scale (with 3 showing the opposite pattern), and 15 subjects showed this pattern when the transformation was an increase in scale (with 6 showing the opposite pattern). A 2 × 2 chi-square test was performed on the number of patterns, aggregated over subjects, that fell into each change in scale (increase or decrease) × change in CODE threshold (lower or higher) category. This test was significant [$\chi^2(1) = 60.59$, p =.01], confirming that the influence of scale on the subjects' judgments was systematically related to changes in CODE thresholds.

EXPERIMENT 3

Experiments 1 and 2 yielded a greater degree of match for repetitions, relative to transformations. One possible explanation for this result is that it is a memory effect. Perhaps, when subjects recognize a repetition or a transformation as being the same as or similar to one they have already seen and grouped (i.e., the related standard), they tend to group it in the same way. To the extent that the subjects were influenced by their prior experience with the patterns (as standards) when they made grouping judgments for repetitions, the reliability of grouping judgments between standards and repetitions might be overestimated.

To determine what effect, if any, recognition of the standards might have had on the grouping judgments for repetitions, a third experiment was conducted. The first block of Experiment 3 was similar to that of Experiment 2: The subjects made grouping judgments for full-scale and reduced-scale standards. In the second block, a surprise recognition memory test was announced: The subjects were to discriminate patterns they had grouped from patterns they had not seen before. Using this approach, it was possible to obtain an estimate of the possible influence of memory for standards on the grouping judgments for repetitions.

Method

Subjects. The subjects were 30 introductory psychology students at the University of Illinois, who received course credit for their participation.

Apparatus and Stimuli. The apparatus was the same as that used for Experiment 2, with the addition of the use of the computer keyboard for entering recognition judgments. As before, a single stimulus set was generated for all the subjects. The stimulus set contained 48 standard patterns, 16 filler patterns, and 48 new patterns. The 48 standard patterns consisted of 24 full-scale and 24 reduced-scale patterns and were generated by the same procedures that had been used to create the standards in Experiment 2. For use in the recognition test, 48 new patterns were created, 6 at each combination of scale (full or reduced) × numerosity (7, 8, 9, or 10 dots). Finally, 16 filler patterns were created, with 2 patterns at each scale × numerosity combination.

Procedure. In the first block, the subjects made grouping judgments for the 48 standard patterns and then for the 16 filler patterns. The second block began with a screen that read "Return to experimenter for further instructions before continuing." At this point, the experimenter informed the subjects that they would be presented with a series of dot patterns and that, rather than making grouping judgments, they were to indicate whether or not each pattern was one they had previously grouped in the first block. For half of the subjects, if the pattern was one they had seen before, they were to press the "z" key, and if the pattern was one they had not seen before, they were to press the "/" key, both of which are located on the bottom row of the computer keyboard. For the other half of the subjects, this key assignment was reversed.

In the second block, the subjects made recognition judgments for half of the standards and half of the new patterns (three patterns at each scale × numerosity combination, for both the standard and the new patterns). In the third block, the subjects made recognition judgments for the remaining standard and new patterns.

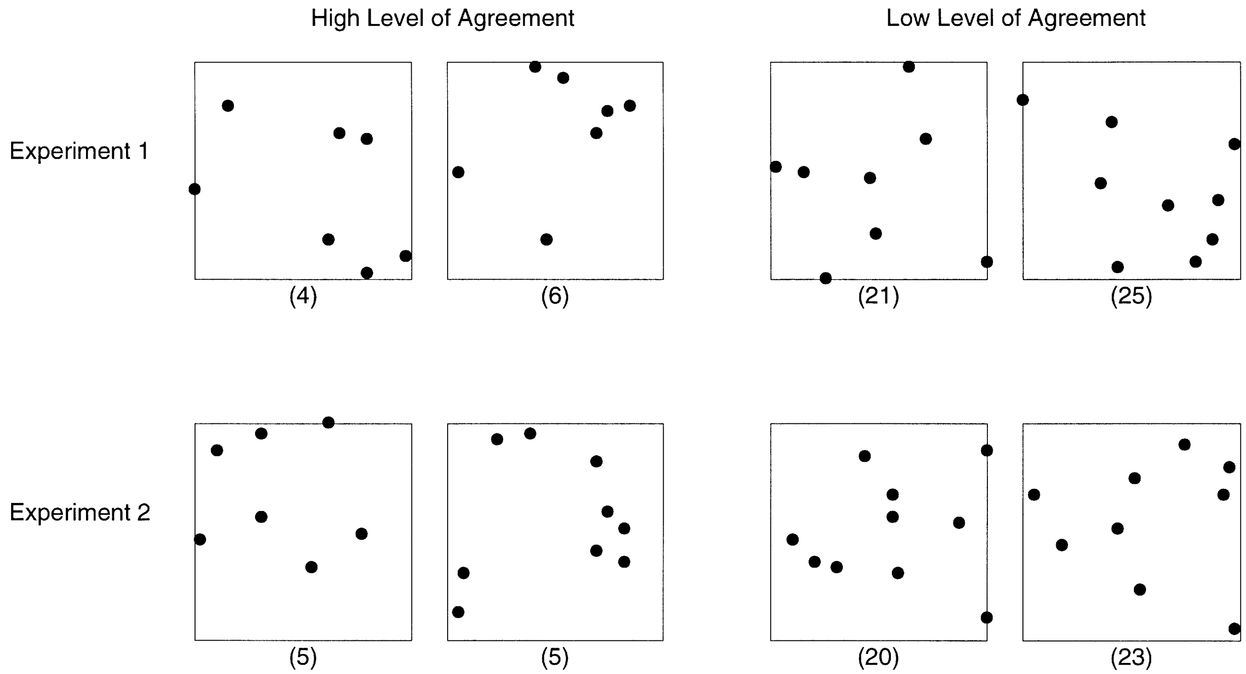

Figure 6. For Experiments 1 and 2, the two stimuli that had the highest levels of intersubject agreement and the two stimuli that had the lowest levels of intersubject agreement are shown. The number of different ways in which the 30 subjects grouped each pattern is shown in parentheses.

Results and Discussion

The results showed poor recognition for the patterns that the subjects had grouped. The mean hit rate was .43, .48, .53, and .42 for numerosities 7–10, respectively. The corresponding mean false alarm rates were .34, .31, .36, and .32. Collapsed across numerosity, d' for individual subjects ranged from 0.16 to 1.08, with a mean of 0.49. Because d' was significantly above zero [$t(2) = 6.53$, $p < .001$], it indicates that there was some recognition for standards. However, recognition was quite limited, which is perhaps not surprising, given the number of elements in each pattern and the random basis of their construction.

These results suggest that the grouping of repetitions was not based entirely on memory for the way in which the pattern was grouped during its prior presentation as a standard. Instead, it appears that the grouping judgments for repetitions relied, to some extent, on the same grouping processes that underlay the grouping of the standards. However, some caution should be taken in interpreting these results, because it is possible that the ability of subjects to retrieve a grouping judgment that was made on a previous encounter with a pattern and then to use it as the basis for their response is not perfectly indexed by the recognition memory test we used.

GENERAL DISCUSSION

The purpose of this article was to assess the reliability of subjects' grouping judgments of random dot patterns, both within and between subjects. The assessment of reliability over reflection, rotation, and changes in scale allows a test of the CODE algorithm's assumption that grouping judgments should be invariant over such transformations.

The reliability of comparable judgments was much greater within than between subjects, even when transformations within subjects were compared with standards between subjects. This difference indicates the presence of individual differences in the way in which subjects grouped patterns. The patterns used were random, and as a result, nearest-neighbor distances tended to be relatively homogeneous, leading to patterns whose organizations were low in goodness (i.e., ambiguous, or lacking in "inner coherence"; Wertheimer, 1923/1967, p. 83). Judgments for patterns with relatively more heterogeneous nearest-neighbor distances would be expected to lead to more reliable grouping judgments, both within subjects and between subjects.

Figure 6 shows, for Experiments 1 and 2, the two standard patterns that had the lowest level of intersubject agreement and the two standard patterns that had the highest level of intersubject agreement. In this case, agreement is defined in terms of the number of different judgments that the subjects made. For example, the 30 subjects in Experiment 1 grouped one of the patterns shown in Figure 6 in only four different ways (upper left). In contrast, another pattern from Experiment 1 was grouped in 25 different ways (upper right).

The CODE algorithm was able to successfully predict the subjects' judgments slightly less than half of the time

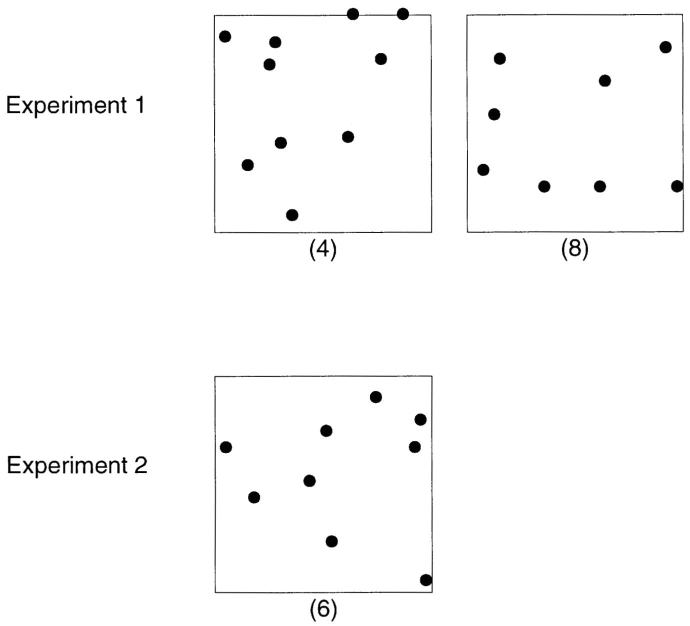

Figure 7. The three patterns (two from Experiment 1, one from Experiment 2) for which there were fewer than 10 CODE-legal judgments, with the number of CODE-legal judgments shown in parentheses.

in Experiment 1 and slightly more than half the time in Experiment 2. Patterns that were CODE legal as standards were likely to also be CODE legal as repetitions and transformations, indicating that CODE was more successful at predicting judgments for some patterns than for others.

For which patterns was CODE most successful? The ability of CODE to predict judgments declined with numerosity, as did between- and within-subjects agreement. Because the stimulus patterns were random, goodness should decrease with pattern numerosity, because adding more elements increases the number of different organizations that are likely to be seen (see Garner, 1970). To the extent that this occurs, it would be expected that both the performance of CODE and between- and within-subjects agreement would decrease with numerosity. The standard textbook demonstrations of grouping by proximity are designed to be unambiguous to viewers, and for these types of patterns, CODE is able to match subjects' judgments quite successfully (van Oeffelen & Vos, 1983). However, when intersubject agreement was very high, subjects still produced some CODE-illegal judgments.

Across Experiments 1 and 2, CODE matched at least 10 of the subjects' judgments for all but three patterns. The three patterns (two from Experiment 1, one from Experiment 2) for which CODE matched fewer than 10 subjects are shown in Figure 7.

Multiple-Threshold Extension of CODE

It is possible that as pattern goodness decreases, the single-threshold version of CODE becomes less appropriate. Would CODE be more successful if different thresholds could be applied to different regions of a pattern? To assess this possibility, a multiple-threshold version of CODE was created. According to the multiple-threshold version of CODE, a subject's judgment is defined as being matched if every group of two or more dots that it contains can be found at one or more CODE thresholds. (Note that if all the groups within a subject's judgment can be found at a single CODE threshold, the judgment would be considered CODE legal by the single-threshold version of CODE that was used throughout the present article.)

For Experiments 1 and 2, the multiple-threshold version of CODE was applied to all the subject judgments that were not matched by the single-threshold version of CODE. Table 9 shows the percent of judgments that were matched by the multiple-threshold version of CODE, but

Table 9
Percent of CODE-Illegal Judgments That Were Matched by the Multiple-Threshold Version of CODE, but not by the Single-Threshold Version, in Experiments 1 and 2

	Trial Type		
Numerosity	Standard	Repetitions	Transformations
	Experiment 1		
7	11	9	8
8	15	17	8
9	13	20	18
10	29	27	22
	Experiment 2		
7	9	8	6
8	10	9	8
9	12	10	11
10	9	8	8

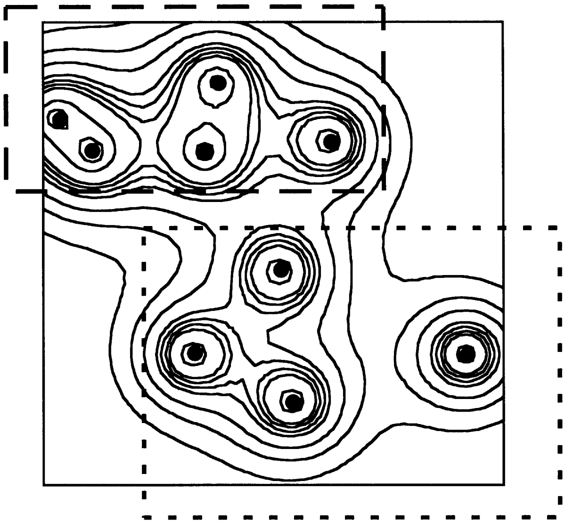

Figure 8. The most popular CODE-illegal judgment from Experiments 1 and 2. The nine dots indicate the dot locations, and the relevant CODE thresholds are indicated by contour lines. The judgment, which was made by 20 subjects, is indicated by two dotted-line boxes.

not by the single-threshold version, by pattern type and numerosity. For each experiment, the multiple-threshold version of CODE was able to handle more judgments than was the single-threshold version. However, the vast majority of CODE-illegal patterns could not be matched by the multiple-threshold version of CODE. This finding indicates that a large proportion of the judgments that the subjects produced contained groups that violated the hierarchical nesting of CODE (Palmer, 1977). It should be noted that the CODE-illegal judgments that could not even be matched by the multiple-threshold version of CODE contained violations of CODE at the level of individual groups, and not just at the level of the pattern as a whole.

An example of such a pattern is shown in Figure 8, which presents the pattern for which a CODE-illegal judgment was made by the greatest number of subjects, across Experiments 1 and 2. The relevant CODE thresholds are indicated by contour lines. The judgment was made by 20 of the 30 subjects and is indicated by two dotted-line boxes. The lower box indicates the group that violates even the multiple-threshold version CODE. The rightmost dot in the pattern is grouped with the other dots only at the lowest threshold, which specifies a single group containing all of the dots. Consequently, CODE will never group the rightmost dot with some but not all of the other dots, as did the 20 subjects who produced the judgment shown in the figure.

Reliability of Repetitions and Transformations

A difference was found between repetitions and transformations in the proportion of judgments that matched their standards. In addition, in Experiment 2, the direction of the change in scale influenced the subjects' judgments in a way that corresponded to systematic changes in CODE thresholds. These results appear to contradict a basic design assumption of the CODE algorithm, which characterizes grouping by proximity as being based solely on the relative distances among elements in the pattern (with the result that CODE is insensitive to transformations such as rotation and changes in scale). However, there are several alternative explanations for this finding.

First, CODE describes space in terms of relative interdot distances, and as a result, neither the orientation nor the scale of the pattern can have any effect on the organization it produces. However, it is possible that a different metric of space could more effectively describe the input to CODE. For example, the metric of space could be based on the absolute density of elements at particular regions, and not just on their density relative to the densities of elements at other regions. Krumhansl (1978) proposed a similar approach to characterizing the density of multidimensional similarity space in accounting for similarity data. If the metric of space depended on absolute element density, grouping by proximity would be sensitive to changes in scale (as are numerosity judgments; see Krueger, 1972). Similarly, perhaps the metric of space depends on orientation, so that units in the horizontal versus the vertical dimension are not in 1:1 correspondence. If that were the case, grouping by proximity would be sensitive to rotation. CODE could potentially be modified to use a different metric of space. This could be accomplished by transforming the *x*, *y* coordinates prior to the creation of the CODE surface.

Second, subjects may be paying less attention to elements in certain locations of the display (e.g., the top or the bottom) than to elements in other locations. If this were true, it would predict that transformations, such as rotation or changes in scale, that serve to alter the extent to which subjects attend to different parts of the pattern should reduce the level of agreement for transformations, relative to repetitions.

A third possibility is that grouping by proximity is invariant over rotation but that other grouping principles are in effect that are sensitive to orientation. For example, it may be that some of the dots are grouped by the principle of good continuation, rather than by proximity (see Figure 8), and that good continuation is more sensitive to lines in some orientations than in others (Prytulak, 1974). To the extent that grouping principles, other than grouping by proximity, are in effect that are sensitive to orientation, the degree of agreement between transformations and standards should be less than that between repetitions and standards. Wertheimer (1923/1967) argued that different grouping principles could work cooperatively or against each other and that the extent to which they apply individually and interact with each other can vary continuously with continuous changes in the stimulus. Jackendoff (1983) detailed a related notion of the application of and interaction among different grouping principles. He described "grouping preference rules" that "establish not inflexible decisions about structure, but relative preferences among a number of logically possible analyses" (p. 132). Clearly, it is difficult to determine which grouping principles are in effect for a given pattern.

The results of Experiment 3 suggest that a large part of the difference between repetitions and transformations in their degree of match to standards was due to a sensitivity of grouping processes to rotation and changes in scale, and not just to a greater level of memorability for repetitions.

In sum, a small but significant difference was found in the ability of repetitions versus transformations to match their standards, indicating that grouping judgments were not completely invariant over transformation. The CODE algorithm was successful in matching a large proportion of subjects' judgments. CODE can be used to define relations between patterns other than identity (CODE relatedness) that allow a greater proportion of repetitions and transformations to be seen as matching their standards. CODE can be used to confirm or disconfirm hypothesized organizations of stimuli that are to be used in experiments and to indicate whether different organizations of a given pattern are CODE related. Finally, the approach used here can be extended to the investigation of a range of grouping principles and to the possible interactions among them.

REFERENCES

Compton, B. J., & Logan, G. D. (1993). Evaluating a computational model of perceptual grouping by proximity. *Perception & Psychophysics*, **53**, 403-421.

Garner, W. R. (1970). Good patterns have few alternatives. *American Scientist*, **58**, 34-42.

Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, **87**, 398-410.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.

Krueger, L. E. (1972). Perceived numerosity. *Perception & Psychophysics*, **11**, 5-9.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, **83**, 445-463.

Kubovy, M. (1994). The perceptual organization of dot lattices. *Psychonomic Bulletin & Review*, **1**, 182-190.

Logan, G. D. (1996). The CODE theory of visual attention: A theoretical integration of space-based and object-based attention. *Psychological Review*, **103**, 603-649.

Logan, G. D., & Bundesen, C. (1996). Spatial effects in the partial report paradigm: A challenge for theories of visual spatial attention. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 35, pp. 243-282). San Diego: Academic Press.

Palmer, S. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, **9**, 441-474.

Prytulak, L. S. (1974). Good continuation revisited. *Journal of Experimental Psychology*, **102**, 773-777.

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

van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory & Cognition*, **10**, 396-404.

van Oeffelen, M. P., & Vos, P. G. (1983). An algorithm for pattern description on the level of relative proximity. *Pattern Recognition*, **16**, 341-348.

Wertheimer, M. (1967). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 71-88). New York: Humanities Press. (Original work published 1923)

NOTE

[1] These proportions meet the definition of agreement indices that was given in the section discussing between-subjects agreement for Experiment 1, because they are the percent of grouping judgments in a category (in this case, repetitions or transformations) that match a particular judgment (in this case, the standard).

(Manuscript received December 20, 1994;
revision accepted for publication July 21, 1998.)