Cognition 209 (2021) 104515

Contents lists available at ScienceDirect

Cognition

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

Parts beget parts: Bootstrapping hierarchical object representations through visual statistical learning

Alan L.F. Lee [a,b,*], Zili Liu [c], Hongjing Lu [c,d]

[a] Department of Applied Psychology, Lingnan University, Hong Kong
[b] Wofoo Joseph Lee Consulting and Counselling Psychology Research Centre, Lingnan University, Hong Kong
[c] Department of Psychology, University of California, Los Angeles, United States of America
[d] Department of Statistics, University of California, Los Angeles, United States of America

ARTICLE INFO

Keywords:
Object representation
Compositionality
Hierarchical structure
Statistical learning
Visual chunks

ABSTRACT

Previous research has shown that humans are able to acquire statistical regularities among shape parts that form various spatial configurations, via exposure to these configurations without any task or feedback. The present study extends this approach of visual statistical learning to examine whether prior knowledge of parts, acquired in a separate learning context, facilitates acquisition of multi-layer hierarchical representations of objects. After participants had learned to encode a shape-pair as a chunk into memory, they viewed cluttered scenes containing multiple shape chunks. One of the larger configurations was constructed by combining the learned shape-pair with an unfamiliar, complementary shape-pair. Although the complementary shape-pair had never been presented separately during learning, it was remembered better than other shape pairs that were parts of larger configurations. The greater perceived familiarity of the complementary shape-pair depended on the encoding strength of the previously learned shape-pair. This "parts-beget-parts" effect suggests that statistical learning, in combination with prior knowledge, can represent objects as a coherent whole and also as a spatial configuration of parts by bootstrapping multi-layer hierarchical structures.

1. Introduction

Visual perception has long been characterized as "unconscious inference" (Helmholtz, 1866/1924). Such inference allows humans to see objects not simply as memorized patterns of unstructured wholes, but rather as whole objects constructed from parts (Biederman, 1987; Lowe, 1985). A longstanding question is how humans use visual inputs to learn object representations based on parts. Grouping cues (e.g., curvature, boundary strength, junctions) play important roles in supporting part-based representations (Singh & Hoffman, 2001). Prior knowledge about the functional roles that parts play in objects also contributes to the formation of part-based representations (Tversky & Hemenway, 1984).

It remains unclear, however, whether humans can learn part-based representations from visual experience even in the absence of such low-level cues or prior knowledge of functional roles. Previous research suggests that, based on mere exposure to the sensory environment, humans are able to exploit statistical regularities to acquire the building blocks of a hierarchy to represent objects (Fiser & Aslin, 2001; Saffran, Aslin, & Newport, 1996). In the context of learning the representation of a visual object, the relevant statistical regularities correspond to joint probabilities of different visual features that exhibit a spatial relationship. Through visual statistical learning, humans can group features that co-occur consistently to form representations of "visual chunks", the building blocks of latent variables in a layer of hierarchical structures (Fiser & Aslin, 2001, 2002, 2005; Miller, 1956; Newport & Aslin, 2004; Turk-Browne, Junge, & Scholl, 2005). Orban, Fiser, Aslin, and Lengyel (2008) showed that human statistical learning of visual chunks can achieve close-to-optimal performance as predicted by a probabilistic chunking model, which uses the visual inputs and a generic prior favoring simple structure for a hierarchy (e.g., smaller number of latent variables) to infer chunks as representation units for recurring feature combinations.

Probabilistic chunking provides a general computational framework to capture inferences about the hierarchical structure of object representations based on visual experience. The goal of probabilistic chunking is to identify the hierarchy with latent variables (i.e., chunks) that best explain all visual inputs from learning experience. As the

* Corresponding author at: Department of Applied Psychology, Lingnan University, Tuen Mun, Hong Kong.
E-mail address: alanlee@ln.edu.hk (A.L.F. Lee).

https://doi.org/10.1016/j.cognition.2020.104515
Received 25 July 2019; Received in revised form 7 November 2020; Accepted 9 November 2020
Available online 23 December 2020
0010-0277/

computation is built on Bayesian inference, priors on possible hierarchical structures play an important role in model predictions. The inductive biases incorporated into priors are vital for the inference process because many different forms of hierarchical structures may be consistent with the observed inputs.

As illustrated in Fig. 1 (an example adapted from Biederman, 1987), we perceive a watering can as a coherent perceptual object, but also as a spatial configuration of smaller elements including the handle, vessel, spout, and nozzle. Fig. 1 depicts a simplified example of the visual inputs in a learning situation, where we sometimes observe the whole object (e.g., the watering can), but sometimes only observe a partial view of the object (e.g., when some parts are occluded from certain viewpoints, or when the visible part itself could be a meaningful object, such as a "bucket" in the illustration). To represent the object, many hierarchical structures are consistent with the visual inputs. Fig. 1 (right panel) shows three possible structural representations, all having simple visual features represented as elements in the bottom layer. The first two hierarchical structures (H1 and H2) include different numbers of units in the top layer of the chunk representation. For example, structure H1 includes a chunk of the whole object, and a chunk of the "bucket" part (each consisting of two elements). Structure H2 includes four chunks in the top layer to represent a range of part combinations in addition to the whole object. Structure H3 includes two layers of chunks: a mid-layer with two-part chunks including bucket body (i.e., handle and vessel) and funnel (i.e., spout and nozzle), and a top layer with a single chunk of the watering can including all four elements. The first two structures (H1 and H2) increase the breadth of chunks in the top layer of a shallow hierarchy, whereas the third structure (H3) increases the depth of the hierarchy by introducing intermediate groupings of parts. These alternative hierarchies shown in Fig. 1 are simply examples; in general, there exist many more possible hierarchical representations.

Among these many possible hierarchical representations, how does the visual system select a structure to represent an object? Specifically, what are the information and learning processes involved in building a flexible representation of an object, so that we can see the object not only as a coherent perceptual whole, but also as a spatial configuration of parts? Fiser and Aslin (2005) showed that, after learning to see the "whole" object, the representations of chunks of small parts embedded in that object were either suppressed or non-existent (likely because they were unneeded to accomplish the experimental task). Similar results have been found in auditory statistical learning of words from syllables (Giroux & Rey, 2009). However, when more statistical cues are provided, a large chunk for a "whole" object can be broken into small parts. Fiser and Aslin (2005, Experiment 5) found that conditional probabilities between elements play an important role in determining what elements form a cohesive unit representing a single part, and what elements are separated into different parts. In order for small parts to be represented by separate chunks (e.g., bucket, funnel) in addition to a chunk representing the whole object (e.g., watering can), the conditional probabilities should indicate that the two parts will not always share the same boundary across all instances and that one part may appear without the other. Under these conditions, different conditional probabilities between the two parts may serve as a cue for part segmentation.

In the present study, we examine how prior knowledge about a certain part acquired in other learning contexts influences part segmentation within an object and the formation of a multi-layer hierarchical representation of the object. For example, before encountering a watering can for the first time, a child may have already seen instances of bucket-like objects, thereby forming a stable representation of a bucket. Such prior knowledge of bucket-like objects can be "recycled" to use for identifying a bucket part as a segmented component in a new object (e.g., watering can), which in turn facilitates the learning of other parts as representational units in order to form a multi-layer hierarchical structure of the watering can (e.g., Froyen, Feldman, & Singh, 2015; Kersten, Mammassian, & Yuille, 2004; Tu & Zhu, 2002). Such a representation allows reusable features to form a statistical distribution that tolerates estimation errors due to partial information (e.g., occlusion), yielding what is known as robust statistics (Fidler, Berginc, & Leonardis, 2006; Yuille & Mottaghi, 2016).

The present study sought evidence to test the role of reusable parts in learning both part and whole representations of a visual object within a multi-layer hierarchy. Grouping cues (e.g., curvature, junctions) were removed by using spatial configurations of novel shape elements, so that the formation of object representations can only rely on statistical regularities of co-occurrence of shape elements. We employed a standard visual statistical-learning paradigm, with a critical extension of adding pre-exposure of a part prior to training. Specifically, we inserted a familiarization phase, during which participants were exposed to a part (e.g., EG composed of two shape elements) that would later be embedded within a complex structure (e.g., EFGH composed of four shape elements) during subsequent training. We hypothesized that such prior exposure to the part (EG) would facilitate the formation of a representation unit of the *other* embedded part (FH), even though FH has

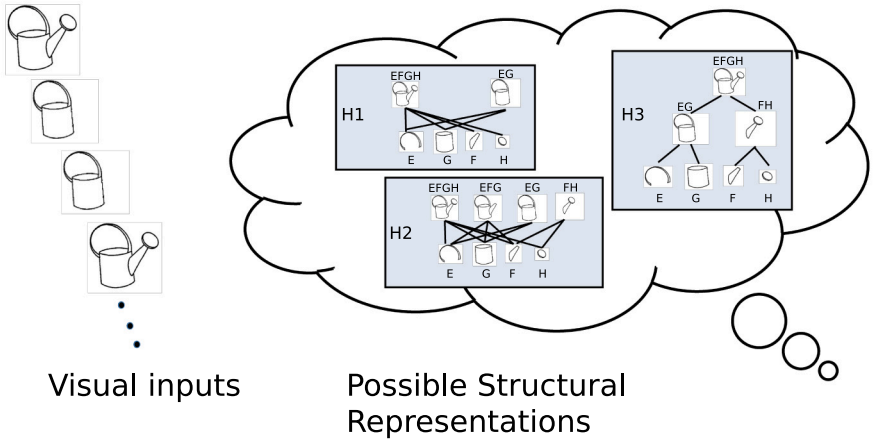

Fig. 1. Ambiguity in structural representations of visual objects. Many hierarchical structures are consistent with the visual inputs of an object, given that we sometimes observe the whole object and sometimes only observe a partial view of the object. The line drawing of the watering can is adopted from Biederman (1987). The letters denote shape elements that can be combined in different ways to form different structural representations of the object.

only co-occurred with part EG and is not a previously-known part by itself. We term this prediction the *parts-beget-parts* effect. In other words, via exposure to statistical regularities at different time points during learning, previously-learned knowledge of a part can serve as a "seed" to bootstrap a multi-layer, hierarchical structure with mid-level parts represented in the structure.

Our study consisted of four experiments using combinations of initially-unfamiliar shapes. In Experiment 1, we employed a "prior-familiarity" paradigm to first pre-expose participants to a pair of shapes (e.g., EG), and then present them with a larger quadruple configuration that was composed of the pre-exposed pair of shapes (EG) and a complementary pair of shapes (FH). We measured the subjective familiarity of a set of shape configurations, including the pre-exposed pair (EG), the complementary pair (FH), and the entire quadruple (EFGH) to assess the parts-beget-parts effect—inducing complementary pairs via a part-based representation. We then examined whether learning of the whole chunk and the part chunks occur together to form a multi-layer hierarchy via statistical learning (Experiment 2). Experiment 3 tested whether the parts-beget-parts effect can still survive when prior familiarity was induced via implicit learning rather than explicit supervision. In Experiment 4, we eliminated prior familiarity of a part, and instead varied co-occurrence frequencies of embedded parts during statistical learning. This manipulation allowed us to compare the effectiveness of two strategies that might facilitate formation of a hierarchical object representation: (1) reusing previously-learned parts (tested in Experiments 1–3), and (2) simultaneous learning based on co-occurrence frequencies (tested in Experiment 4).

2. Experiment 1

2.1. Participants

Sixty-one undergraduate students at the University of California Los Angeles (UCLA) participated for course credits. All participants had normal or corrected-to-normal vision, and were naïve to the purpose of the experiment. The experiments were approved by the UCLA Institutional Review board.

2.2. Stimulus and apparatus

We adopted the 24 shape units from Turk-Browne et al. (2005) (Fig. 2, left). Following the design in the study by Fiser and Aslin (2005), the study included two types of chunks: pairs and quadruples. Pairs were formed by diagonally positioning two shape units. Quadruples were formed by putting two pairs together. Because the choice of shape units and the assignment of the units to chunks were randomized across participants for counterbalancing, in what follows, we use letters to refer to individual shapes (e.g., A, B), and strings of letters to refer to chunks (e.g., quadruple ABCD, pair KL).The inventory of the shape elements is shown in Fig. 2.

These stimuli were fitted into a 5 × 5 grid to form a scene. The grid and the shapes were black (0 cd/m^2) on white background (146.5 cd/m^2). The whole grid subtended a visual angle of 8.4°. Each shape was about 1.2° × 1.2° in size, and was located at the center of its square cell within the grid. Stimuli were presented on a Viewsonic CRT monitor, with a 75 Hz refresh rate and a 1024 × 768 pixels resolution. The viewing distance was maintained at 57 cm using a chin rest. The experiment was run using MATLAB (MathWorks, Inc., Natick, MA) and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

2.3. Procedure

For each participant, 16 out of the 24 shapes were randomly chosen for use in the experiment. These 16 shapes were then randomly assigned to different chunks, as shown in the middle panel of Fig. 2. Chunks of shape pairs in the present paper were in the oblique spatial configuration.

Fig. 3 illustrates the procedure of Experiment 1. We first familiarized participants with a specific shape pair (Phase 1). Afterwards, participants went through a standard visual statistical learning procedure (Fiser & Aslin, 2001, 2005) that included repeated presentations of visual scenes cluttered with quadruple and pair chunks (Phase 2). For easier reference, in the following description we consistently use the example of EG as the pre-exposed embedded pair in Phase 1 and EFGH as the target quadruple in Phase 2. We investigated whether prior familiarity of an embedded pair (EG) within a quadruple (EFGH) would facilitate the subsequent learning of the *complementary* embedded pair (FH) in the quadruple.

As shown in the middle panel of Fig. 2, the chunk inventory also included a control quadruple (ABCD), which consisted of two embedded pairs (AC and BD). These two embedded pairs in the control quadruple had similar oblique structure as the complementary embedded pair (FH), were always presented together in the quadruple format, and had never been presented in stand-alone format during training. In the other words, both the complementary embedded pair FH and control embedded pairs AC/BD were *always* presented as embedded chunks within their respective quadruples with the same frequency, and had never been separated from their respective quadruples in any scenes during training. In Phase 2 training, the pre-exposed embedded pair EG were presented as an embedded pair within its parent quadruple EFGH and also as a stand-alone chunk presented with other chunk(s) in the scenes. Note that the assignment of shape units as pre-exposed embedded pair (EG) was randomized and counterbalanced across participants.

2.3.1. Procedure in phase 1

The goal of Phase 1 was to train participants to gain high familiarity of a shape pair EG, the pre-exposed embedded pair. Phase 1 consisted of

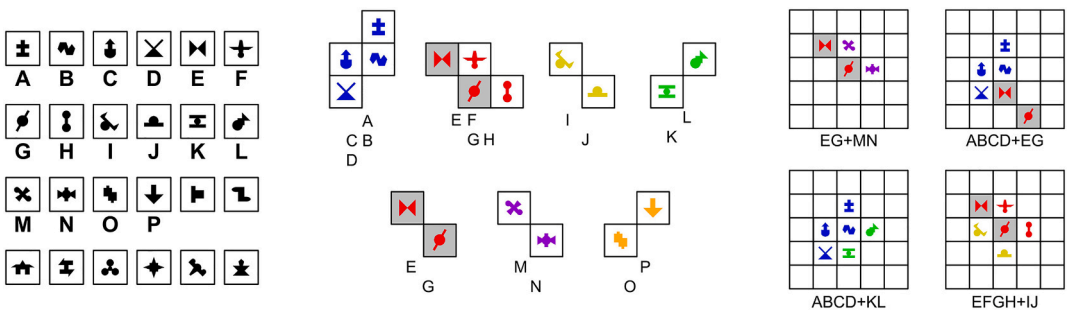

Fig. 2. Illustration of the stimuli used (shape colors and cell shading are for illustration purpose only; all stimuli were black and white). The shading highlights the trained embedded pair (EG in this example). Left: The 24 shape units adopted from Turk-Browne et al. (2005). The letters A to P illustrate 16 units that were randomly drawn from the 24 shape units for one participant. Middle: Examples of visual chunks in the training inventory, which were either quadruples (ABCD and EFGH) or pairs (e.g., IJ, KL, EG). Right: Examples of training scenes formed by combining two chunks.

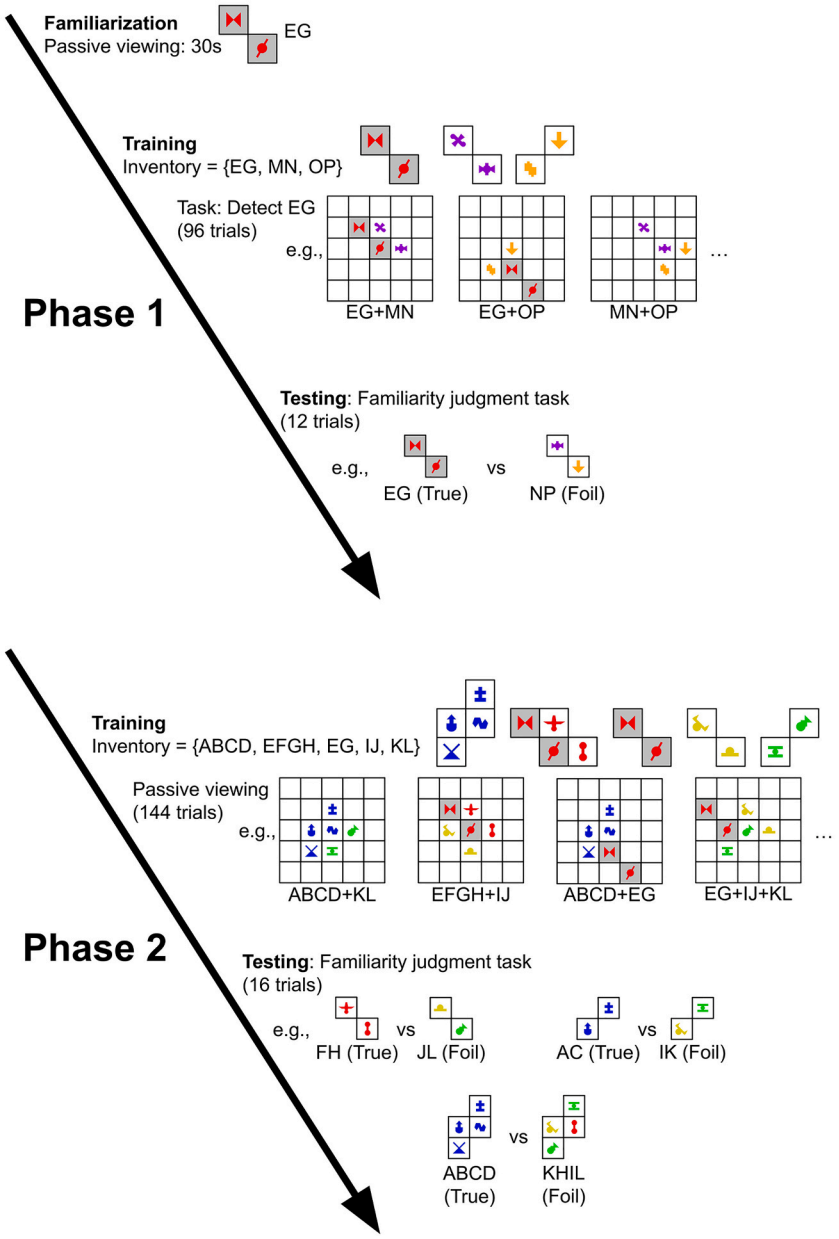

Fig. 3. Illustration of procedure for Experiment 1. To better illustrate the design, we highlight the chunks using colors, and the trained embedded pair (EG in this example) using shading. In all experiments here, all stimuli were black and white. Top: In Phase 1, participants learned the embedded target pair (EG) through a detection task. The result of such learning was measured during Phase 1 Testing. Bottom: Phase 2 was statistical learning. Training scenes contained six shapes, each consisting of either a quadruple plus a pair or three pairs. Familiarity was assessed for all the trained chunks and their embedded pairs during Phase 2 Testing.

three blocks: Familiarization (30 s), Training (5 min), and Testing (1 min). During Familiarization, participants were presented for 30 s with the target: the pre-exposed pair EG. They were asked to pay attention to this pair of shapes and were told that they would be performing some task related to this target afterwards.

The subsequent Phase 1 training session consisted of 96 trials. On each trial, participants viewed a scene containing four shape elements from two pairs, which were chosen from an inventory consisting of three pairs: EG, MN, and OP. The two pairs within each scene were spatially connected as they shared at least two common edges in the grid (see Fig. 3, top, for some example training scenes). Depending on the specific orientation of the trained embedded pair EG, the maximum number of possible configurations to place two spatially-connected pairs within a 5 $\times$ 5 grid was either 48 or 52. For each participant, we randomly sampled 16 scenes each with a combination of the two pairs. This resulted in three types of scenes being presented during Training: EG + MN (16 scenes), EG + OP (16 scenes), and MN + OP (16 scenes). These 48 distinct scenes were repeated to form two blocks. Scene orders were block-randomized. Accordingly, each of the three pairs appeared 64 times, or 2/3 of the total 96 trials.

On each training trial of Phase 1, the scene was presented for two seconds, followed by a one-second pause. Participants performed a detection task by indicating whether the scene contained the pair of shapes that had been presented during Familiarization (EG in this example). After Phase 1 Training, participants were shown their detection accuracy over the 96 trials, followed by a one-minute rest.

During Phase 1 Testing, there were two types of testing scenes for measuring participants' familiarity judgment. A "true" testing scene contained a chunk from the training inventory. A "foil" testing scene contained a chunk with shape combinations that had never been presented during Training. As an example shown in Fig. 2, a "true" testing scene would be one of the three pairs in the training inventory (i.e., EG, MN, or OP); the "foil" scenes contained two shapes that formed a pair (e.g., MO, EP, or NG) that was not included in the training inventory. In each trial during Testing, the "true" and "foil" testing scenes were simultaneously presented side by side for 4 s. The true and foil chunks in the same test trial shared the same spatial configuration, were located at the same position within their respective grids, and did not shared any

common shape. Assignment of the left or right positions was randomized and counterbalanced for the true and foil testing scenes. After the testing scenes disappeared, participants performed a discrimination task (Fiser & Aslin, 2001, 2005), indicating which test scene (left or right) appeared more familiar. There were 12 test trials in total, with each training pair being tested four times.

2.3.2. Procedure in phase 2

The goal of Phase 2 was to train participants with more complex visual scenes generated from an inventory that contained quadruples and pairs. Two quadruples (ABCD and EFGH) were included in the inventory. One of these two quadruples (EFGH) was the target quadruple, which contained the pre-exposed embedded pair (EG) that the participants had been trained in a supervised manner during Phase 1. The other quadruple (ABCD) served as a control quadruple that was presented with equal occurrence frequency as the quadruple (EFGH), but none of the embedded pairs were presented stand-alone in the training scenes. As a result, excluding the pair EG, participants would view all other embedded pairs within these quadruples (i.e., FH, AC, and BD) with an equal frequency during training. If the prior familiarity of EG had no effect on the learning of its complementary part FH, participants would show the same degree of familiarity with the complementary pair FH as with the control embedded pairs AC/BD.

Phase 2 (Fig. 3, bottom) consisted of a Training (10 min) and a Testing (2.5 min) session, with a one-minute rest in between. The training inventory of Phase 2 contained four types of chunks: (1) the target quadruple EFGH, (2) the control quadruple ABCD, (3) the pairs IJ and KL, and (4) the pre-exposed pair EG, termed as *trained embedded pair* in the following text. Note that the shapes M, N, O, and P from Phase 1 were not used in Phase 2.

The training scenes consisted of six shape elements. Each scene was constructed with a quadruple and a pair, or three pairs. The selected chunks were spatially connected within the 5 × 5 grid. The number of all possible ways to place a quadruple and a pair within a 5 × 5 grid in a spatially-connected manner is either 24 or 34 depending on the relative orientations between the quadruple and the pair. The number of possible ways to place three pairs together is 80. Four types of training scenes were randomly sampled from all the possible scenes: control quadruple + a pair (ABCD + IJ and ABCD + KL, sampled 8 scenes each), target quadruple + a pair (EFGH + IJ and EFGH + KL, sampled 16 scenes each), control quadruple + trained embedded pair (ABCD + EG, sampled 16 scenes), and three-pair scenes (EG + IJ + KL, sampled 8 scenes). This design resulted in 72 distinct training scenes, which were repeated to form two blocks with scene orders independently block-randomized. Table 1 shows the frequency of each training scene and the frequency for each inventory chunk to be presented out of the 144 total training trials. Note that the two quadruples (target and control) were presented with equal frequencies, yielding the same presentation frequency for the embedded pairs (i.e., 0.44 for both FH in the target quadruple EFGH and AC/BD in the control quadruple ABCD). Each training scene was presented for two seconds, followed by a one-second pause. Then, a text prompt was presented. There was no task involved, and participants responded to the prompt by pressing the space bar, which then triggered the presentation of the next training scene.

Phase 2 Testing included the familiarity-discrimination task on the following eight types of chunk. We tested all the four types of chunks that were in the inventory: (1) EFGH, (2) ABCD, (3) IJ/KL, and (4) *trained embedded* pair EG. The critical testing trials of "true" chunks focused on the following embedded pairs that were within the two quadruples, but had not been presented as separate chunks in the training phases: (5) the *complementary pair* FH embedded in the target quadruple, (6) a control embedded pair AC or BD in the control quadruple, (7) a target non-oblique pair EF or GH, and (8) a control non-oblique pair AB or CD. The key comparison was on the familiarity between the complementary pair FH and the control embedded pair AC or BD, neither of which had been presented as separate chunks during training. The testing block included two trials for each of the eight types of chunks. For each participant, these 16 testing trials were presented in random order, and the familiarity judgment accuracy for each type of chunk was computed by averaging the accuracy over the two trials.

In each test trial, one true chunk was compared with a foil chunk consisting of random shape combinations that had never been presented during Training. In addition to the constraints imposed on Phase 1 Testing, the foil chunks were generated with the following constraints: (1) Shapes in the "true" and "foil" chunks were taken from different chunks in the training inventory; (2) The "foil" chunks had to contain shapes from at least *two* different chunks; and (3) A "foil" chunk could not contain any shape from the trained embedded pair (i.e., E or G in our example).[1]

It should be noted that a new foil chunk was generated for each new trial. To minimize the impact of familiarization during Testing, we included only two test trials for each chunk type. Therefore, participants would see the true chunk only *one* time more than the *second* foil chunk during Testing. While this extra exposure during Testing could increase participants' familiarity with the true chunks, the effect, if any, would be matched across all chunk types—critically between the complementary pair FH and control embedded pair AC/BD. Thus the repetition of chunks in Testing cannot account for the parts-beget-parts effect.

2.4. Results and discussion

One participant was removed from the analysis due to a very low detection accuracy of 0.72 in the Phase 1, which was more than four standard deviations below the mean. For the remaining 60 participants, the average detection accuracy during Phase 1 Training was high (M = 0.98, SD = 0.04). For the familiarity-discrimination task in Phase 1 Testing, the mean accuracy for the trained embedded pair EG was M = 0.98 (SD = 0.12), with 57 out of 60 participants obtained a perfect score. These results indicate that the participants encoded and remembered the trained embedded pair successfully after Phase 1.

Fig. 4 shows the results in Phase 2 Testing. The trained embedded pair (EG) maintained ceiling accuracy (M = 0.98, SD = 0.14) in familiarity discrimination. Participants also showed well above-chance accuracies for quadruples (EFGH and ABCD), together with the other two pairs IJ, KL in the training inventory (averages ranging from 0.70 to 0.78; all p's < 0.001), indicating successful acquisition of visual chunks in the inventory from statistical learning.

To examine the familiarity of the embedded pairs, we considered three critical embedded pairs (complementary pair FH, control embedded pair AC or BD, and the target non-oblique pair EF or GH) in a repeated-measures ANOVA. Importantly, we found a parts-beget-parts effect, as participants identified the complementary pair (FH; M = 0.73, SD = 0.36) with significantly higher familiarity than the control embedded pair (AC or BD; M = 0.59, SD = 0.40; Tukey test: $t(59) = 2.38$, $p = .049$, Cohen's $d = 0.31$), despite their equal presentation frequency and within-pair element predictability throughout training. Performance for non-oblique pair (EF or GH) was poor without showing significant difference from the chance performance (M = 0.57, SD = 0.40, $p = .17$). As predicted by probabilistic learning (Orban et al., 2008), we replicated the finding that the familiarity accuracy for the complementary pair (FH) was significantly higher than non-oblique control pair (EF or GH) (Tukey test: $t(59) = 2.80$, $p = .016$, Cohen's $d = 0.36$). These results suggest that the superior performance for the complementary

[1] There was an exception to constraint 3 when the true chunk was the target quadruple EFGH or the non-oblique target pair EF or GH. Because participants were very familiar with the trained embedded pair EG and the individual shape units E and G, they could be biased to choose the true chunk that contained either of the shapes (or both) over any foil chunk that did not contain them. In order to minimize such bias, when the true chunk contained E or G or both, the shape(s) would appear in exactly the same positions within the foil chunk.

Table 1
Frequencies of training scenes and chunks for Phase 2 in Experiments 1–3. For each type of chunk, the non-parenthesized numbers represent the frequencies for the chunk to appear as a separate chunk (i.e., not embedded within any chunks). Numbers in parentheses represent the frequencies at which the chunk was displayed as a part embedded within a complex chunk. Specifically, for the trained embedded pair EG, the overall occurrence frequency was 48 (separate) + 64 (embedded) = 112, out of the 144 total trials.

		Control quadruple	Target quadruple	pair	pair	Trained embedded pair	complementary pair	Control embedded pair
		ABCD	EFGH	IJ	KL	EG	FH	AC/BD
Training Scenes	Freq.							
ABCD + IJ	16	16		16				(16)
ABCD + KL	16	16			16			(16)
ABCD + EG	32	32				32		(32)
EFGH + IJ	32		32	32		(32)	(32)	
EFGH + KL	32		32		32	(32)	(32)	
EG + IJ + KL	16			16	16	16		
Total	**144**	**64**	**64**	**64**	**64**	**48 (64)**	**(64)**	**(64)**
Rel. freq.		**0.44**	**0.44**	**0.44**	**0.44**	**0.33 (0.44)**	**(0.44)**	**(0.44)**

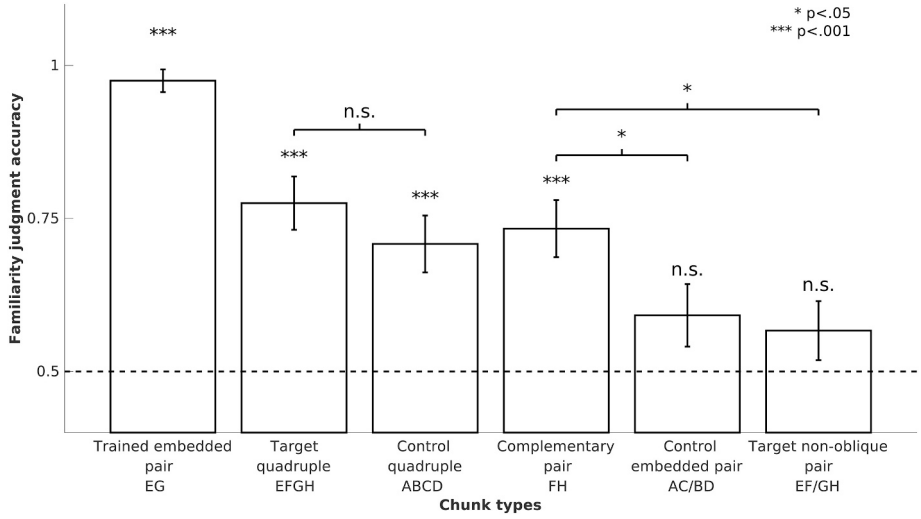

Fig. 4. Results of familiarity judgments in Phase 2 Testing for different chunk types in Experiment 1. Labels on the horizontal axis are based on the example in which EG is the trained embedded pair (see text). Error bars indicate ±1 S.E.M.

pair FH and the trained embedded pair EG resulted from encoding the spatial configuration of the two pairs as two unitary components in the hierarchical representation.

One potential explanation for the parts-beget-parts effect could be from the difference in familiarity of the respective quadruples. Namely, the better performance for the complementary pair FH may be from a "spill-over" association from the higher familiarity of the target quadruple (EFGH) than of the control quadruple (ABCD), because EFGH includes the well-trained embedded part EG. However, Experiment 1's data showed that the familiarity accuracies were not different between the target and control quadruples ($t(59) = 1.24$, $p = .22$, Bayes Factor = 0.29, favoring the null that the two were equal). In summary, these results from comparisons among pairs and among quadruples are consistent with a multi-layer hierarchical representation that included the whole chunk (quadruple) at one level and the two part pairs at the other level (i.e., the trained-embedded pair and previously-unseen complementary pair) in a hierarchical structure. Table S1 in Supplementary Material shows the descriptive statistics for each chunk type tested in Phase 2 of Experiment 1.

3. Experiment 2

Findings from Experiment 1 provide evidence to show the part-beget-part effect, and familiarity of various chunks consistent with learning a multi-layer hierarchical representation. However, participants may have chosen the target quadruple EFGH to be more familiar than the foil because they recognized the parts of EG and FH, rather than representing EFGH as a unitary, "whole" chunk composed of these two parts. If this were the case, participants would be equally familiar with the true quadruple EFGH in which part EG is on the left of FH as included in training trials, and with a quadruple that had the embedded pairs spatially swapped (FEHG) in which part EG is on the right of FH, as illustrated in Fig. 5. However, if participants had learned the target quadruple EFGH as a "whole", they would be sensitive to its spatial layout of the parts and be able to distinguish between the true quadruple with its "swapped foil". To address this possible explanation, we included a "true vs swapped" test in Experiment 2 (top panel of Fig. 5) for the target quadruple (i.e., EFGH vs FEHG)), in addition to measuring the familiarity of the complementary pair FH.

3.1. Methods

Eighty UCLA undergraduate students participated for course credits. The stimuli and procedure were identical to those in Experiment 1, except for Phase 2 Testing, when we tested four chunk types only: 1) the complementary pair (FH) versus a foil pair composed in the same way as in Experiment 1, 2) the control embedded pairs (AC and BD) versus a foil pair composed in the same way as in Experiment 1, 3) the target quadruple (EFGH) versus a part-swapped quadruple (FEHG), and 4) the control quadruple (ABCD) tested against a swapped foil (BADC). The inclusion of the control quadruple was to balance out the appearance frequency across shape elements during testing. Same as Experiment 1,

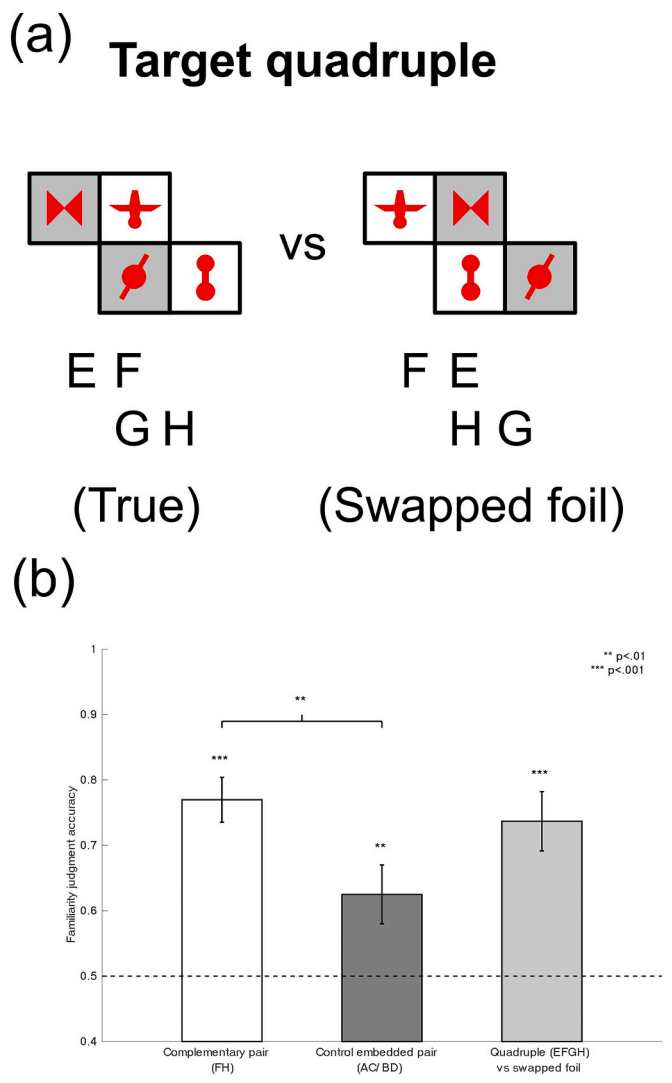

Fig. 5. (a): Illustration of the swapped-foil test (on the target quadruple EFGH as an example; trained embedded pair EG highlighted with shading) in Experiment 2. (b): Results of Experiment 2. Familiarity-discrimination accuracies of the complementary pair (white bar), the control embedded pair (dark gray bar), and the the target quadruple (light gray bar) tested against the swapped foil. All error bars indicate ±1 S.E.M. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

participants completed two trials of familiarity-judgment task for each type of chunks, and the familiarity-judgment accuracy was computed by averaging the accuracy over these two trials. Among the 80 participants, 20 participants were tested on the two pair chunks (four trials in total, with order randomized). The other 60 participants were tested on all four chunk types (eight trials in total, with order randomized).

3.2. Results and discussion

The analysis excluded the data of four participants. Two of them had exceptionally poor detection accuracy in Phase 1 Training (0.41 and 0.47), which was more than four SD's below the mean. The other two indicated at the end of Experiment 2 that they had participated in a similar experiment of visual statistic learning at UCLA. The remaining 76 participants (19 tested with two chunk types, 57 tested with four chunk types) yielded close to perfect performance in the detection task during Phase 1 Training ($M = 0.995$, $SD = 0.0110$) and in familiarity discrimination during Phase 1 Testing (trained embedded pair: $M = 0.987$, $SD = 0.056$).

Fig. 5 shows the results of Phase 2 Testing for the complementary pair, the control embedded pair, and the target quadruple (tested against swapped-foil) in Experiment 2. As in Experiment 1, the familiarity accuracy for the complementary pair (M = 0.77, SD = 0.30) was significantly different from chance ($t(75) = 7.86$, $p = 2e\text{-}11$, Cohen's $d = 0.90$). Unlike in Experiment 1, the accuracy for the control embedded pair (M = 0.63, SD = 0.39) was also significantly above chance ($t(75) = 2.78$, $p = .007$, Cohen's $d = 0.30$), indicating that participants may have learned the embedded pair in the control quadruple. This result could be due to learning during the test of the control quadruple displayed with its swapped foil, which gave away the embedded structure of the control quadruple. Most importantly, despite the above-chance familiarity of the control embedded pair, we found again the parts-beget-parts effect, as the average familiarity accuracy for the complementary pair was significantly higher than that for the control embedded pair (paired t-test: $t(75) = 2.70$, $p = .009$, Cohen's $d = 0.31$). We compared the parts-beget-parts effects revealed in Experiments 1 and 2 in a 2 × 2 mixed-factorial ANOVA with one within-subjects factor of embedded pair type (complementary vs control pairs) and one between-subjects factor of experiment (Experiment 1 vs 2). The result did not reveal a significant interaction effect between the experiment and the difference between complementary pair and control pair ($F(1, 134) = 0.001$, $p = .97$, partial $\eta^2 = 1e\text{-}5$). The main effect of chunk type was significant ($F(1, 134) = 11.91$, $p = .0007$, partial $\eta^2 = 0.082$), but the main effect of experiment was not ($F(1, 134) = 0.552$, $p = .46$, partial $\eta^2 = 0.004$). These results suggest that both experiments showed qualitatively similar parts-beget-parts effect, i.e., the familiarity accuracy was higher from complementary pairs than control pairs.

Another key finding is that the familiarity accuracy for the target quadruple in the "swapped test" was significantly above chance (Target: M = 0.74, SD = 0.34, $t(56) = 5.23$, $p = 3e\text{-}6$, Cohen's $d = 0.69$). This result shows that participants were able to reliably discriminate between the target quadruple (EFGH) and its swapped foil (FEHG), suggesting that their familiarity on the target quadruple (EFGH) cannot be explained solely by combining the separate familiarity of the two embedded pairs (EG + FH).

4. Experiment 3

The first two experiments show that prior familiarity to an embedded part (e.g., EG) within a more complex structure (EFGH) is sufficient to induce the encoding of the complementary part (FH) as a representation unit in the subsequent testing. However, in order to introduce the prior familiarity of the embedded part to participants, the reference to the embedded part was *explicit* throughout the Familiarization block and the detection task in Phase 1 training. Although our results show that this procedure led to high familiarity of the complementary part, the effect could be due to explicit deduction from the knowledge of the pre-exposed chunks. It is possible that participants were biased by the tasks in Phase 1 and consciously deduced the existence of the complementary part based on their knowledge about the trained embedded part.

In Experiment 3, we aimed to test whether explicit knowledge of the embedded pair (EG) is necessary for the subsequent learning of the complementary part (FH) to show the parts-beget-parts effect. If explicit familiarity of an embedded part was necessary, the effect obtained in the previous experiments would disappear when Phase 1 learning was made implicit without explicitly defining the trained embedded pair through tasks. It should be clarified that, in the context of Experiment 3, we use "implicit learning" only to refer to the training process of the embedded pair (EG) that was "without an explicit task" as opposed to that in Experiments 1 and 2.

4.1. Methods

Eighty-five UCLA students participated for course credits. The stimuli and procedure were identical to those used in Experiment 2, except the following. In Phase 1 of Experiment 3, we omitted the following steps and tasks: the entire Familiarization, the detection task during Training, and Testing. To ensure Phase 1 Training was adequate, the number of training scenes was increased from a total of 96 to 144 (i.e., 24 scenes sampled from each of the three types of two-pair scenes; repeated to form two blocks with trial order block-randomized). In Phase 2 Testing, we omitted the swapped-foil tests for quadruples, i.e., we focused on the parts-beget-parts effect by testing only the complementary pair (FH) and the control pairs (AC/BD). In summary, the procedure of Experiment 3 was as follows: Phase 1 Training, Phase 2 Training, and Phase 2 Testing. Importantly, during Training in both Phases 1 and 2, participants were only asked to passively view the stimuli without performing any task. They were not informed about anything related to the subsequent familiarity-discrimination task.

4.2. Results and discussion

As shown in Fig. 6, we replicated the parts-beget-parts effect that familiarity accuracy for the complementary pair FH was significantly higher than for the control embedded pair AC/BD ($t(84) = 2.47$, $p = .016$, Cohen's $d = 0.27$), even when embedded pairs in the prior exposure were learned implicitly without performing any tasks and tests. Familiarity discrimination accuracy for the complementary pair ($M = 0.61$, $SD = 0.38$) was significantly above chance ($t(84) = 2.58$, $p = .012$, Cohen's d = 0.28), but the accuracy for the control embedded pair ($M = 0.46$, $SD = 0.38$) was not different from chance-level performance as expected ($t(84) = 0.87$, $p = .39$, Bayes Factor = 0.172, moderate evidence favoring the null).

To evaluate the impact of explicit tasks on learning trained embedded pairs on the parts-beget-parts effect, data from Experiments 1–3 were compared using a mixed-factorial ANOVA, with embedded pair type (complementary vs control) as the within-subjects factor and learning type (explicit in Experiments 1 and 2, and implicit in Experiment 3) as the between-subjects factor. The two-way interaction between embedded pair type and learning was not significant ($F(1, 219) = 0.001$, $p = .97$). The results revealed significant main effects for embedded pair type ($F(1, 219) = 17.10$, $p < .001$, partial $\eta^2 = 0.072$) and learning type ($F(1, 219) = 15.33$, $p < .001$, partial $\eta^2 = 0.065$). Specifically, the post-hoc Tukey tests revealed that familiarity accuracy for complementary pair was significantly higher than that for the control embedded pair for both types of learning (for explicit learning: $t(219) = 3.36$, $p = .005$; for implicit learning: $t(219) = 2.62$, $p = .047$). These results suggest that high prior familiarity to an embedded chunk, regardless of whether the familiarity has been acquired explicitly with specific tasks or implicitly through statistical learning, facilitated the learning of part-based hierarchical structure by encoding its complementary part as a representation unit.

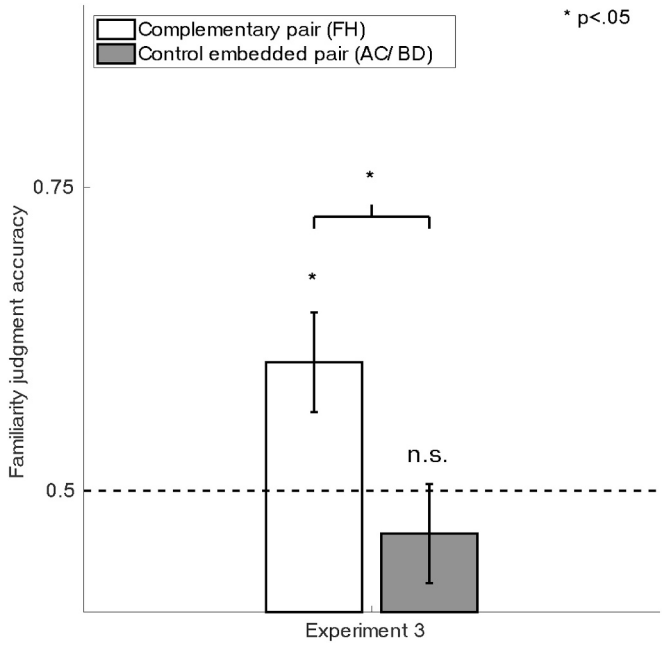

Fig. 6. Familiarity accuracies of the complementary pair (white bars) and the control embedded pair (gray bars) in Experiment 3. Error bars indicate ±1 S.E.M.

5. Experiment 4

In Experiments 1–3, high familiarity on the embedded pair EG was acquired *prior* to the learning of quadruples and pairs. The learning of the embedded pair EG was, therefore, temporally separable from the training of the quadruple EFGH. Would the parts-beget-parts effect still maintain when learning of the part and the whole occurred together?

To address these question, Experiment 4 removed the entire Phase 1 in previous experiments, but introduced the embedded pairs with different co-occurrence frequencies in a standard, one-phase statistical learning procedure. If simultaneous learning of the embedded pair and the quadruple based on co-occurrences could facilitate building of a part-based representation, a similar parts-beget-parts effect in Experiments 1–3 would be observed. However, if high familiarity to the embedded pair was critical to induce part-based representations, the lack of strong familiarity to the embedded pair would mitigate the effect of encoding the complementary pair as a representation unit, resulting in a weaker or nonexistent parts-beget-parts effect.

5.1. Methods

Eighty undergraduate students at UCLA participated for course credits. This sample size would yield a power of 0.80 for detecting an effect with the same effect size (Cohen's d = 0.28) observed in Experiment 3 for the complementary pair (FH). To be conservative, we chose the effect size in Experiment 3 in this power analysis because it was the smallest effect size observed among Experiments 1–3 for the parts-beget-parts effect (Cohen's $d = 0.64$ for Experiment 1; 0.90 for Experiment 2).

This experiment used the training inventory, which included two base quadruples (ABCD, EFGH) and two pairs (IJ, KL). In addition, we included two embedded pairs (e.g., AC, EG) within the quadruples (ABCD and EFGH, respectively). These embedded pairs were presented as separate chunks in some training scenes. We refer to them as trained embedded pairs, and their counterparts (e.g., BD, FH) in the quadruples as complementary pairs. The choice of trained embedded pairs within a quadruple (i.e., AC or BD within ABCD; EG or FH within EFGH) was similarly randomized and counterbalanced across participants as in Experiments 1–3.

Training scenes were constructed by putting together a quadruple and a pair within the 5 × 5 grid with the same spatial constraint used in Experiments 1–3. There were 24 or 34 total ways to place a quadruple and a pair within the grid, depending on their relative orientations. The pair could be an untrained pair (IJ or KL), or a trained embedded pair (AC or EG). We used all possible 164 distinct scenes with the combinations of one quadruple and one pair as training stimuli. Training frequencies of each inventory chunk were constrained by spatial configuration limits, but were also set to maintain the equal frequency between the two quadruples (ABCD and EFGH: 0.50), between the two untrained pairs (IJ and KL: 0.35), and between the two embedded pairs (e.g., AC and EG: 0.15). These training frequencies were different from those used in Phase 2 Training in Experiments 1–3 because we attempted to maintain the same inventory while only train one

embedded pair in each quadruple. These constraints required balancing the training frequency for the trained embedded pair, which limited the number of possible training scenes. Table 2 shows the training frequency of each chunk appearing in the training scenes.

Fig. 7 shows the procedure of Experiment 4, which consisted of Training and Testing. During Training, the 164 training scenes were presented in a randomized order across participants. All other aspects were the same as the procedure in Phase 2 Training in Experiments 1–3.

To minimize the effect of learning during testing, each unique true chunk was tested only once per participant. This resulted in 12 testing trials in total, including two trials for the quadruples (ABCD and EFGH), two trials for the untrained pairs (IJ and KL), two trials for the trained embedded pairs (AC and EG), two trials for the complementary pairs (BD and FH), and four trials for embedded triplets extracted from the quadruple (ABC, BCD, EFG, and FGH). We added the tests for the embedded triplets in Experiment 4 so that our results could be compared with the following finding in Orban et al. (2008): although familiarity on pairs embedded within a quadruple was at chance level, familiarity on triplets embedded within a quadruple was significantly above-chance. If we could not replicate this "triplet-but-not-pair" result pattern, it would suggest that our training procedure may not be sufficient for inducing familiarity on any embedded parts. Otherwise, we could compare our results with theirs in the context of training embedded parts. The foils were generated in the same way as in Phase 2 Testing of Experiments 1.

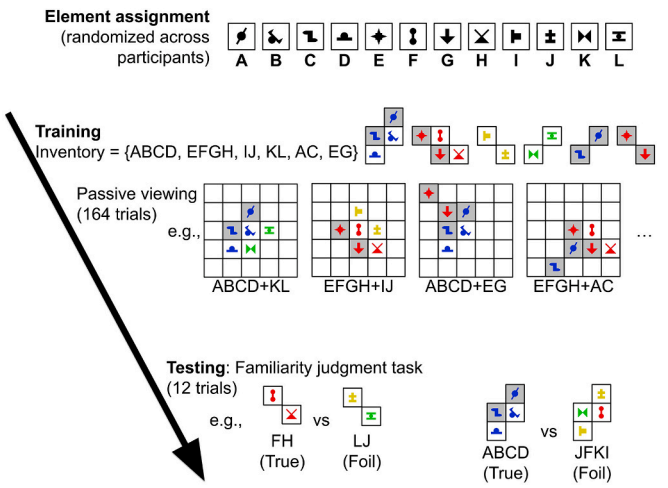

Fig. 7. Illustration of the shape units, training inventory, and procedure for Experiment 4. Colors and shading are for illustration purpose only (as in Figs. 2 and 3). All stimuli were presented in black and white in the experiment.

5.2. Results and discussion

Fig. 8 shows the average familiarity accuracy across participants for each tested chunk type in Experiment 4. Table S2 in Supplementary Material shows the statistics for one-sample *t*-test against chance for the familiarity accuracy for each chunk type. The familiarity discrimination accuracy was significantly above chance for the quadruples ABCD/EFGH ($M = 0.73$, $SD = 0.31$, $t(79) = 6.55$, $p = 5e\text{-}9$), the pairs IJ/KL ($M = 0.62$, $SD = 0.34$, $t(79) = 3.13$, $p = .002$), the trained embedded pairs AC/EG ($M = 0.63$, $SD = 0.32$, $t(79) = 3.46$, $p = 9e\text{-}4$), and the triplets ABC/BCD/EFG/FGH ($M = 0.66$, $SD = 0.27$, $t(79) = 5.44$, $p = 6e\text{-}7$). But, critically, despite the familiarity of the trained embedded pairs was significantly above chance, we did not find the parts-beget-parts effect, as the accuracy for their complementary pairs BD/FH was not different from chance ($M = 0.54$, $SD = 0.35$; $t(79) = 0.97$, $p = .334$, Bayes Factor = 0.194, favoring the null hypothesis that the mean = 0.50). We compared recognition accuracy of complementary pairs BD/FH with that for the trained embedded pair (AC/EG). The difference between the two conditions, as shown in Fig. 8, were trending but not significant ($t(79) = 1.80$, $p = .075$, Cohen's d = 0.20); nor did the comparison between complementary pairs (BD/FH) and pairs (IJ/KL) ($t(79) = 1.53$, $p = .13$, Cohen's d = 0.17).

Results from Experiment 4 were consistent with previous findings that participants were able to gain familiarity for the quadruples, as well as an embedded part that took up more than half of the chunk (i.e., embedded triplets) (Fiser & Aslin, 2005; Orban et al., 2008). However, without strong familiarity on the embedded pair within a quadruple, participants did not show the parts-beget parts effect to induce the familiarity of the complementary pairs.

It should be noted that there were other differences between Experiment 4 and Experiments 1–3 in addition to the number of phases in procedure. First, in Experiment 4, one embedded pair was taken from each of the two quadruples (AC from ABCD; EG from EFGH) to be presented as a separate pair during training. But, in Experiments 1–3, one quadruple (ABCD) was designated to be a control quadruple while the trained embedded pair (EG) was only taken from the other, target quadruple (EFGH). Second, training frequencies among the chunks were not identical between Experiment 4 and Experiments 1–3 (for quadruples and the complementary pairs: 0.50 in Experiment 4 and 0.44 in Experiments 1–3; for trained embedded pairs: 0.65 in Experiment 4 and 0.78 in Experiments 1–3).

Despite such differences, the familiarity accuracies for the quadruples in both Experiment 1 ($M = 0.75$) and in Experiment 4 ($M = 0.72$) were similar and significantly above chance. These values were also comparable to that found in Experiment 4 in Fiser and Aslin's (2005) study (mean ~ 0.72). This converging result suggests that the quadruples were still successfully learned in Experiment 4. However, the lack of the parts-beget-parts effect, i.e., the chance-level familiarity for the complementary pair (BD/FH), indicated that the learning setup in Experiment 4 did not suffice to trigger formation of a multi-layer hierarchical representation for the quadruple (ABCD/EFGH). Taken together, results from Experiments 1–4 suggest that the strength of familiarity of a part significantly impacted the formation of a part-based, hierarchical representation of the whole, as demonstrated by the familiarity for the complementary part.

Table 2
Frequencies of training scenes and chunks for Experiment 4. All within-experiment comparisons were ensured to be equal-frequencies, such that appearance frequency could not be confound for any effect.

		Quadruple	Quadruple	Pair	Pair	Trained embedded pair	Trained embedded pair
		ABCD	EFGH	IJ	KL	AC	EG
Training Scenes	Freq.						
ABCD + IJ	24	24		24		(24)	
ABCD + KL	34	34			34	(34)	
ABCD + EG	24	24				(24)	24
EFGH + IJ	34		34	34			(34)
EFGH + KL	24		24		24		(24)
EFGH + AC	24		24			24	(24)
Total	**164**	**82**	**82**	**58**	**58**	**24 (82)**	**24 (82)**
Rel. freq.		**0.50**	**0.50**	**0.35**	**0.35**	**0.15 (0.50)**	**0.15 (0.50)**

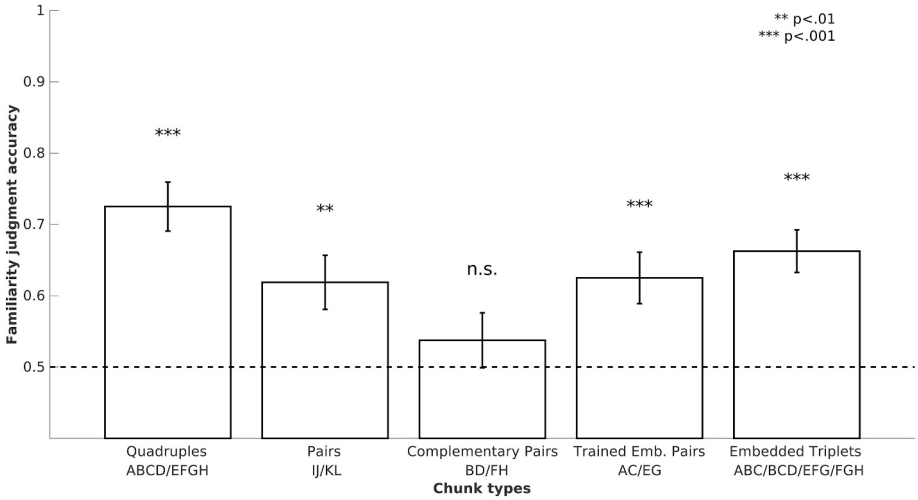

Fig. 8. Average familiarity judgment accuracies in Experiment 4. Error bars indicate ±1 S.E.M.

To further highlight this effect of strong part familiarity on learning of a multi-layer hierarchical representation, we compared the effect sizes in terms of Hedge's *g* (Hedges, 1981) across different procedures for learning the trained embedded part (EG) in the present study. The effect was defined as the difference between familiarity judgment accuracy for the complementary pair (FH) and the chance-level performance (0.50 for two-alternative forced-choice). Because the procedure for Phase 1 Training was explicit in both Experiments 1 and 2, we grouped the data of those two experiments. As shown in Fig. 9, the effect size was largest when Phase 1 Training was explicit, was reduced but still significantly above zero when it was implicit in Experiment 3, and was the weakest when the training was absent in Experiment 4.

6. General discussion

Using a two-phase training paradigm, we found that prior familiarity with a part of a complex chunk facilitates the formation of a multi-layer hierarchical representation of complex configurations. Specifically, after participants were pre-exposed to one part of a complex and novel configuration, subsequent learning enabled them to acquire familiarity of the complementary part of the complex configuration, even when the complementary part had not been presented separately from the whole.

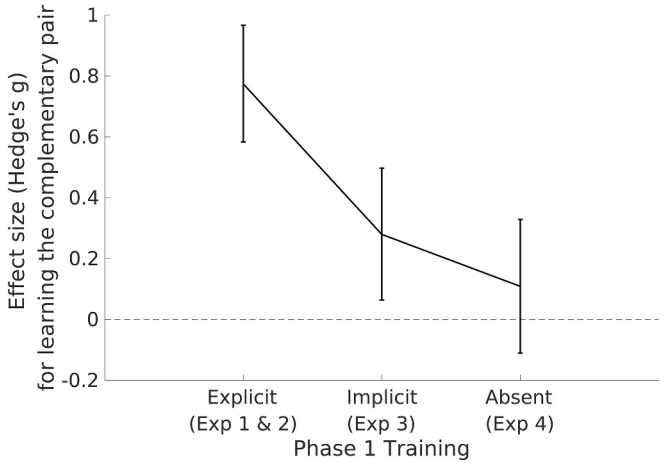

Fig. 9. Effect sizes of familiarity judgment accuracies for the complementary pair as a function of Phase 1 Training procedures. Error bars represent 95% confidence intervals (computed using Hentschke and Stüttgen's (2011) toolbox for measuring effect size.

We refer to this representation-formation process from one part to its complementary part as the parts-beget-parts effect.

Our study is not the first to demonstrate that parts embedded within a whole can be encoded and remembered via visual statistical learning. Fiser and Aslin (2005, Experiment 5) showed that a large structure can be broken into parts according to conditional probabilities of these parts during learning, while the joint probabilities among the elements were controlled. For example, the conditional probability of a part appearing next to another part was either $P(Y \mid X) = 1$ or $P(Z \mid X) = 2/3$ during training. Part X, thus, had higher predictability (defined as conditional probability) on Y than on Z in terms of their occurrence during training. In the subsequent testing, participants judged the part pair with higher conditional probability (i.e., XY) to be more familiar than the pair with lower conditional probability (i.e., XZ). This result indicated that predictability among parts can serve as a cue to divide a large structure of "whole" into smaller "parts".

The parts-beget-parts effect reported in the present study can be viewed as an extension of encoding smaller parts via visual statistical learning over time and across learning contexts. Specifically, prior knowledge of certain parts acquired in an early learning phase provides a "seed" or a starting point, enabling the subsequent learning of multi-layer hierarchical structures. What are the essential components in acquiring both the whole and the part representations in a multi-layer structure? We will address this question in the next two subsections from perspectives of computational modeling and learning processes involved in the paradigm.

6.1. Computational extension of probabilistic chunking

Based on the computational framework of probabilistic chunking, we will use a simplified example to show how statistical coherence in the training data and prior preference favoring simple structural representations determine what is learned. Suppose we view a set of scenes, denoted as D, in which some scenes include a quadruple with four shapes EFGH and some other scenes include the embedded pair EG separately (i.e., in the absence of the other two shapes F and H). We compare two relevant structural representations S_1 and S_2, each of which has a two-layer structure, with chunks in the top layer and element shapes in the bottom layer: structure S_1 consisting of three chunks {EFGH, EG, FH}, and structure S_2 with one quadruple and one pair {EFGH, EG}. According to the Bayes rule, the posterior probability of a structure in light of the observed data can be defined as $P(S \mid D) \propto P(D \mid S)\, P(S)$. For the first term, the likelihood $P(D \mid S)$, Orban et al. (2008) demonstrated the Occam's-razor property of Bayesian models: complex structures with more chunks yield smaller likelihood. Intuitively, more complex

structures can generate a larger variety of possible scenes. This fact naturally leads to a smaller probability of generating the specific set of scenes observed in an experiment. As confirmed in our simulation (see Supplementary Material for details of the simulation), for the experiment paradigm used in the current study, although both structures can account for the observed set of scenes, S_1 structure with three chunks {EFGH, EG, FH} yields a smaller likelihood than does the simpler structure S_2 with two chunks {EFGH, EG}, see Fig. 10 left panel. The second term, the prior P(S), further penalizes complex structures by assigning smaller prior probabilities to them. For example, Orban et al. (2008) adopted geometric distributions as the prior to incorporate a preference favoring chunk inventories with fewer and smaller chunks. As shown in Fig. 10 left panel, S_1 structure with three chunks imparts a smaller prior probability than does the simpler structure S_2 with two chunks. Accordingly, if the structural form is constrained to a two-layer hierarchy with chunks in the top layer, S_1 structure that includes *both* the quadruple (EFGH) and the complementary part (FH) as chunks is unlikely to yield the highest posterior probability (i.e., the product of likelihood and prior) in light of the observed data D. The Bayesian analysis reflects a rational strategy to maintain the efficiency of visual representations for objects or multi-object scenes without adding numerous extra chunks that are redundant given the observed data.

However, when the structural representation can be extended in depth to include additional layers, other preferences can be included in the structural priors to express inductive biases about what kinds of multi-layer hierarchical structures are likely to occur in the world, and what constraints to incorporate in a prior distribution over structures.

Here we suggest to use a nonparametric distribution over tree structures, known as *the nested Chinese restaurant process* (nCRP) (Blei, Griffiths, & Jordan, 2010) to derive the prior distribution for a multi-layer hierarchy. The nCRP prior is flexible enough to accommodate different structures (e.g., different layers, different number of nodes in a tree structure) while also probabilistically favoring simpler structures that provide a parsimonious account of observed data. We chose the nCRP prior rather than other structural priors, such as the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2006), or the nested-IBP (Chien & Chang, 2014) for several reasons. First, the IBP is not applicable to the present experiment as this stochastic process is unable to generate multi-layer hierarchical structures. Second, although the nested-IBP can be applied to infer multi-layer structures, this prior process is less constrained in the sense that it allows a shape element to be included in multiple chunks. In contrast, the nCRP prior only allows a shape element to appear in one chunk at any level of the hierarchy. This "exclusive" constraint in nCRP limits the structural hypothesis space, which helps learning converge with relatively small training samples. For the current simulation, we define a three-layer tree structure consisted of EFGH as a whole chunk in the top layer, EG and FH as part chunks in the middle layer, and the individual shapes in the bottom layer (analogous to the H3 structure in Fig. 1). This structure complies with the "exclusive" constrain of the nCRP prior. We denote this multi-layer structure as S_3, which can account for the set of scenes D including some scenes with a shape pair EG, and some other scenes consisted of a quadruple EFGH.

Now, consider a competing two-layer structure S_2 consisting of one quadruple and one pair chunks {EFGH, EG}, in which a shape constituent can be used in two different chunks (i.e., E and G). As there is no nested tree structure involved in S_2, the prior for this structure can be the standard Chinese restaurant process (Aldous, 1985; Pitman, 1995), a distribution over partitions of objects into chunks. The prior for the three-layer tree structure S_3 can be greater than the prior for the two-layer structure S_2, as long as the parameter in nCRP that controls a penalty for adding deeper layers is not too large. Fig. 10 right panel shows the prior probability as a function of the parameter in nCRP, which controls the penalty of more layers in the structure. Hence, at the computational level with Bayesian inference, it is possible for the posterior probability for the three-layer nested tree structure S_3 to be higher than the posterior probability for the two-layer structure S_2, as long as the prior for S_3 can compensate for the differences in the likelihood term. At the algorithm level, using a sampling approach to exploit the structure space, the strong chunking for the trained embedded part (EG) in early learning likely provides a cue to guide the efficient sampling of structures. The empirical finding of better recognition for the complementary pair (FH) provides evidence that the three-layer structure is the most supported representation based on the statistical coherence of the constituents and prior preference for structural representations.

The above simulation provided evidence that the present findings are better explained using a multi-layer, hierarchical representation than by using a two-layer, feature-to-chunk representation. However, the simulation results do not necessarily imply that our human participants had built such representations of the complex chunks, or that the parts-beget-parts effect can *only* be explained by such representation. Still, our analysis suggests that, compared with a two-layer, chunk-to-feature model, a multi-layer hierarchical representation provides a better

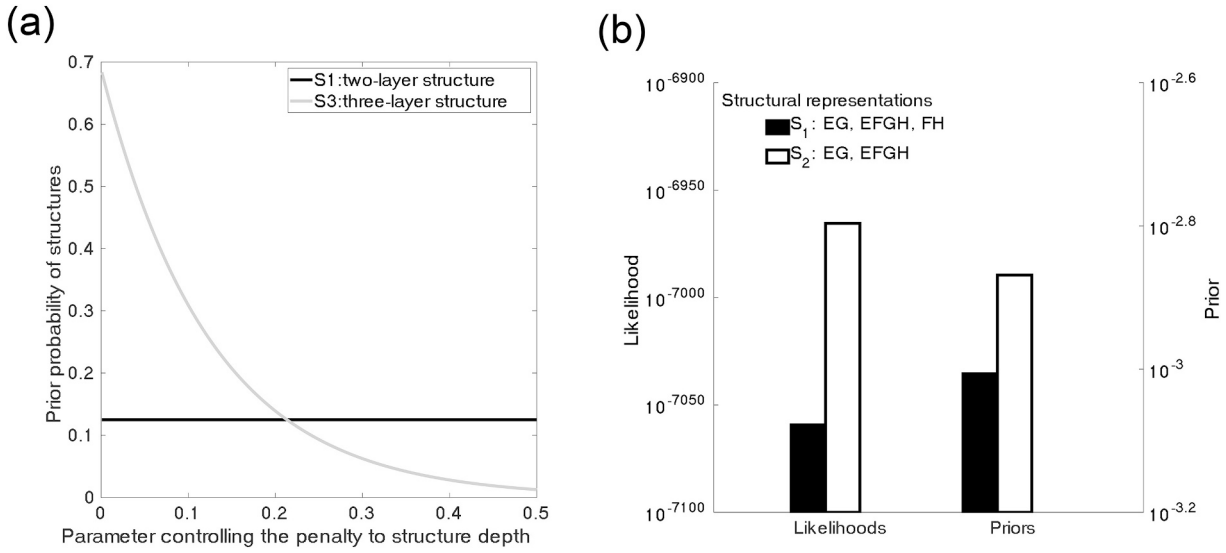

Fig. 10. Simulation results. (a): likelihood and prior probabilities for the two-layer structures (S_1 and S_2) according to the BCL model proposed by Orban et al. (2008). See Supplementary Material for details. (b): prior probability of two structures (S_2 and S_3) as a function of nCRP parameter that controlls the penalty to structure depth.

description of the underlying representation that can account for the present behavioral findings.

6.2. Learning processes

We present below possible learning processes involved in the learning context of the present study. We then discuss each process component in detail along with alternative interpretations and outstanding issues. First, in Phase 1 of Experiments 1–3, participants learned the trained embedded pair (EG) as an individual component, so that they formed a chunk representation of EG (component 1). Then, during Phase 2 when participants were exposed to training scenes containing the quadruple, the existence of the part representation EG may have enabled the formation of other part representations within EFGH (EFGH = EG + some other part(s); component 2). This facilitated formation of the representation for the complementary part (FH; component 3). Based on the part representations (EG and FH), the quadruple was then represented as a composition of these parts (EFGH = EG + FH; component 4).

For component 1 (learning of the trained embedded pair EG as a chunk), our findings demonstrate that this learning component can be achieved either via explicitly presenting the chunk to participants in a supervised manner (Experiments 1 and 2) or via passive exposure to implicitly learn the statistical regularities (Experiment 3). Previous studies suggest that there are many possible ways to acquire representations of parts within a complex chunk. Everyday objects often contain cues that allow observers to segregate meaningful parts from one another, such as the part boundaries arising from the minima of curvature (Hoffman & Richards, 1984), the ability for a part to move within the object (e.g., a hand rotating around the wrist), and, at a more conceptual level, the specific function that a part serves irrespective of its shape (e.g., support of a chair). It is possible that learning the trained embedded part via these means could also initiate the following processes that facilitate the learning of other parts.

For component 2 (supporting the formation of complementary pair as an individual component in representations), our findings suggest that this learning of embedded pair did not happen for the control quadruple ABCD, because participants showed near-chance discrimination in their familiarity judgments on the embedded pairs for the control quadruple ABCD. As mentioned above, previous studies have suggested other means for acquiring part representations (e.g., Fiser & Aslin, 2005). In the context of the present study, we discuss the facilitation of forming part representations specifically via extracting statistical regularities from visual scenes.

One question that concerns this component is the extent to which a complementary-part representation is favored. Although the present study found evidence for the learning of a complementary pair, we believe that this "complementarity rule" does not always apply, and probably depends on other factors. One such factor could be the number of low-level features that the complementary part contains. In the present study and previous visual statistical learning studies, visual features were operationalized as the shape elements. In our experiments, the complementary part contained only two shapes and took up half of the quadruple. If the quadruple contains many features and the trained embedded part contains only a small proportion, the remaining, complementary portion will be containing many features. In this situation, the embeddedness constraint (Fiser & Aslin, 2005) may still limit the complementary portion to be further grouped as a part by itself to avoid the curse of dimensionality.

Another factor that could affect the induction of the complementary-part representation is the relative amount of training between the part and the whole. It is possible that, if participants are only exposed to significantly more instances of the training scenes containing the whole complex chunk (EFGH), the facilitation effect from the pre-existing part representation EG can be overridden, so that the complex chunk would be represented as an inseparable whole without embedded parts. However, in daily life, objects are more likely viewed repeatedly with partial whole due to occlusion and viewpoint, more analogous to the learning situations in the present study that parts and wholes are intermixed in the learning data. The impact of the relative amount of training examples between the part and the whole is also revealed in the results of Experiment 3 and 4. In Experiment 3, the part representation (EG) was learned via passive exposure and, thus, could be weaker relative to the whole representation (EFGH), which contributes to the weakened parts-beget-parts effect in Experiment 3. When Experiment 4 significantly reduced the training examples for the part (EG), the parts-beget-parts effect was not revealed.

Component 3 (the formation of representation of the complementary part) involves two important subprocesses. The first is the grouping of the remaining elements in the complementary portion (F and H) into a unit (FH). This grouping of elements embedded within a whole chunk supports the idea that the embeddedness constraint is flexible (Fiser & Aslin, 2005), and the present study demonstrated another condition under which such flexibility would allow the formation of part representation. The second subprocess is the segmentation of the complementary portion (FH) from the trained embedded part (EG), so that it can be represented as a chunk by itself. We believe that participants learned the border between part EG and the complementary part FH, and successfully transfer the border ownership from EG to FH, enabling segmentation of EFGH (EFGH = E + G + F + H) that groups the lowest-level shape elements into two separate parts. Here, the "border" refers to the conceptual boundary that separates internally represented chunks. If there was no such border formed for FH, it would be possible that FH is represented as a background pattern but not an isolated chunk. It is important to emphasize that this border ownership by FH was never taught to the participants; rather, the establishment of such ownership by both F and H makes it possible for FH to be represented as a meaningful unit by itself. In their Experiment 5, Fiser and Aslin (2005) demonstrated that imbalanced conditional probabilities in occurrence among embedded elements could facilitate the "cutting" of parts. Findings from the present study expands on this by demonstrating that such facilitation on segmentation could happen based on imbalanced conditional probabilities between parts, which potentially operates at a part level instead of the element level.

Component 4 concerns the compositional representation of the quadruple (EFGH = EG + FH): was there a whole representation of the complex chunk EFGH. In Experiment 1, the above-chance familiarity with the whole quadruple (EFGH) could be explained by separate familiarity with the two embedded parts (EG and FH). If so, there would have been no need for the top-level representation of EFGH as a whole to present in the hierarchical structure. However, in Experiment 2, participants consistently chose the true whole EFGH as more familiar than a spatially swapped foil (FEHG) which used the same part constituents. These findings suggest that participants did not simply derive the familiarity with EFGH based on their familiarity with EG and FH. Instead, they learned the spatial configuration by putting together these two parts into a whole.

6.3. Future directions

There are specific findings that remain to be explored in future studies. In this section we provide interpretations of these findings, and suggest potential directions for future studies to pursue.

In Experiment 3, when the pre-exposed embedded pair was trained implicitly without a task, prior familiarity of this embedded pair was reduced. This weakened prior familiarity resulted in lower familiarity accuracy on the complementary embedded pair in the subsequent learning. In Experiment 4, when prior familiarity of an embedded part was eliminated by removing the pre-training familiarization, the complementary pair was no longer seen as more familiar than chance. This lack of familiarity with the complementary pair is consistent with previous findings in statistical learning for visual chunks (Fiser & Aslin,

2005) and pseudowords from syllables (Giroux & Rey, 2009). The absence of the parts-beget-parts effect in Experiment 4 may have been due to the weak learning of trained embedded pair, or to inhibition between learning of the whole and of its parts at the same time. The present study cannot tease apart these possibilities. Future studies should explore the degree of part familiarity needed to trigger generation of complex structure by bootstrapping based on familiar parts.

Nonetheless, one could argue that participants also learned the trained embedded pair during the training phase in Experiment 4. If so, it would imply that when the learning of the whole and of its part happens within the same training session (i.e., not separate in time), the complementary part cannot be learned. Therefore, instead of prior familiarity on a part, it might have been the time-separation between Phase 1 and Phase 2 that produced the parts-beget-parts effect.

An interesting empirical question that arises from this interpretation is whether the same parts-beget-parts effect would be observed if the temporal order of Phase 1 and Phase 2 were swapped. If participants had only been exposed to training scenes in Phase 2 of Experiments 1–3 in the present study, they should only form a whole, non-compositional representation of the target quadruple EFGH, as in Experiment 4 in the present study and in previous studies, e.g., Fiser and Aslin (2005). If participants then went through the training in Phase 1 (i.e., learning EG), would the visual system retrospectively turn the non-compositional representation of EFGH into a compositional one based on later-formed representation of EG? If so, this would suggest that the formation of a compositional representation is flexible and can be updated based on newly-acquired statistical evidence. Otherwise, it would suggest that timing is critical for the parts-beget-parts effect to be observed, as familiarity on a part needs to be acquired before the learning of the whole in order for the complementary part to be learned. These possibilities remain to be explored in future studies.

In a recent study, Plaut and Vande Velde (2017) demonstrated that human learning of wholes and parts through statistical learning can be modeled by learning in artificial neural networks. Their neural network models captured both the whole-part suppression (Fiser & Aslin, 2005; Giroux & Rey, 2009) and the facilitation of part representation by element predictability (Fiser & Aslin, 2005). In this context, our finding may pose a challenge to this model, because the enhanced familiarity on the complementary part could not be explained by element predictability. Future models of statistical learning need to consider prior familiarity of a subpart as an important factor modulating the likelihood of combining other subpart(s) to form intermediate-level part representations.

Furthermore, the generalizability of the parts-beget-parts effect remains to be explored. Specifically, despite their popularity in previous studies in visual statistical learning, the extent to the findings based on abstract shape elements can be generalized to other types of shapes (e.g., silhouettes, boundary outlines, or real-world objects) remains unclear. Although there has been evidence for visual statistical learning using other types of elements (e.g., Brady and Oliva (2008) used real-world scenes; Otsuka, Nishiyama, Nakahara, and Kawaguchi (2013) used everyday objects), future studies can explore the learning of parts and wholes using other types of elements. Also, similar to many previous statistical-learning studies on adults, we assessed learning using a familiarity-discrimination task. Although above-chance performance in such a task could imply the formation of a chunk representation, it is possible for an object part to be rated as familiar without being represented as a separate chunk. To further evaluate the robustness of the parts-beget-parts effect, future studies could explore other testing methods. Examples of such methods include reaction-time tasks (e.g., Turk-Browne et al., 2005) for measuring the flexibility and automaticity of the part representation and remember/know tasks (e.g., Batterlink et al., 2015) for measuring the awareness of the knowledge about the part representation.

7. Conclusion

In summary, one of the most remarkable aspects of perception is the acquisition of representations of novel objects to achieve flexible and adaptive object recognition (Tanaka & Farah, 1993; Ullman, 2007; Yuille, 2011). The present study provides evidence that compositionality in object representations can be learned from mere exposure via statistical learning, which promotes reuse of prior knowledge of parts that occurred in a different learning context. Critically, prior familiarity with one part of a complex object facilitated the formation of a hierarchical representation of that complex object based on an additional, novel part. These findings demonstrate the power of statistical learning to bootstrap the acquisition of novel parts in order to form compositional representations of objects. Future studies should explore other learning mechanisms fostering compositional object representations, as well as the interactions between different learning mechanisms.

Acknowledgment

This research was supported by NSF grant BCS-165530 to Hongjing Lu.

Appendix A. Supplementary data

Supplementary materials can be found online at [INSERT LINK TO SUPPLEMENTARY MATERIALS]. Data of individual experiments are available online on OSF at https://osf.io/bj5d2/. Each data file (one file per experiment) contains responses form individual participants in all tasks described in the present study. Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104515.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *Lecture Notes in Mathematics École D'Été De Probabilités De Saint-Flour XIII — 1983* (pp. 1–198). https://doi.org/10.1007/bfb0099421.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of memory and language, 83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM, 57* (2), 1–30. https://doi.org/10.1145/1667053.1667056.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes. *Psychological Science, 19*(7), 678–685. https://doi.org/10.1111/j.1467-9280.2008.02142.x.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Chien, J.-T., & Chang, Y.-L. (2014). The nested Indian buffet process for flexible topic modeling. In *Proc. Annu. Conf. Int. Speech Commun. Assoc.* (pp. 1434–1437).

Fidler, S., Berginc, G., & Leonardis, A. (2006). Hierarchical statistical learning of generic parts of object structure. In *Paper presented at the computer vision and pattern recognition, 2006 IEEE computer society conference*.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*(6), 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, USA, 99*(24), 15822–15826.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General, 134*(4), 521.

Froyen, V., Feldman, J., & Singh, M. (2015). (2018). Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review, 122*(4), 575–597. https://doi.org/10.1037/a0039540.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science, 33*, 260–272.

Griffiths, T., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *, 18. Advances in neural information processing systems* (pp. 475–482). Cambridge, MA: MIT Press.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.

Helmholtz, H.v. (1866/1924). *A treatise on physiological optics. 1*. New York: Optical Society of America.

Hentschke, H., & Stüttgen, M. C. (2011). Computation of measures of effect size for neuroscience data sets. *European Journal of Neuroscience, 34*, 1887–1894.

Hoffman, D., & Richards, W. (1984). Parts of recognition. *Cognition, 18*(1–3), 65–96. https://doi.org/10.1016/0010-0277(84)90022-2.

Kersten, D., Mammassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.
Lowe, D. G. (1985). *Perceptual organization and visual recognition*. Norwell, MA, USA: Kluwer Academic Publishers.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.
Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127–162.
Orban, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, USA, 105*(7), 2745–2750.
Otsuka, S., Nishiyama, M., Nakahara, F., & Kawaguchi, J. (2013). Visual statistical learning based on the perceptual and semantic information of objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 196–207. https://doi.org/10.1037/a0028645.
Pelli, D. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.
Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields, 102*(2), 145–158. https://doi.org/10.1007/bf01213386.
Plaut, D. C., & Vande Velde, A. K. (2017). Statistical learning of parts and wholes: A neural network approach. *Journal of Experimental Psychology: General, 146*(3), 318–336.
Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.
Singh, M., & Hoffman, D. D. (2001). Part-based representations of visual shape and implications for visual cognition. In *, 130. Advances in psychology* (pp. 401–459) (North-Holland).
Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology. A, 46*(2), 225–245.
Tu, Z., & Zhu, S. (2002). Image segmentation by data-driven markov chain Monte Carlo. *IEEE Tranactions on Pattern Analysis and Machine Intelligence, 24*(5), 657–673.
Turk-Browne, N. B., Junge, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552–564.
Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113*(2), 169–193.
Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences, 11*(2), 58–64. https://doi.org/10.1016/j.tics.2006.11.009.
Yuille, A. (2011). Towards a theory of compositional learning and encoding of objects. In *Paper presented at the computer vision workshops (ICCV workshops), 2011 IEEE international conference*.
Yuille, A., & Mottaghi, R. (2016). Complexity of representation and inference in compositional models with part sharing. *Journal of Machine Learning Research, 17*(1), 292–319.