Journal of Experimental Psychology: General
2017, Vol. 146, No. 3, 318–336

0096-3445/17/$12.00 http://dx.doi.org/10.1037/xge0000262

Statistical Learning of Parts and Wholes: A Neural Network Approach

David C. Plaut and Anna K. Vande Velde
Carnegie Mellon University

Statistical learning is often considered to be a means of discovering the units of perception, such as words and objects, and representing them as explicit "chunks." However, entities are not undifferentiated wholes but often contain parts that contribute systematically to their meanings. Studies of incidental auditory or visual statistical learning suggest that, as participants learn about wholes they become insensitive to parts embedded within them, but this seems difficult to reconcile with a broad range of findings in which parts and wholes work together to contribute to behavior. Bayesian approaches provide a principled description of how parts and wholes can contribute simultaneously to performance, but are generally not intended to model the computations that actually give rise to this performance. In the current work, we develop an account based on learning in artificial neural networks in which the representation of parts and wholes is a matter of degree, and the extent to which they cooperate or compete arises naturally through incidental learning. We show that the approach accounts for a wide range of findings concerning the relationship between parts and wholes in auditory and visual statistical learning, including some findings previously thought to be problematic for neural network approaches.

Keywords: statistical learning, neural networks, parts and wholes, connectionist modeling

Learners of all ages are highly sensitive to the statistical structure of their experience in the world. Early empirical studies of statistical learning (Saffran, Aslin, & Newport, 1996) focused on sensitivity to transitional probabilities within a stream of auditory syllables. Subsequent work extended this to a broad range of types of distributional information within the auditory (e.g., Maye, Werker, & Gerken, 2002; Saffran & Griepentrog, 2001), visual (e.g., Fiser & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002), and even tactile (Conway & Christiansen, 2005) modalities (for reviews, see Aslin & Newport, 2012; Smith, Suanda, & Yu, 2014; Thiessen, Kronstein, & Hufnagle, 2013).

Much of this work has been concerned with how learners discover the relevant units of perception. Thus, in a typical auditory experiment (e.g., Saffran et al., 1996), the syllable stream might be composed of randomly ordered "words," each consisting of two or three syllables that always occur together in a fixed order. In the visual modality (e.g., Fiser & Aslin, 2001), a display might consist of one or more "objects," each composed of a set of contiguous elements in a fixed spatial arrangement, possibly co-occurring with distracting elements. Successful learning of the words or objects is demonstrated by a preference for these structures (or against them, in the case of an infant novelty bias) when paired with random combinations of syllables or elements. Indeed, in each modality, the structures acquired through statistical learning of this sort have been shown to facilitate subsequent word learning (Graf Estes, Evans, Alibali, & Saffran, 2007; Hay, Pelucchi, Graf Estes, & Saffran, 2011) or object processing (Zhao, Cosman, Vatterott, Gupta, & Vecera, 2014).

Although early accounts of statistical learning attempted to distinguish segmentation via clustering versus bracketing (Goodsitt, Morgan, & Kuhl, 1993; Swingley, 2005), most recent formulations combine multiple cues to segmentation in the service of *chunking* (Frank, Goldwater, Griffiths, & Tenenbaum, 2010; French, Addyman, & Mareschal, 2011; Kibbe & Feigenson, 2016; Otsuka, Koch, & Saiki, 2016; Perruchet & Pacton, 2006; Perruchet & Vinter, 1998; Thiessen et al., 2013)—that is, combining multiple smaller scale units (e.g., syllables, visual elements) into an explicit representation of a larger scale unit (e.g., word, object). Although the various approaches differ in specifics, they all involve a discrete decision as to whether a given bit of the input in a given context is or isn't treated as a chunk.

Certain issues arise with the notion of explicit chunking when one recognizes that entities such as words and objects are not undifferentiated wholes, but are commonly composed of parts which themselves can be treated as wholes in other contexts. Thus, many words (e.g., UNTEACHABLE) are composed of morphemes (UN- TEACH, -ABLE) that contribute in systematic ways to their meaning. Similarly, many objects (e.g., a bicycle) are composed of parts (wheels, seat, handlebars, etc.) that contribute in systematic ways to their properties or functions. The question is, how are the parts of complex wholes treated during learning? When a chunk is formed for the whole, is this chunk added to the chunks for its parts, or does the larger scale chunk for the whole replace or supersede the smaller-scale chunks?

Fiser and Aslin (2005) examined this issue in a series of studies in the visual modality (see also Fiser & Aslin, 2001; Jun & Chong, 2016). In their first experiment, participants were exposed to a series of visual displays, each of which consisted of a contiguous

This article was published Online First January 12, 2017.

David C. Plaut and Anna K. Vande Velde, Department of Psychology, Carnegie Mellon University.

We thank Marlene Behrmann, Erik Thiessen, Jay McClelland, members of the VisCog research group, and reviewers for helpful discussion and comments.

Correspondence concerning this article should be addressed to David C. Plaut, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890. E-mail: plaut@cmu.edu

cluster of six distinct visual elements positioned in a 5 × 5 array. Unbeknownst to participants, clusters were created by combining two out of four possible "objects," each of which consisted of a fixed arrangement of three specific elements. When later tested, participants preferred each object to random triples of elements. By contrast, participants showed no preference for element pairs embedded in objects relative to random element pairs, even though the embedded pairs occurred as often as the triples, suggesting that the chunks for wholes supersede those for their parts. The fact that "not all the embedded features that are parts of a larger whole are explicitly represented once a representation of the whole has been consolidated" was referred to by Fiser and Aslin (2005, p. 532) as the *embeddedness constraint* of statistical learning, and was viewed as important for reducing the computational complexity of deriving efficient perceptual organizations of scenes.

Giroux and Rey (2009) carried out an analogous study in the auditory modality. Participants were exposed to syllable sequences composed of randomly ordered tri- and disyllabic "words" (ABC, DEF, GH, IJ, KL and MN, where letters designate distinct syllables; none of the stimuli were actual words) for either two or 10 min. They then tested participants' preference for the disyllabic words (e.g., GH) and for pairs embedded in the trisyllabic words (e.g., AB)—which were matched to the disyllabic words on frequency and transition probabilities—relative to "partword" pairs constructed from the end of one word and the beginning of another (e.g., HI). After two minutes of exposure, participants showed a mild and equivalent preference for both the disyllabic words and for embedded pairs relative to the partword pairs. With extended exposure, however, only the preference for disyllabic words continued to strengthen, suggesting that acquisition of the trisyllabic words prevented further strengthening of pairs embedded in those words. Giroux and Rey showed that an explicit chunking mechanism in which larger chunks supersede smaller ones (PARSER; Perruchet & Vinter, 1998) replicated the results. By contrast, a simple recurrent network (SRN; Elman, 1990) trained on syllable prediction treated embedded and nonembedded pairs equivalently, thereby failing to match the empirical findings.

The idea that the representation of a whole inhibits the representation of its parts would seem at odds with a broad range of empirical findings indicating that, in many contexts, parts and wholes work together in perception (e.g., morphological priming, Marslen-Wilson, Tyler, Waksler, & Older, 1994; Rastle & Davis, 2008; Schreuder & Baayen, 1995; Taft, 1994; and the word superiority effect, Reicher, 1969; Wheeler, 1970). The question then becomes, how could an approach in which both parts and wholes contribute simultaneously to perception explain the above findings that participants fail to exhibit sensitivity to parts when embedded in more complex wholes?

In considering their findings, Fiser and Aslin (2005) articulated a number of key insights that ultimately lead to a more graded, context-dependent view of segmentation (see also Hoffman & Singh, 1997). Indeed, some of their later experimental results revealed a mixture of influences of parts and wholes, consistent with a more graded account (see also Baker, Olson, & Behrmann, 2004; Turk-Browne, Isola, Scholl, & Treat, 2008). Fiser and Aslin proposed that the representations of parts within wholes—and of "features" at every scale more generally—could be a matter of degree, as a function of their relative predictability (although the exact nature of a partial representation was not specified). Moreover, this predictability, and hence the resulting organization, might vary as a function of context (e.g., other parts and wholes in the scene). Note that, on this view, the implications of embeddedness are not so much a constraint on statistical learning as a consequence of it.

Fiser and Aslin (2005) offered a simple Bayesian formulation of statistical learning that is consistent with this graded view of segmentation (see also Feldman & Singh, 2006; Frank et al., 2010; Froyen, Feldman, & Singh, 2015). Orbán, Fiser, Aslin, and Lengyel (2008) later developed the Bayesian perspective more fully in the form of the Bayesian chunk learner (BCL). The BCL determines the likelihoods of all possible chunk inventories given the entire collection of displays as well as prior probabilities on parameters concerning the number and sizes of chunks in an inventory. It then calculates the posterior probabilities of various test displays by marginalizing over the posterior distribution of both inventories and parameters. Thus, all possible wholes and parts contribute to performance, combined in a way that depends on the statistical structure to which the system is exposed and priors concerning what sorts of structure to prefer. Orbán and colleagues showed that, with an appropriate set of spatially constrained priors, the resulting posterior probabilities mirror participants' preferences in a wide range of studies (including some from Fiser & Aslin, 2005).

Despite the clear advance that a Bayesian perspective on segmentation offers relative to all-or-none formulations of chunking, it does not provide a full, mechanistic account of the performance of participants in statistical learning tasks. The BCL does not actually assign a specific representation to a given scene that could participate in other cognitive processes. It is not a *process model*—the computations it uses to calculate probabilities, which involve iterating over all displays and all possible chunk inventories, are not thought to be carried out by participants in any direct or literal sense. Rather, the BCL can be understood as a means of scoring test displays based on their consistency with priors and experience, without attempting to approximate the actual representations, processes, and learning that give rise to the corresponding choice preferences in participants. Moreover, concerns about computational complexity are unresolved, particularly if embeddedness serves only to downweight potential parts but not eliminate them from consideration entirely.

In the current work, we develop a computationally efficient process model of statistical learning of parts and wholes. We adopt an approach, based on learning in neural networks, that is capable of capturing statistical structure at multiple levels of representation simultaneously and yet eschews the notion of explicit chunking. Through the use of learned distributed representations, there is no notion of discrete "units" of perception; rather, the extent to which a particular subset of the input in a particular context is represented in a coherent manner is a matter of degree, and the extent to which learned structure at one level of analysis cooperates or competes with learned structure at other levels is not prespecified but arises naturally as a consequence of incidental learning in the domain (see Gonnerman, Seidenberg, & Andersen, 2007; Plaut & Gonnerman, 2000; Seidenberg & Gonnerman, 2000, for a similar perspective applied to derivational morphology). Our approach is, of course, closely related to other efforts to apply neural networks to statistical learning (Christiansen, Allen, & Seidenberg, 1998; Cleeremans & McClelland, 1991; Dominey & Ramus, 2000; French et al., 2011; Mirman, Graf-Estes, & Magnus, 2010; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Sirois, Buckingham, & Shultz, 2000) although, as we

discuss at various points, we adopt somewhat different assumptions about how such networks are applied in this domain.

The current work focuses largely on visual statistical learning, as issues concerning the representation of parts and wholes have been explored in the greatest depth in this context. We begin, though, by modeling Giroux and Rey's (2009) findings in the auditory modality, in part to establish the generality of the relevant computational principles. We then show how an analogous model of visual processing can account for the findings from Fiser and Aslin's (2005) series of studies, along with those of a critical follow-up study (Orbán et al., 2008), concerning the treatment of parts and wholes in visual statistical learning. A final simulation establishes the simultaneous operation of parts and wholes in the network and serves to generate predictions for future empirical studies. We conclude by considering relations to other approaches, limitations of our work, and important directions for future research.

Simulation 1: Giroux and Rey (2009)

As mentioned in the introductory paragraphs, Giroux and Rey (2009) exposed adult participants to an auditory stream composed of two trisyllabic "words" (here denoted ABC and DEF) and four disyllabic words (GH, IJ, KL, and MN), presented in random order for either two or 10 min (corresponding to 400 or 2,000 syllables in total) with no pauses or other acoustic cues to word boundaries. Following this exposure, participants were presented with each of the disyllabic words and the embedded pairs within the trisyllabic words, paired with a partword pair which straddled a boundary between words during exposure. Importantly, the words and embedded pairs were matched on frequency of occurrence, and both types of pairs had a transitional probability of 1. As shown in Figure 1a, when asked to select the syllable pair that sounded more similar to what they had listened to in the exposure period, participants showed a small and equivalent preference for the disyllabic words and the embedded pairs over partwords. After 10 min of exposure, however, the preference for the disyllabic words had further strengthened but the preference for embedded pairs had not.

Giroux and Rey (2009) also trained an SRN to predict the next syllable in an equivalent stream of 400 or 2,000 syllable presentations. They then calculated the network's prediction error from the first to second syllable of disyllabic words (e.g., G→H), embedded pairs (e.g., A→B), and partwords (e.g., H→I), assuming that lower relative error implies greater preference. Unlike participants, the network's preferences for embedded pairs and disyllabic words (relative to partwords) behaved identically, increasing the same amount from 400 to 2,000 syllable presentations. Thus, the network's performance reflected the equivalent frequency and transition probabilities of these pair types but was insensitive to the larger trisyllabic words in which the former were embedded.

However, when comparing the performance of participants with that of the network, it may be that the absence of information is as important to consider as its presence. That is, the silence that both precedes and follows test stimuli may be relevant to participants' judgments. For example, having been exposed to a trisyllabic word like ABC, the presence of silence (instead of C) after AB, and the onset of B from silence (instead of from A) for BC, may reduce participants' sense of familiarity of these embedded pairs. By contrast, the presence of silence around disyllabic words is not disruptive because expectations at their boundaries are far weaker. The first simulation shows that, when silence precedes each test stimulus and the familiarity measure includes the prediction of silence following the stimulus, an SRN's performance is a much closer match to that of participants.

Method

An SRN was trained to predict the next syllable in a sequence of 2,000 syllables, constructed exactly as in Giroux and Rey's (2009) empirical study and SRN simulation (as just described). The simulation was run using the Lens neural network simulator developed by Doug Rhode (http://tedlab.mit.edu/~dr/Lens/). All training and testing files for this and subsequent simulations are available for download at http://www.cnbc.cmu.edu/~plaut/PlautVandeVelde/.

The network had 14 input and output units (one for each syllable) and 30 hidden and context units, with the input and context units fully connected to the hidden units which, in turn, were fully connected to

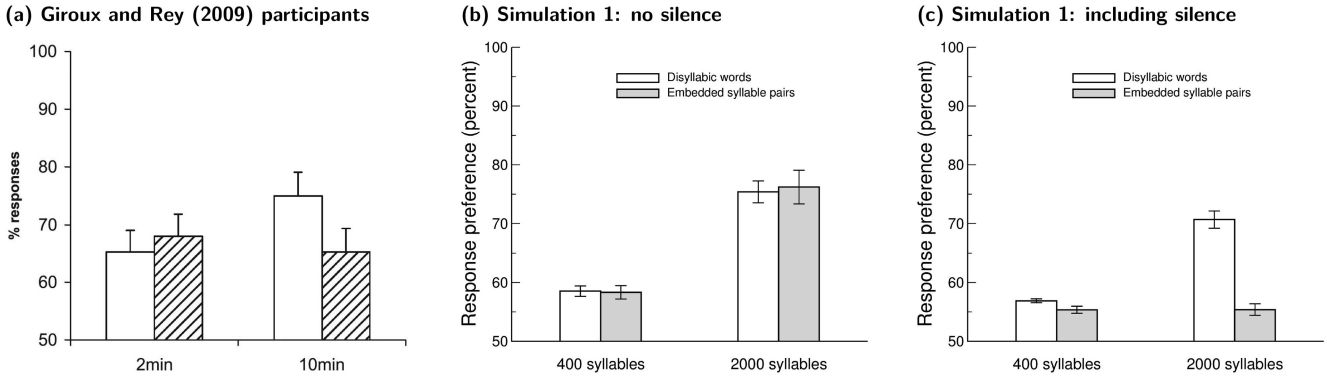

Figure 1. (a) Percentage responses to disyllabic words (open bars) and embedded syllable pairs (hashed bars) made by participants in Giroux and Rey's (2009) experiment, (b) analogous results from Simulation 1 when silence is not included during testing, and (c) Simulation 1 results when including silence during testing. Adapted with permission.

the output units. (Note that, by using a unique unit for each syllable, we are assuming that the syllables are equally discriminable to participants and that their relative similarities do not influence performance.) Hidden and output units also had a bias connection (from a unit with fixed activation of 1.0) and used a sigmoid activation function. Initial weights were set to random values sampled uniformly between ±1 except for output unit biases, which were initialized to –3.0. Results were calculated over 20 samples of random initial weights. As is standard in an SRN, hidden activations were copied to the context units prior to the presentation of each input in the sequence. Each syllable was presented to the network by setting the activation of its corresponding unit to 1.0 and all others at that position to 0.0, and then adding Gaussian noise ($SD = 0.3$) to these activations. The network was trained with back-propagation (Rumelhart, Hinton, & Williams, 1986) using momentum descent in cross-entropy error (Hinton, 1991), with a learning rate of 0.05, momentum of 0.9, batch size of 1 (weight updates after each syllable), and a bound of 1.0 on the magnitude of the premomentum weight step vector.

The network was tested after 400 and 2,000 syllable presentations. Testing involved presenting a series of two-syllable inputs, each preceded and followed by silence (i.e., an input of all zeros). Performance on each test stimulus was measured by summing the error produced when the first syllable predicted the second and when the second predicted silence (targets of 0.0 for all output units). This error was averaged separately over all disyllabic words (GH, IJ, KL, MN), over all embedded pairs within trisyllabic words (AB, BC, DE, EF), and over all partwords formed by combining a word-final syllable (C, F, H, J, L, N) with a word-initial syllable (A, D, G, I, K, M).[1]

We assume that participants (and the network) prefer a stimulus that has a better internal representation, which can be operationalized as one which produces lower error. Hence, the network's preference for words over partwords was computed as $E_p/(E_w + E_p)$, where E_p is the mean error on partwords and E_w is the mean error on words. The preference for embedded words over partwords was computed analogously.

For comparison purposes, and to replicate Giroux and Rey's (2009) simulation results, we carried out the identical simulation except without including silence before each test stimulus and without including the prediction of silence at the end.

Results and Discussion

Figure 1b shows the network's relative preferences for disyllabic words and for embedded pairs, compared with partwords, after 400 and 2,000 syllable presentations, when not including silence during testing. As Giroux and Rey (2009) found, and unlike participants, the network becomes increasingly sensitive to the stronger transition probabilities within both words and embedded pairs irrespective of whether or not the pair occurs within a larger whole.

Figure 1c shows the equivalent data under the condition in which testing included silence. Here, as is true of participants (Figure 1a), the network shows a slight and nearly equivalent preference for the words and embedded pairs after 400 syllables. By 2,000 syllable presentations, the preference for words has grown much stronger, whereas the preference for embedded words remains unchanged.

The initial increase in preference for words and embedded pairs is, of course, due to their much higher transition probabilities (1.0) compared with partwords (0.167). Performance on words continues to improve as the network becomes increasingly sensitive to transition probabilities in making predictions. Thus, when testing disyllabic words, the correct second syllable is activated moderately well (mean 0.417, error 2.04), and most units are relatively inactive when compared with silence following the second syllable (mean 0.053, error 0.886), leading to relatively low total error ($M = 2.92$). However, as the network's preference for trisyllabic words (ABC and DEC) relative to random triples increases—from 58.7% after 400 syllables to 79.7% after 2000 syllables—performance on embedded pairs suffers, although this manifests differently for initial (e.g., AB) versus final (e.g., BC) pairs. For initial pairs, as in disyllabic words, the second syllable is appropriately activated (mean 0.405, error 2.08), but the third syllable in the triple (e.g., C) is strongly but incorrectly activated (mean 0.684, error 2.23) when compared with silence, leading to high total error (mean 4.31). For final pairs, the preceding silence disrupts activation of the second syllable from the first (mean 0.104, error 3.82), but activation is relatively low when predicting silence (mean 0.070, error 1.11), also resulting in high error (mean 4.93).[2] By comparison, partwords suffer from both of these effects, and have a mean total error of 6.10. Thus, relative to partwords, the network shows a much stronger preference for disyllabic words over embedded pairs.

There might be some concern that the results are biased by the fact that the network was never exposed to silence during training. To evaluate this possibility, we replicated the simulation but added silence after 30 randomly selected occurrences of each of the six words during training (180 occurrences, amounting to 9% of all syllable presentations)—note that we are assuming that silence never occurs in the middle of words. That is, 180 partword transitions in the training corpus, such as C→D, were replaced by two transitions, C→silence and silence→D, where silence was represented by inputs or targets of all zeros. When tested in the same way as above (including silence), the network's preference for disyllabic words increased substantially, from 56.5% ($SE = 0.45$) after 400 syllables, to 77.9% ($SE = 1.41$) after 2,000 syllables. By contrast, the networks preference for embedded pairs held nearly constant, from 56.0% ($SE = 0.56$) to 58.9% ($SE = 1.32$). Thus, the network replicated the original findings. Because silence is relatively rare and unpredictable, its inclusion during training has little impact on the network's preferences.

In summary, Giroux and Rey (2009) found empirically that learning trisyllabic words prevented the learning of the two-syllable pairs embedded within them. They interpreted these findings as inconsistent with an SRN-based account, and instead took them to implicate an explicit chunking mechanism, such as the one implemented in PARSER (Perruchet & Vinter, 1998), in which learning a larger chunk blocks learning of smaller ones. Our modeling results show that Giroux and Rey (2009) were premature

[1] Giroux and Rey (2009) evaluated words and embedded pairs against particular sets of partwords, but we used all of them to provide a more reliable estimate of error.

[2] The specific choice of error function influences the exact numeric relationship between error on initial versus final embedded pairs, but not the fact that each produces more error than disyllabic words.

in rejecting SRNs as capable of providing an account of their findings and, moreover, that the findings do not implicate an explicit chunking mechanism in which the representations of wholes (trisyllabic words) supersede the representations of their parts (embedded syllable pairs).

Simulation 2: Fiser and Aslin (2005) Experiment 1

Although Giroux and Rey's (2009) results can be explained by a standard SRN trained on syllable prediction, this shouldn't be taken to imply that such a model is the most appropriate way to account for language learning more generally. SRNs are often viewed as a specific type of neural network that is distinct from other types that have been applied to developmental data (see Yermolayeva & Rakison, 2014), and some researchers consider the application of SRNs to tasks other than prediction to be "a fundamental change in the way in which SRNs are conceptualized" (French et al., 2011, p. 621). We think these views are too narrow. Rather, an SRN is more appropriately understood as a convenient simplification of a fully recurrent network with potentially unrestricted connectivity, and, as such, can be applied to the full range of cognitive tasks involving information about the past (maintenance), present (reconstruction), and future (prediction).

Similarly, *autoencoders*—networks that copy inputs to outputs via one or more layers of learned, internal representations (Ackley, Hinton, & Sejnowski, 1985; Mareschal, French, & Quinn, 2000)—are also sometimes considered to be a distinct class of network, but again are better thought of as a computational simplification of a fully recurrent network in which hidden units both receive input from, and reconstruct input over, the same lower level representation (see LeCun, Bengio, & Hinton, 2015). Thus, although we adopt a different network architecture in shifting from auditory statistical learning (SRN) to visual statistical learning (autoencoder), we view these as two approximations of the same underlying computational approach. Indeed, we could have used an SRN version of our autoencoder by including context layers for each of its hidden layers, but doing so would be irrelevant in the current context as the training environments we consider have no temporal dependencies for the network to learn.

In their first experiment, Fiser and Aslin (2005) carried out what is in many ways the visual analog of Giroux and Rey's (2009) auditory experiment. As summarized in the in the introductory paragraphs, Fiser and Aslin constructed four objects (termed *base-triplets* in Figure 2), each consisting of a fixed, contiguous spatial arrangement of three elements. They then created 112 six-element visual displays by combining pairs of these objects into contiguous arrangements at various positions within a 5 × 5 grid, an example of which is shown in the center of Figure 2 (see Barenholtz & Tarr, 2011, for evidence of the importance of contiguity). Participants viewed an 11 min movie in which each of these displays was presented twice, in random order, for 2 s with a 1-s interstimulus interval.

In the testing phase, participants were presented with a sequence of two displays containing either single elements, pairs, triples, or quadruples presented in the center of each grid, and asked which of the two seemed more familiar based on the earlier exposure phase. There were two critical comparisons: (a) the base objects were compared against random contiguous triples of elements that had never appeared in that spatial configuration in the exposure phase, and (b) contiguous pairs of elements that were embedded in the base triples were compared against random contiguous pairs that also had never previously appeared in that spatial configuration. The results of these comparisons are shown in Figure 3a. Following exposure, participants showed a clear preference for the base triples over random triples, but showed no reliable preference for element pairs embedded in those triples over random pairs, even though the embedded pairs occurred as often as the base triples and with the same conditional probabilities among their elements (1.0). These results, like those of Giroux and Rey (2009), suggest that learning larger scale structures somehow impedes the learning of smaller-scale structures within them. The current simulation tests whether a network trained in an analogous manner shows the same pattern of effects.

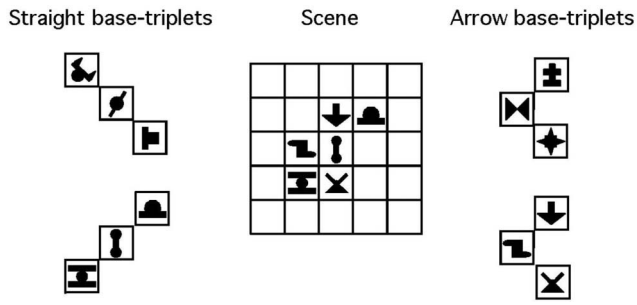

Figure 2. Base objects and an example display from Fiser and Aslin's (2005) Experiment 1. Adapted with permission.

Method

The network architecture, parameters, and training procedure used in the current simulation were also used in all of the remaining simulations reported in this article, with the exception of details of stimulus construction and testing procedures that are specific to each experiment.

The network was a feedforward autoencoder with 300 input units (12 elements at each position in a 5 × 5 array) that were fully connected to a first layer of 80 hidden units, which in turn were fully connected to a second layer of 40 hidden units, which were fully connected to 300 output units.[3] All hidden and output units also had bias connections. Weights were initialized to small random values (sampled uniformly between ±0.5) except for the biases of output units, which were initialized to –3.0. Results were calculated over 20 instances of random initial weights. An element was presented at a particular position by setting the activation of its corresponding unit to 1.0 and all others at that position to 0.0, and then adding Gaussian noise ($SD = 0.15$) to these activations. Empty positions had all activations set to 0.0. The network was trained with back-propagation using momentum descent in cross-entropy error, with a learning rate of 1.0, momentum of 0.9, batch size equal to the number of examples, and a bound of 1.0 on the magnitude of the pre-momentum weight step vector.

[3] Analogous to the auditory simulation, by using a unique unit for each shape element at each position, we are assuming that each element can be easily discriminated from all the others at each position, and that the relative similarities among their retinotopic visual representations are not relevant to accounting for participants' performance.

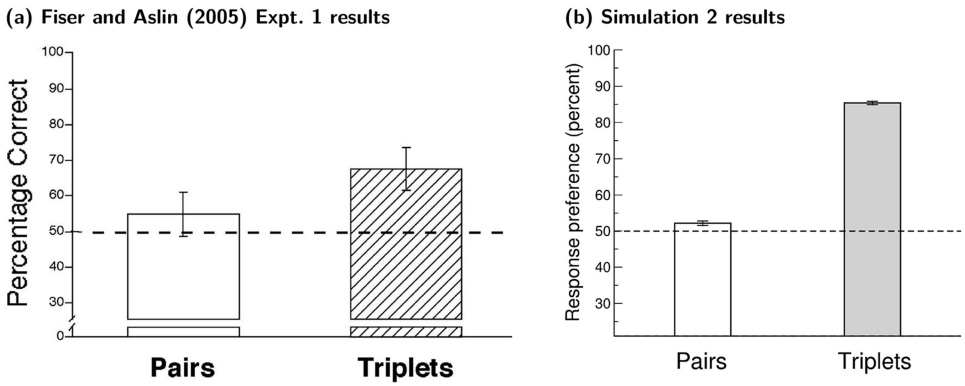

Figure 3. (a) Response preferences of participants in Fiser and Aslin's (2005) Experiment 1 for embedded versus random pairs ("pairs") and for base objects versus random triples ("triplets"); (b) analogous response preferences in Simulation 2. Adapted with permission.

For the current simulation, the network was trained on stimuli analogous to the visual displays used in Experiment 1 of Fiser and Aslin (2005). An issue immediately arises, however, with regard to how to handle positional variation. Participants come to the experiment with a visual system that is largely invariant with respect to the retinal position of stimuli, at least within the range of central vision used in the experiment. The network, by contrast, comes with no such invariance—to the network, shifting a configuration by a single position in any direction generates an entirely nonoverlapping input (apart from effects of noise).

One way of rectifying this discrepancy is to structure and pretrain the network such that it exhibits the requisite positional invariance prior to undergoing experimental training. For instance, convolutional neural networks (CNNs; Fukushima, 1980; LeCun et al., 1989; Riesenhuber & Poggio, 1999) employ multiple banks of units in each layer, such that all of the units within each bank detect the same feature (i.e., have the same incoming weights) but at different spatial positions. Typically, a "pooling" layer follows each convolutional layer in which a given unit is active if any of the equivalent feature units over a range of positions is active. Such networks are typically applied to recognition tasks known to be invariant to absolute image position.

We chose not to use a CNN in the current work for a number of reasons. First, for an autoencoder, the increasing positional invariance across successive layers in the network makes it problematic to reconstruct the position-specific aspects of the input. Second, CNNs achieve positional invariance by having identical receptive fields tiled across the input (and across each successive topographic layer), but the specified size of these receptive fields introduces a bias favoring structure at the corresponding spatial scale (relative to larger or smaller scales). And finally, a CNN would require extensive pretraining to achieve pre-experimental positional invariance, it is difficult to formulate a pretraining environment that does not incidentally bias the network toward learning certain types of structure in the experimental stimuli at the expense of other types.

Instead, we adopted an approach that yields positionally invariant performance and requires training only on the experimental stimuli, but that sacrifices an exact correspondence between the nature of exposure for participants and for the network. Specifically, during training, whenever a particular configuration is presented to the network, it is presented at each of the 25 positions, wrapping horizontally and/or vertically when necessary in order to avoid edge effects. Edge effects are problematic because they give rise to unbalanced frequencies with which certain embedded combinations occur across positions, and it is not possible to construct a training regime that is unbiased with respect to all relevant position-specific frequencies. Wrapping allows every subset of every configuration to occur at every spatial position equally often, thereby avoiding any possibility of positional biases in the results. We think of the training regime as an integration of pre-experimental visual experience that induces position invariance and experimental experience with particular types of statistical structure. Pretraining with non-experimental stimuli would, as just discussed, inappropriately bias the network toward learning some types of structure over others.[4]

During testing, we present each configuration at a single, central position, as was true for participants. Because of this, most wrapped presentations during training are not directly relevant as those positions are never involved in testing. Even so, they are included to ensure that no positional biases were introduced during training. We will evaluate a version of the simulation in which stimuli are not wrapped during training to determine the degree to which the results depend on wrapping per se.

As was true for participants, display configurations during initial exposure consisted of spatial configurations of two of the four base

[4] An alternative approach would be to assume participants make multiple fixations of each display, and to approximate this by positioning each configuration such that each of its elements falls in the center of the display. We explored this approach but ultimately decided against it. It leads to the same general patterns of performance as found in the current work (and empirically), but the effects are much weaker. The reason is that the limited fixations give rise to too much idiosyncrasy in how the relevant configurations are positioned during training, and this waters down the extent to which the network learns their (sub)structure. Note that this is not an issue of bias in positioning during testing, as all potential components (with the same shape) are positioned equivalently during both training and testing. Rather, representations need to be largely positionally invariant in order to give rise to the observed effects, and fixated training of the experimental stimuli is, by itself, insufficient to induce this.

objects in which at least two pairs of elements are adjacent horizontally (n = 28; see Fiser & Aslin, 2005, for details). The network was trained for 100 epochs in which each of the 28 possible configurations was presented at each spatial position.[5] Although this amounts to much greater exposure to each configuration than participants received, the training also serves as a proxy for pre-experimental visual experience, and much of it is only indirectly relevant as it involves positions that are not involved in testing. Even so, we recognize that the lack of intrinsic spatial invariance in the network means that we cannot draw close comparisons between the amount of training for the network and for participants. We will, at least, hold the amount of training (100 epochs) constant across simulations, as the amount of exposure given to participants across experiments was approximately equal.

Participants were tested for forced-choice preferences among singletons, pairs, triples, and quadruples in order not to draw their attention to a particular level of structure. As that is not an issue for the network, we restrict testing of the network to just the two critical experimental comparisons: base triples against random triples, and embedded pairs within base triples against random pairs (where random triples and pairs were constructed in the same way as in Fiser & Aslin, 2005, Experiment 1). Choice preferences in each of these comparisons were calculated as in Simulation 1, that is on the basis of the relative mean reconstruction error for the different stimulus types.

Results and Discussion

Figure 3b shows the network's preference for base objects ("Triplets") over random triples, and for embedded pairs ("Pairs") over random pairs. Like participants, the network shows a very strong preference for the base objects over random element triples, but virtually no preference for element pairs embedded in those objects over random pairs. The latter result is, technically, reliably different from chance, but this is because the error bars reflect variability (due to initial random weights) in a real-valued measure of choice probability, whereas participants' performance has much greater variability because, on our account, it reflects using this measure to make probabilistic choices on a trial-by-trial basis.

The explanation for the lack of a preference for embedded pairs is in some ways similar to that for the Giroux and Rey (2009) findings. When presented with an embedded pair, the network generates an internal representation that is similar to the one generated by the full triple, and thus the network generates error by incorrectly activating the missing element. The magnitude of this error is comparable to that produced for a random pair where each element weakly activates both of the elements with which it typically co-occurs.

As mentioned, we also trained a version of the network without any wrapping during training, which limits the range of positions in which each configuration can be presented. To determine the degree to which performance was affected by position-specific biases, we tested the trained network with components centered at each spatial position separately (centering toward the top or left of components with even width or height). In all other respects, the training and testing procedures were identical to those of the main simulation.

Table 1 shows the response preference of the network for pairs and triples as a function of spatial position. Data for some positions are lacking because not all components could be centered at all positions without wrapping. Overall, the network shows a strong preference for triples but not for pairs, replicating the main result. There is, however, a fair amount of variability across testing position, particular for pairs. This variability is caused by the fact that, without wrapping, the position-specific frequencies of particular pairs and triples vary, and these impact the network's performance. This issue is exacerbated for the simulations to follow, as many of them involve larger components. As the size of the relevant configurations increases—and, hence, the available presentation positions become more constrained—the variability in the position-specific frequencies of potential parts increases as well. To avoid these biases, we will continue to employ wrapping of configurations during training for the remainder of the simulations reported in this article.

Table 1
Response Preferences for Pairs/Triples Presented at Each Spatial Position for the Network Trained Without Wrapping

	1	2	3	4	5
1	.55/—	.52/—	.54/—	.52/—	.58/—
2	.60/—	.55/.87	.55/.92	.58/.91	.60/.91
3	.58/.93	.49/.91	.49/.92	.54/.91	.62/.94
4	.56/.91	.54/.90	.52/.91	.57/.90	.54/.92
5	.49/—	.51/—	.51/—	.51/—	—/—

Note. Column and row headings indicate x and y coordinates, respectively, of spatial positions within the display. Dashes indicate there were no data available at that position.

Simulation 3: Fiser and Aslin (2005) Experiment 3

Fiser and Aslin's (2005) second experiment simply confirmed that participants can learn disyllabic words when displays are composed of combinations of them instead of triples. Experiment 3 went further by confirming that participants can learn structures at multiple scales simultaneously—that is, both two- and three-element objects through the same exposure. If this were not the case, the lack of learning of embedded pairs within triples would have nothing to do embedding per se.

A series of six-element displays were composed from either two three-element objects or three two-element objects, in various spatial arrangements that preserved the property that at least two pairs of elements were adjacent horizontally. The configurations were constructed in such a way that their outlines were identical to those of configurations in their first experiment. Exposure and testing procedures were analogous to that experiment as well. As show in Figure 4a, participants reliably preferred both two-element objects ("Pairs") relative to random element pairs, and three-element objects ("Triplets") relative to random triples of elements.

When the network from Simulation 2—with exactly the same parameters, training, and training procedures—is trained on the analogous stimuli, it also shows clear preferences for both pairs

[5] Fiser and Aslin (2005, p. 525) were not explicit about the distribution of configurations of their stimuli, stating only that scenes were generated by "randomly positioning" pairs of triples so that the constraint was satisfied. We chose to present all 28 configurations to ensure all relevant statistics were balanced.

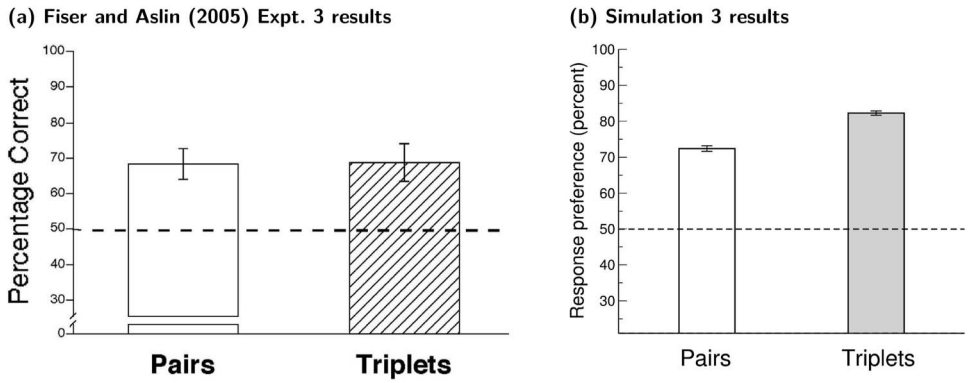

Figure 4. (a) Response preferences of participants in Fiser and Aslin's (2005) Experiment 3 for two-element objects versus random pairs ("pairs") and for three-element objects versus random triples ("triplets"); (b) Analogous response preferences in Simulation 3. Adapted with permission.

and triples relative to the corresponding random configurations (see Figure 4b), although the effect is stronger for triples than for pairs, which is not true of participants. We think this latter discrepancy arises because the network, unlike participants, was trained solely on the experimental stimuli, and this induces a bias favoring the specific larger-scale structure that occurs in the experiment. In other words, the network's representations are somewhat more tailored to the idiosyncrasies of the experimental stimuli than are those of participants.

Simulation 4: Fiser and Aslin (2005) Experiment 4

To this point, the network, like participants, shows sensitivity to both pairs and triples when they each behave independently, but not when the pairs are embedded in the triples. Experiment 4 of Fiser and Aslin (2005) provided two additional relevant findings to understanding the treatment of parts and wholes in visual statistical learning. The first is to extend the nature of learned structure to four-element objects (quadruples) and their embedded pairs; the second is to provide at least a limited measure of the timecourse of learning, by reporting forced-choice preferences in the middle of the exposure period as well as the end.

In this experiment, displays were composed of one of two quadruples and one of two pairs (see Figure 5). Figure 6a shows the results for participants after being exposed to 120 displays in a 6 min movie ("1st round"), and then again after the same amount of exposure ("2nd round"). Participants learned both the two- and four-element objects, with only a slight (but not reliable) numeric increase in preferences from the 1st round to the 2nd round. By contrast, and replicating their Experiment 1, participants did not prefer pairs that were embedded within the four-element object (compared with random pairs), despite having the same frequency and conditional probabilities as the two-element objects.

The same network and procedures that were used in Simulations 2 and 3 was trained on displays analogous to those used in Experiment 4 by Fiser and Aslin (2005). Because each round of exposure to participants was about half of that used in the other experiments, we tested the network after 50 and 100 epochs of training. As shown in Figure 6b, the network replicates the findings from participants of strong preferences for two- and four-element objects (compared with random versions) at both points in training, but no preference for embedded pairs (compared with random pairs). From 50 to 100 epochs, the network does show a distinct increase in its preference for nonembedded pairs and, to a lesser extent, for quadruples, whereas these effects are only trends in the empirical data. Moreover, the network's preference for quadruples is stronger than it is for pairs, which is not true of participants, echoing the discrepancy between triples and pairs in Simulation 3 (and likely due to the same bias favoring larger-scale structure). In other respects, though, the network's pattern of performance provides a reasonable match to that of participants.

Simulation 5: Fiser and Aslin (2005) Experiment 5

Fiser and Aslin's (2005) empirical results establish that participants can discover "objects" (pairs, triples, quadruples) consisting of a fixed spatial arrangement of elements with conditional probabilities of 1.0, in displays in which they co-occur with a variety of other such objects. However, subsets of these objects, such as pairs within triples or quadruples—despite having equal frequencies and elementwise conditional probabilities—do not appear to be represented as independent entities at all, insofar as participants do not consider them to be more familiar than random pairs when presented in isolation. That is, to this point, participants (and the network) are behaving in accordance with a strict, all-or-none version of the embeddedness constraint.

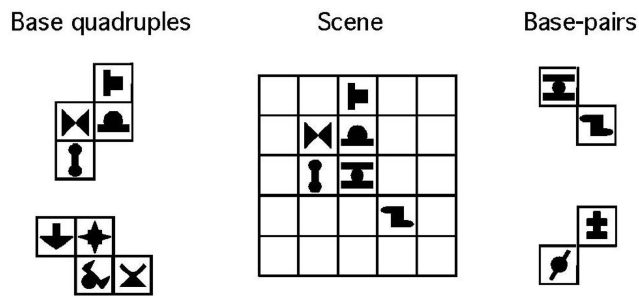

Figure 5. Two- and four-element objects used to construct displays in Experiment 4 of Fiser and Aslin (2005), and an example display. Adapted with permission.

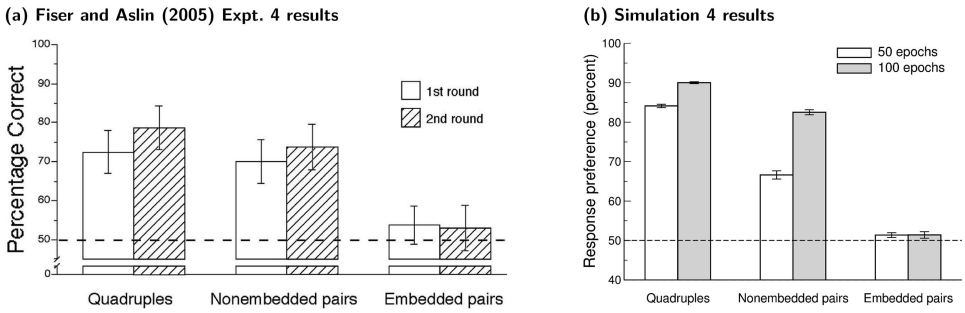

Figure 6. (a) Response preferences of participants in Fiser and Aslin's (2005) Experiment 4 for four-element objects versus random quadruples ("quadruples"), for 2-element objects versus random pairs ("nonembedded pairs"), and for embedded pairs within quadruples versus random pairs ("embedded pairs"), after being exposed to a 6 min video of 120 displays ("1st round") and then after an equivalent amount of additional exposure ("2nd round"); (b) analogous response preferences in Simulation 4 (after 50 and 100 epochs of training). Adapted with permission.

It is in this context that Experiment 5 of Fiser and Aslin (2005) is particularly informative. Fiser and Aslin constructed displays containing seven elements, six of which were one of two base sextuplets (see Figure 7). The seventh element—positioned anywhere adjacent to the sextuple—was selected from among two variable or "noise" elements from the other sextuple (shown in gray). This manipulation created two kinds of embedded pairs that were matched in frequency: *strong* pairs with no noise element and, hence, conditional probabilities of 1; and *weak* pairs that contain a noise element and, thus, have lower conditional probabilities but higher element frequencies (as the noise element also occurs in half of the displays containing the other sextuple).

Figure 8a shows participants' preferences for four types of comparisons following exposure. First, participants showed a mild but reliable preference for single elements with higher frequency (i.e., the noise elements) relative to other elements. Second, they showed a strong preference for strong pairs over random pairs. Third, although participants did prefer weak pairs over random pairs, this preference was significantly weaker than that for the strong pairs, indicating that they were sensitive to elementwise conditional probabilities. Also, this preference holds despite the fact that the weak pairs had higher mean element frequencies than the strong pairs, and that participants were sensitive to these frequencies (in the first comparison). Finally, participants also showed a strong preference for embedded over random quadruples (even though the majority of these contained at least one noise element). Figure 8b shows that the network, when trained in an analogous fashion, shows the same pattern of results.

These results provide clear evidence that a more graded interpretation of the embedded constraint is needed. Although it was not tested, it seems safe to assume that participants would have shown a strong preference for the base sextuples against random sextuples. In fact, if the network is tested on this comparison using random sextuples with the same overall outline, its preference for the base sextuples is 92.1% (SE = 0.102). If something similar held for participants, then all of their preferences in this experiment reflected learned sensitivity to structure embedded within larger wholes.

As suggested by Fiser and Aslin (2005), these findings can be understood within a framework that retains explicit representations of wholes and parts, but in which these vary in strength depending on their relative predictiveness. The network provides a similar account, but without the need to posit explicit chunks of any sort. In a situation in which larger-scale wholes are somewhat less coherent (due to the noise elements), the network still learns representations that capture their structure but now also captures their substructure to a greater extent. In this way, the network treats something as both a whole and as parts when doing so better captures the underlying statistical structure in the domain.

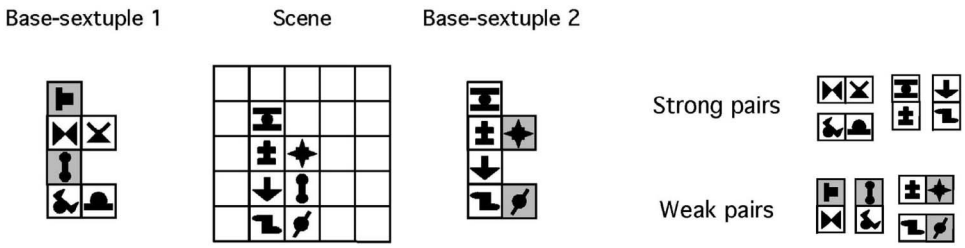

Figure 7. Design of stimuli used in Experiment 5 of Fiser and Aslin (2005). An example seven-element display consisted of one of two base sextuples and one of two variable elements (shown in gray) from the other base sextuple. "Strong pairs" are adjacent elements within a sextuple that only occur as part of that sextuple; "weak pairs" are also are adjacent elements within a sextuple but contain a noise element that occurs with the other sextuple. Adapted with permission.

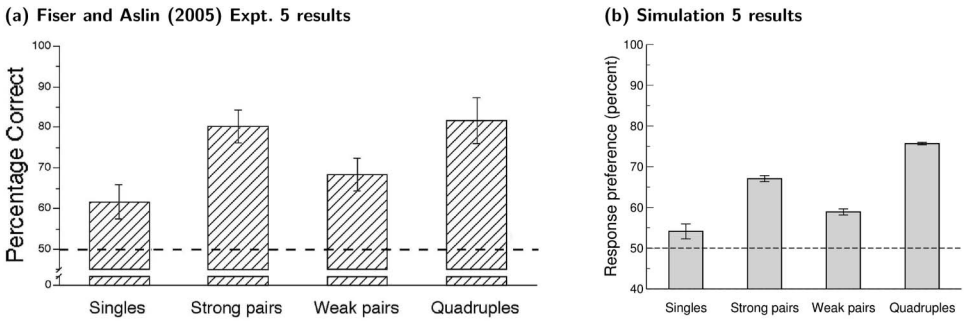

Figure 8. (a) Response preferences of participants in Fiser and Aslin's (2005) Experiment 5 for noise elements (with elevated frequencies) relative to other elements ("singles"), embedded pairs with no noise elements relative to random pairs ("strong pairs"), embedded pairs containing a noise element relative to random pairs ("weak pairs"), contiguous quadruples embedded in the base sextuples relative to random quadruples ("quadruples"); (b) analogous response preferences in Simulation 5. Adapted with permission.

Simulation 6: Orbán, Fiser, Aslin, and Lengyel (2008)

As discussed in the Introduction, Orbán et al.'s (2008) BCL is an elaboration of the idea that wholes and parts are not treated as mutually exclusive alternatives, but that both make graded contributions to performance. Orbán and colleagues were primarily concerned with evaluating the BCL against simpler learning mechanisms, including an associative learner (AL) that adopted the same general structure as the BCL but used Hebbian learning of pairwise correlations between elements without an explicit notion of chunks. Both the BCL and AL were capable of replicating the results of two studies from Fiser and Aslin (2001) as well as those of Experiments 1 and 4 from Fiser and Aslin (2005), although the AL's match to the latter was somewhat poor. To distinguish between the two accounts, Orbán and colleagues carried out an experiment in which two types of triples were matched in terms of elementwise and pairwise statistics and yet varied in their higher-order structure (to which only the BCL should be sensitive).

Each display contained six elements constructed from the components shown in Figure 9, such that each individual component occurred on the proportion of trials listed for its type. This was accomplished by randomly picking a triple, pair, and singleton on two thirds of trials, and randomly picking a pair to be presented with the quadruple on the remaining one third of trials. Note that the *true* triples on the left are constructed from four elements in a fixed spatial arrangement, although this quadruple is never presented to participants. The so-called *false* triples are similar in that they consist of three of the four elements of a quadruple, but in this case the quadruple itself was presented but not the triples (except as part of the quadruple, of course). True and false triples are matched on first- and second-order statistics; the difference is that false triples never occurred as triples per se—they always occurred embedded in the quadruple, and their elements could also occur independently as singletons. During testing, these two types of triples were compared against each other and against *mixed* triples composed of elements belonging to different components.

Figure 10a shows the performance of participants, and of the BCL and AL, on these comparisons. Participants preferred true triples to both mixed and false triples, but showed no preference between false and mixed triples (essentially exhibiting a type of embeddedness constraint). The BCL performed similarly. Unlike participants, however, the AL learned false triples nearly as

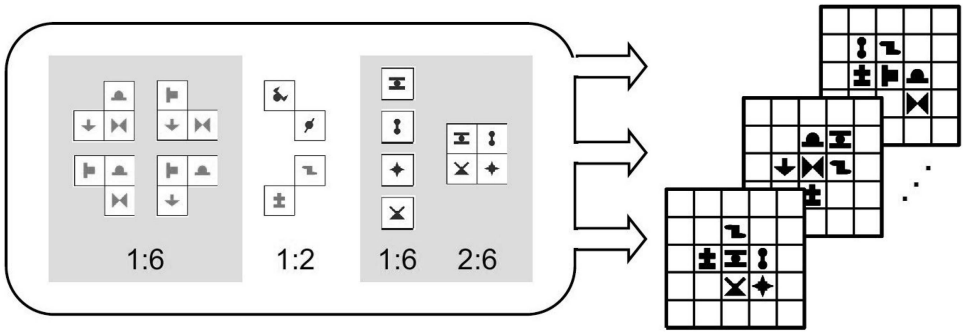

Figure 9. Design of stimuli used by Orbán et al. (2008). On the left are the components used to construct display—triples, pairs, singletons, and a quadruple—and, for each type, the proportions of all trials in which each individual component occurs; on the right are example displays. Triples on the left are termed "true"; triples formed from three of four elements of the quadruple are termed "false" (relative to random triples, termed "mixed"). Adapted with permission.

strongly as true ones, showing no preference between them, and also preferred false triples to mixed ones.

In applying the network from the previous simulations to the current study, we decided to reduce the learning rate to 0.05 because there are many more examples per epoch compared with the previous experiments. In all other respects, the same parameters and procedures were applied. Figure 10b shows the performance of the network when trained and tested on stimuli analogous to those used in the Orbán et al. (2008) study. Like the BCL and unlike the AL, the network shows the same overall pattern as participants, with the exception that it shows a slight dispreference for false compared with mixed triples—in part because it has learned the lone quadruple somewhat more strongly than participants and therefore predicts the missing element when presented with a false triple (although this dispreference would not be reliable with equivalent variance as participants). Critically, the network treats true and false triples as entirely different, even though they are matched on low-order statistics. Thus, despite being trained in something like an "associative" way, the network learns hierarchical structure in a way that is fundamentally different from the assumptions built into the AL—that is, simple pairwise correlations between elements—and more in line with participants.

Simulation 7: Simultaneous Representation of Parts and Wholes

Our theory of visual statistical learning is based on the idea that the visual system learns structure at multiple levels simultaneously, and that whether or not a given level influences performance depends on the details of the statistical structure to which participants are exposed (see also Turk-Browne et al., 2008). We have shown that a neural network that learns internal representations that capture the structure in a set of displays—as evidenced by its ability to reconstruct those displays—provides a good match to a range of empirical findings on the relationship between objects of different sizes (wholes) and the structure embedded within those objects (parts). To this point, though, none of the results have directly established that the network simultaneously represents a given subset of the input both as a whole and as a collection of parts, and that its behavior is influenced by both of these levels of structure simultaneously. Accordingly, we thought it worthwhile to carry out a final simulation that establishes this property of the network directly. The results also serve as predictions of the pattern of performance that participants would exhibit in an analogous study.

Our starting point was Simulation 4, which involved both quadruples and pairs as objects. The basic idea of the simulation is that embedded pairs within one of the quadruples, labeled "Parts" in Figure 11, will be trained as independent two-element objects and thus come to be treated as parts of that quadruple. We will then compare how well the network represents a configuration that is both a whole and parts (the first quadruple) to those that are either only a whole (the second quadruple) or only parts (combinations of the base pairs).

The difficulty is that, by adding additional training of parts of one of the quadruples, its elementwise and pairwise frequencies are increased relative to the other quadruple and the base pairs, and thus it would be uninteresting if it showed an advantage. Accordingly, we constructed displays in such a way that the elementwise and pairwise frequencies are matched across the various comparisons (as reflected in the factors shown in parentheses in Figure 11). To accomplish this, it was necessary to add displays consisting of both base pairs and one of the parts (in addition to displays in which a quadruple occurs with a base pair or part). Specifically, the first (part-based) quadruple co-occurred with each of its own parts with a frequency of 1, and with each of the base pairs with a frequency of 2. The frequencies for the second (non-part-based) quadruple were double these. In addition, the two base pairs co-occurred with each of the parts with a frequency of 3. Eight configurations of each display type were created that satisfied Fiser and Aslin's (2005) criterion that at least two horizontally adjacent pairs of elements were present in each display. Because of the number of displays involved, the learning rate in the network was reduced to 0.2, but all other parameters and procedures from the previous simulations were the same.

Although this design matches on lower-order statistics, it does not match the frequencies of the two quadruples, nor the frequencies of the base pairs and parts—indeed, this cannot be done without disrupting the conditional probabilities of these components by presenting their constituent elements in other contexts. Thus, in interpreting the results, one must keep in mind that the part-based quadruple has only half the frequency of the non-part-

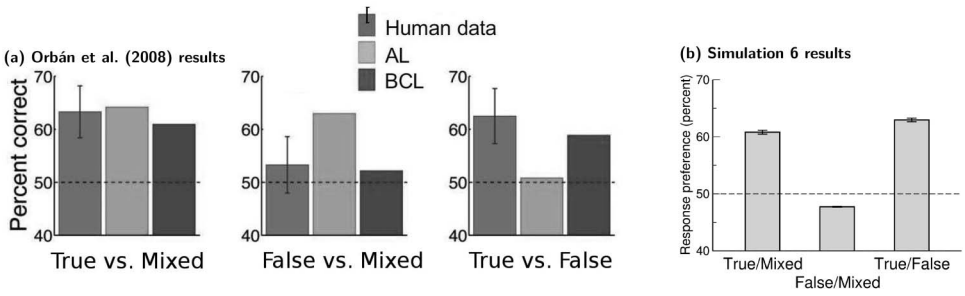

Figure 10. (a) Percentage of trials in which participants in Orbán et al.'s (2008) study, the Associative Learner (AL), and the Bayesian chunk learner (BCL) preferred true triples (shown as triples during exposure) over mixed triples (with elements drawn from different components), false triples (shown only embedded in a quadruple) over mixed triples, and true over false triples; (b) analogous response preferences in Simulation 6. Adapted with permission.

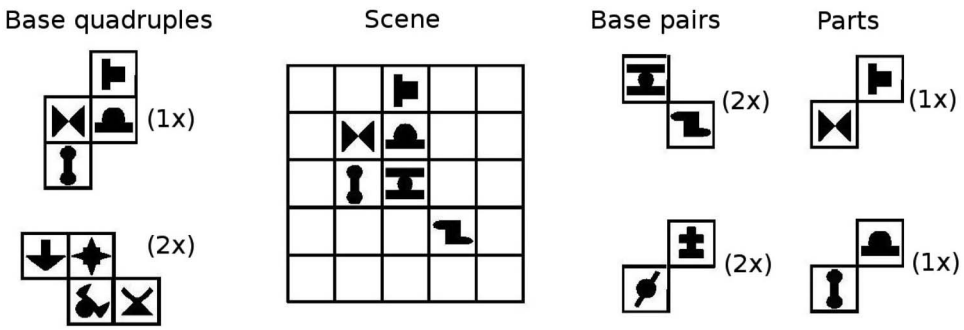

Figure 11. Components used in constructing displays in Simulation 7. Factors in parentheses are the relative frequencies of each component. Components labeled "Parts" are the only new components compared with Simulation 4 and are taken from the first quadruple.

based one, and only half the pairwise frequencies of configurations of the two base pairs.

Figure 12 shows the networks preference for the part-based quadruple ("Whole+Parts"), for the non-part-based quadruple ("Whole only") and for configurations composed of both base-pairs together in a pseudoobject ("Parts only"), each compared against random quadruples. After 100 epochs of training, the network's preference for the part-based quadruple is only slightly below that for the non-part-based quadruple, despite the latter's greater frequency. To avoid possible ceiling effects, we also presented data from earlier in training (70 epochs), where preferences are weaker but the same basic pattern holds. Thus, a part-based representation of the former provides nearly as much benefit as having twice the whole-object frequency. Moreover, although both quadruples have elevated pairwise frequencies, this alone cannot explain the results because the network's preference for four-element configurations matched in pairwise frequency ("Parts only") is much lower. The contrast between the "Whole+Parts" and "Parts only" conditions establishes a substantial benefit from having a whole-based representation, and the similarity of the "Whole+Parts" and "Whole only" conditions (in the face of a difference in whole-object frequency) establishes a substantial benefit from having a part-based representation. Thus, the network shows evidence of simultaneously benefiting from whole and part-based representations and, in this way, is very different from theoretical frameworks in which wholes and parts are alternative interpretations of a given input.

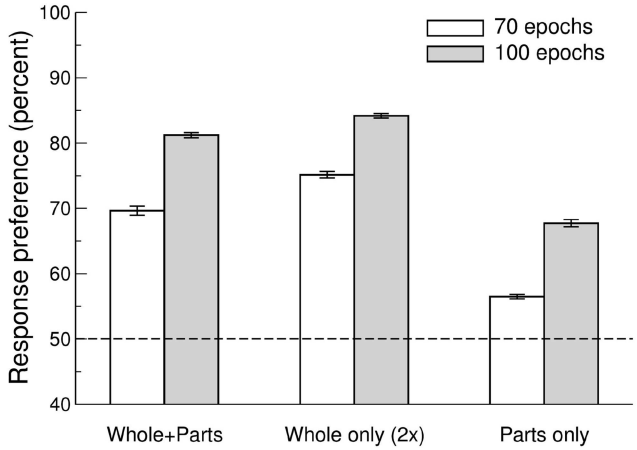

Figure 12. Response preferences of Simulation 7 for the quadruple composed of parts ("Whole+Parts"), for the quadruple not composed of parts ("Whole only [2x]") and for configurations combining the two base pairs ("Pairs only"), each compared against random quadruples, after 70 and 100 epochs of training. The "(2x)" indicates that the whole-object frequency in that condition is twice the whole-object frequency in the first condition.

General Discussion

Learning about entities such as spoken words and visual objects is not simply a matter of segregating them out of the input stream and then mapping them to their meanings as undifferentiated wholes, but must include learning their internal structure and how this structure contributes to their meanings. Among the multiple cues to discovering and representing words, objects, and their parts, the statistical distribution of their occurrence and co-occurrence makes a critical contribution (Conway & Christiansen, 2005; Fiser & Aslin, 2001; Kirkham et al., 2002; Saffran et al., 1996; Smith et al., 2014; Swingley, 2005).

Empirical studies that have directly examined the relationship between wholes and their parts in statistical learning (Fiser & Aslin, 2005; Giroux & Rey, 2009) have generally found that learning larger-scale structure tends to inhibit learning smaller-scale structure embedded within it (see also Zhao, Ngo, McKendrick, & Turk-Browne, 2011). Fiser and Aslin (2005) proposed that such an embeddedness constraint helps avoid the combinatorial explosion of having to consider all combinations of lower level components when representing higher order structure. As they put it,

> [I]f there is a reliable mechanism that is biased to represent the largest chunks in the input in a minimally sufficient manner, rather than using a full representation of all possible features, this constraint can eliminate the curse of dimensionality. (Fiser & Aslin, 2005, p. 532)

They recognized, though, that this constraint had to be graded rather than all-or-none in order to be consistent with the full range of relevant findings.

Orbán et al. (2008) elaborated on this graded view by developing a Bayesian formulation of visual statistical learning. The

Bayesian chunk learner (BCL) combines all possible inventories of parts and wholes, as a function of their likelihoods given priors and the experimental exposure, to provide a measure of the relatively likelihoods of test displays. To be clear, though, there is no a priori embeddedness constraint in operation, and rather than trying to avoid the curse of dimensionality, if anything the approach fully embraces it.[6] Of course, the computations employed by the BCL to calculate probabilities are not intended to be interpreted as being carried out by participants in any literal sense. Rather, the system provides an elegant but abstract characterization of how statistical information at different scales can be combined to account for human performance. It does leave open the question, though, of how a system can learn to be sensitive to the appropriate statistical structure in a computationally efficient and plausible manner.

The current work attempts to address this question. We propose an account based on learning internal distributed representations in artificial neural networks. This type of learning doesn't suffer a combinatorial explosion because it doesn't consider all possible subsets as explicit, discrete representations (Bengio & Bengio, 2000). Rather, representations are shaped by capturing the information in the input in an efficient manner and, where applicable, by activating relevant downstream representations in the service of supporting effective behavior. The learned representations are potentially sensitive to all combinations at all levels of structure simultaneously, but are pressured to be efficient due to limited representational resources (e.g., numbers of hidden units). The result is that the network develops graded representations of multilevel structure that are largely insensitive to embedded structure when it is completely redundant with larger-scale structure, but which capture both levels of structure when they make somewhat separable contributions to the distribution of inputs. The fact that participants behave similarly across a range of studies suggests that their perceptual systems may operate according to similar principles.

Relation to Other Approaches

Our work continues a broad, extensive effort to apply neural networks to model human learning and performance across a wide range of domains (see Rogers & McClelland, 2014), including auditory statistical learning (e.g., Christiansen et al., 1998; French et al., 2011; Mirman et al., 2010; Sirois et al., 2000). There have been many applications of neural networks to learning visual object representations (e.g., Cadieu et al., 2014; Hinton & Salakhutdinov, 2006; Khaligh-Razavi & Kriegeskorte, 2014; Mel, 1997; O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013) but none that we know of directed specifically at modeling empirical data on visual statistical learning. In this section, we clarify the relationship of the current work to a neural network model of auditory statistical learning: the Truncated Recursive Autoassociative Chunk Extractor (TRACX; French et al., 2011). We also discuss the relationship of our approach to Bayesian approaches more generally, and to what is often termed "associative" learning.

TRACX. TRACX (French et al., 2011) is a particular type of autoencoder neural network known as a Recursive Auto-Associative Memory (RAAM; Pollack, 1990). The input and output layers are composed of two slots of units, where each slot is the same size as the intervening hidden layer. The network is trained to reconstruct pairs of elements (e.g., A+B, one in each slot). When the network has learned to reconstruct a pair effectively, its hidden representation, [A+B], can then be used as an input in one of the slots in order to form larger groupings (e.g., [A+B]+C) which in turn, when learned, can enter into still larger groupings (e.g., [[A+B]+C]+D), and so on. TRACX consists of an RAAM applied to auditory statistical learning, in which a sequence of syllables is streamed across the two input slots from right to left, such that the left input is replaced with the hidden representation of the previous pair whenever that pair's reconstruction error falls below a specified criterion. French et al. (2011) showed that TRACX accounts for a number of challenging phenomena in auditory statistical learning, including the results of Giroux and Rey (2009) (although see the following text).

As a neural network that learns distributed representations of multilevel structure, TRACX clearly shares many properties with the current approach. Within both, the strength with which a particular pair, triple, and so forth, is learned is a matter of degree, and thus different groupings at the same scale can vary in strength of encoding in a way that lines up with human performance. There are, however, some important differences. The architecture and machinery of TRACX is structured specifically to apply to pairs of inputs, and an explicit criterion determines whether a given pairing is or isn't treated as a chunk—which, in turn, determines how its subsequent training experience is organized. Neither of these constraints apply to our approach. Indeed, it's not clear how TRACX could even be applied to visual statistical learning of the sort studied by Fiser and Aslin (2005).

When applied to sequential learning, the TRACX architecture leads to the primacy of pairwise structure over other scales, as well as a preference for leading substructure ([A+B]+C]) over trailing substructure (A+[B+C]). This has problematic implications for its treatment of Giroux and Rey's (2009) findings. French et al. (2011) report that, as TRACX learns the quadruple KLMN as [[K+L]+M]+N, its error on embedded pairs LM and MN increases—consistent with an embeddedness constraint. Unfortunately, however, performance on embedded pair KL was not reported. In fact, because it continues to be treated as a separate chunk (as shown in the organization for KLMN), its error would continue to decrease. The same is true for the initial triple KLM. Thus, TRACX is inconsistent with an embeddedness constraint for leading embeddings. Giroux and Rey (2009) did not report performance separately for leading and trailing embeddings, but there is no reason to think they didn't both show the effect. As reported for Simulation 1, our network's preference for leading embedded pairs (58.6%) is slightly greater than for trailing embedded pairs (55.3%), although both are much weaker than its preference for disyllabic words (67.6%).

Among the results that TRACX did model successfully was participants' ability to learn to segment a syllable stream based solely on backward transition probabilities (Perruchet & Desaulty, 2008)—that is, when certain elements reliably precede (rather than follow) other elements. French and colleagues showed that an SRN trained on syllable prediction, like Simulation 1, fails to show the same behavior. To be clear, this has nothing to do with failing to predict final silence at test. Rather, it is due to training solely on

[6] Orbán et al. (2008) use approximations to make the probability calculations more tractable, and they point out in their supplementary material (in a footnote, p. 1) that "the posterior over inventories was overwhelmingly dominated by the marginal likelihood of a single inventory," so the practical computational demands of their formulation are not as severe as they might be in principle.

prediction. If an SRN is trained not just on prediction, but also on maintenance and reconstruction, it is also sensitive to backward transition probabilities.

To demonstrate this, we applied a modified version of Simulation 1 to the backward grammar used by Perruchet and Desaulty (2008). This grammar uses 12 syllables to create nine disyllabic "words": three high-frequency words (XA, YD, and ZG) which were trained three times more often than six low-frequency words (XB, XC, YE, YF, ZH, and ZI). When these words are ordered randomly in a stream, the resulting partwords (AY, AZ, DX, DZ, GX, and GY) occur as often as the low-frequency words, but the latter have backward transition probabilities of 1 (B and C are always preceded by X, E, and F are always preceded by Y, and H and I are always preceded by Z). The only difference from Simulation 1 was that the network was trained not only to predict the next syllable, but also to reconstruct the current syllable and to recall the previous syllable. We think of all three of these tasks as proxies for the need to remember, represent, and predict information in the course of language processing. Testing also involved all three tasks (except for recall of the previous syllable on presentation of the first syllable) and included final silence.

After 2100 syllable presentations—the same as Perruchet and Desaulty's (2008) participants—with the same learning parameters as in Simulation 1, the network showed a reliable preference for the low-frequency words compared with the partwords (M = 62.4%, SE = 1.13, over 20 samples of initial random weights). Thus, contrary to French et al.'s (2011) claims, sensitivity to backward transition probabilities is not problematic for the application of SRNs to statistical learning, and provides no differential support for their account over ours.

Bayesian approaches. Although neural network and Bayesian approaches are often compared (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McClelland et al., 2010), they typically have different goals and address different questions. Neural networks attempt to approximate the underlying computational mechanisms that learn, represent, and process information in real time. In contrast, Bayesian approaches attempt to understand the structure of a problem, formulate an optimal solution to it, and then examine how the cognitive system might approximate this solution (see Griffiths, Vul, & Sanborn, 2012; Rogers & McClelland, 2014, for further discussion). Even so, the approaches are not incompatible: learning and processing in neural networks—both real (e.g., Fiser, Berkes, Orbán, & Lengyel, 2010; Orbán, Berkes, Fiser, & Lengyel, 2016; Pouget, Beck, Ma, & Latham, 2013) and artificial (e.g., Hinton & Sejnowski, 1983; McClelland, 2013; Rumelhart, Durbin, Golden, & Chauvin, 1995)—can be understood in terms of Bayesian statistical inference. Moreover, some computational approaches—particularly those positing structured, symbolic representations—are also committed to modeling underlying computational mechanisms of performance, and use Bayesian inference to learn over such representations in a principled manner (e.g., Kemp, 2012; Kemp & Tenenbaum, 2009).

The two approaches do typically adopt somewhat different stances with regards to the nature and use of priors. For neural networks, the computational formalism itself incorporates certain commitments—ultimately inspired by basic principles of neural computation—that give rise to specific biases about what kinds of things are easier versus harder to learn, and how performance generalizes to novel items. In particular, the use of multiple layers of units which each pass a linear combination of inputs through a limited (e.g., monotonic) nonlinearity causes networks to be better at learning mappings which preserve (vs. violate) similarity, which depend on lower order (vs. high-order) relationships among inputs, and which involve local (vs. long-distance) dependencies in time and/or space. A given model augments these intrinsic biases with domain-specific ones, expressed largely in terms of the training environment to which the network is exposed, as well as some aspects of the architecture of the network (e.g., receptive-field sharing in a CNN). As a result, some aspects of the behavior of a model are more-or-less unavoidable. For instance, the current simulations of visual statistical learning employ a very generic network architecture—a fully connected, feedforward autoencoder—and involve training only on the experimental stimuli themselves. The resulting behaviors thus reflect the intrinsic biases of the formalism itself and there is very little that could be done to alter these effects. If, for example, participants in Fiser and Aslin's (2005) Experiment 1 had shown a strong preference for embedded over random pairs, the network would be unable to account for this finding.

On a Bayesian approach, biases on performance derive from how the statistical structure of the environment relates to specified priors regarding what types of structure should be preferred. These priors thus play the same role as network architecture and intrinsic formalism properties within the neural network approach. The difference, though, is that the origin of priors, and the constraints on their properties, are not always made explicit or derived from more basic theoretical principles (Jones & Love, 2011). For example, some parameters in the BCL bias the size of the chunks to be formed. It is impressive that Orbán et al. (2008) could fit the results of multiple studies with a single set of priors on the parameters, but it is unclear, apart from fitting the data, why the system should have those priors and not others. Indeed, it would be informative to compare the system's preferences to those of participants when a broader range of chunk sizes are relevant. It is unfortunate, in this regard, that the BCL was not applied to Experiment 5 of Fiser and Aslin (2005), which involved sextuples, quadruples, pairs and singletons. Not only do these chunks span a broad range of spatial scales, but the results indicate that participants did not adhere to a strict form of the embeddedness constraint as they had in most of the studies fit by the BCL.[7]

Given a particular ensemble of displays, the specified priors in the BCL give rise to preferences for certain chunk inventories over others. Other priors would yield other preferences. Without an understanding of why the system has a particular set of priors and not others, the fact that those priors lead to preferences that are similar to those of participants does little to explain why participants behave the way they do. In general, Bayesian accounts—and all other modeling efforts, for that matter—are most informative when the sources of their biases can be independently justified (see, e.g., Geisler, 2008; Griffiths & Tenenbaum, 2006).

"Associative" learning. Some researchers think of neural networks as a type of associative learning mechanism, perhaps be-

[7] There is one result reported by Orbán et al. (2008) that violates the embeddedness constraint but which the BCL (and the associative learner) nonetheless successfully modeled: participants' preference for triples embedded in quadruples, relative to random triples (see Panel D of Figure 2, p. 2747 of their article). The data appear to come from Experiment 4 of Fiser and Aslin (2005), but as that article does not report the performance on triples in that experiment, the exact status of these data is somewhat unclear.

cause they are often trained to "associate" each input with a particular output. Unfortunately, this perspective conflates the structure of the task with the structure of the learning mechanism itself.

Orbán et al. (2008) contrasted the BCL with an AL, which failed to behave like participants when elementwise and pairwise statistics were controlled. Our neural network simulation behaved like the BCL and participants, and unlike the AL. That is, although the network was performing an associative task—mapping inputs to outputs—it was not equivalent to an associative learner.

The critical property that doomed the AL is that its learning was based on direct relationships (correlations) among surface elements; the system had no means of learning new representations with an altered organization or similarity structure. Neural network without hidden units are similarly limited (Minsky & Papert, 1969), but those with hidden units, like our network, can learn to rerepresent the input in a nonlinear way, thereby emphasizing some relationships (e.g., "chunk-like" configurations) and ignoring others (e.g., some types of embedded structure). Thus, in evaluating proposals concerning associative learning, it is critical to examine the detailed properties of the underlying learning mechanisms—particularly with regard to their ability to learn new representations—and not simply assume that all systems that perform an associative task have equivalent properties and limitations.

Limitations and Future Directions

Although the networks examined in the current work are largely successful in accounting for the relevant behavioral findings, like all models they suffer from limitations in their scope and adequacy which are important to understand, and which provide opportunities for improvement in subsequent work.

Positional invariance. The most obvious limitation of the current work concerns our treatment of positional invariance. Participants come to the experiment with an extensive—although not complete (Kravitz, Kriegeskorte, & Baker, 2010)—ability to generalize their visual knowledge across retinotopic position. While the basis for this degree of positional invariance is not fully understood, it presumably results from a combination of patterns of eye movements, the integration of visual information across successive retinotopic representations in the visual system, and learning from extensive visual experience across retinotopic positions (see Kravitz, Vinson, & Baker, 2008, for discussion). Our simulations of visual statistical learning incorporated only the last of these factors, and only for the experimental stimuli directly. Specifically, whereas participants viewed each display at most a couple of times during the exposure phase of each visual statistical learning experiment, we trained the network on each display at every possible position, wrapping the displays vertically and horizontally to avoid edge effects. Although this was intended to capture the impact of both pre- and within-experimental experience, and we illustrated (for Simulation 1) that wrapping was not necessary to account for the relevant effects, the training procedure is clearly a poor approximation to participants' experience. Moreover, it fails to account for conditions in which participants do show sensitivity to absolute spatial position (see, e.g., Fiser & Aslin, 2001).

We believe that the principles that govern the treatment of parts and wholes in visual statistical learning can be studied and understood without a fully adequate treatment of positional invariance in visual perception, and our training regime simply allowed us to do this in a way that avoids positional biases entirely. Even so, a full account of visual statistical learning must ultimately be framed within a more general theory of the visual system which accounts for positional invariance along with a host of other properties, and we view this as an important direction for future work.

Learning procedure. Even setting aside the additional positional training, our visual simulations received a greater number of presentations of each configuration compared with participants. This was due largely to the fact that the network had no preexperimental experience of any sort, in order to avoid possible biases from the structure of such experience. It is possible that, by implementing a better theory of positional variation and by pretraining the network with experience comparable to that of participants, the simulations could have matched not just the outcome of learning but also its timecourse on a trial by trial basis. We suspect, though, that a further discrepancy between participants and the simulations would also need to be addressed—namely, the specific use of back-propagation as the learning procedure in the simulations. Back-propagation is essentially using the chain-rule from calculus to calculate the derivative of performance error with respect to each connection weight in the network; these derivatives are then used to change the weights to improve performance. Although a literal implementation of back-propagation itself is not biologically plausible (Crick, 1989), it is one of a class of error-correcting procedures which have very similar characteristics, some of which are more plausible (e.g., Hinton, Osindero, & Teh, 2006; O'Reilly, 1996).

These procedures differ in their efficiency, however, and so matching the speed of human learning will probably require the application of a learning procedure which is a closer match to the one actually employed by participants. It is interesting to note, though, that despite its implausibility, back-propagation can give rise to hidden representations whose similarity structure is closely related to that of the corresponding neural representations (e.g., object representation in the ventral visual pathway; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015; Yamins et al., 2014), suggesting that the properties of the underlying representations may depend more on the functional demands placed on them than on the specific details of the learning procedure.

Relation to the neural basis of statistical learning. Part of the attraction of neural networks, at least for some researchers, is that they can approximate principles of neural computation, both by employing neuron-like processing units and by having the units organized into groups that correspond to specific cortical regions (O'Reilly & Munakata, 2000). The current work has taken very little advantage of this, using highly simplified units and generic network architectures that, at best, could be interpreted as corresponding to cortical regions that support modality-specific perceptual processing. At a very general level, this stance is consistent with the widespread view that statistical learning reflects domain-general principles that operate throughout neocortex (Frost, Armstrong, Siegelman, & Christiansen, 2015). However, recent results suggest that medial temporal lobe structures may play an unexpectedly important role in the relatively rapid statistical learning that occurs in many experimental paradigms, particularly with regard to temporal structure (Davachi & DuBrow, 2015; Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014; Schapiro,

Turk-Browne, Norman, & Botvinick, 2016; Turk-Browne, Scholl, Chun, & Johnson, 2009). While it may be possible to understand these results as reflecting a mixture of implicit and explicit processing (Bertels, Franco, & Destrebecqz, 2012; Dale, Duran, & Morehead, 2012), a full account of statistical learning performance, both in the laboratory and in more natural settings, will no doubt require consideration of complex interactions between neocortical and subcortical structures (McClelland, McNaughton, & O'Reilly, 1995) as well as between the two hemispheres (Roser, Fiser, Aslin, & Gazzaniga, 2011).

Integration with recognition and comprehension. The current work applies a form of unsupervised learning—sometimes called *self-supervised* because the input is used both as input and output—as the empirical studies involve incidental learning without feedback. Visual representations do not develop solely in the service of reconstructing the input, however, but must also support the recognition and comprehension of words, objects, and other entities. Thus, a full account of the role of statistical learning in language and vision must be developed within a broader theory of learning, representing, and processing words, objects, and their meanings (Emberson & Rubinstein, 2016; Graf Estes et al., 2007; Hay et al., 2011; Koehne & Crocker, 2015; Smith et al., 2014).

Our simulations of auditory statistical learning are particularly impoverished in this regard. We do not really think that word learning is based on explicit training on syllable prediction, reconstruction and recollection. Rather, the demands of language processing in general involve learning to represent and maintain information so it can be integrated properly with upcoming information, and this integration will be aided to the extent that the upcoming information can be anticipated (Federmeier, 2007; Kuperberg & Jaeger, 2016). These properties of the language learning system then manifest in a variety of experimental contexts, including auditory statistical learning. Thus, we need to move from models of statistical learning to models of language (and visual) learning that exhibit the appropriate patterns of performance when faced with statistical learning paradigms (as well as many others).

Conclusions

Statistical learning is often cast as a solution to the problem of how people discover the "units" of perception, but true perceptual organization is far more complicated than finding a simple parse of the input to pass on to recognition and comprehension processes. Entities such as words and objects have rich internal structure that contribute in important ways to their interpretation, and any process that purports to use statistical structure to aid in recognition must be capable of discovering and representing both wholes and parts. Interestingly, empirical studies that have examined the relationship between wholes and parts in auditory (Giroux & Rey, 2009) and visual (Fiser & Aslin, 2005) statistical learning have found many cases in which discovery of the structure of the whole seems to inhibit discovery of substructure of that whole, which seems difficult to reconcile with a variety of empirical findings in which wholes and parts work together to support behavior. Bayesian formulations (Orbán et al., 2008) provide a natural means of expressing how the the integration of parts and wholes manifests in performance, but do not attempt to approximate the mechanisms that actually govern learning and processing in participants.

The current work presents a series of computational simulations in which artificial neural networks that learn internal representations in the service of representing auditory or visual input exhibit the same patterns of performance as participants, despite not forming discrete "chunks" for wholes and/or parts. Importantly, both participants and the networks behave somewhat differently when the large-scale structure is somewhat less reliable relative to the smaller-scale structure. This argues for a more graded "embeddedness constraint" on statistical learning (Fiser & Aslin, 2005) and suggests, in agreement with Bayesian formulations, that the degree to which structure at the whole- and part-levels contribute to behavior depends on their relative informativeness. An important direction for future work is to incorporate the same learning and representational principles into more general models of auditory and visual recognition and comprehension processes.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Science, 9,* 147–169. http://dx.doi.org/10.1207/s15516709cog0901_7

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science, 21,* 170–176. http://dx.doi.org/10.1177/0963721412436806

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15,* 460–466. http://dx.doi.org/10.1111/j.0956-7976.2004.00702.x

Barenholtz, E., & Tarr, M. J. (2011). Visual learning of statistical relations among nonadjacent features: Evidence for structural encoding. *Visual Cognition, 19,* 469–482. http://dx.doi.org/10.1080/13506285.2011.552894

Bengio, S., & Bengio, Y. (2000). Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks, 11,* 550–557. http://dx.doi.org/10.1109/72.846725

Bertels, J., Franco, A., & Destrebecqz, A. (2012). How implicit is visual statistical learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1425–1431. http://dx.doi.org/10.1037/a0027210

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., . . . DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology, 10*(12), e1003963. http://dx.doi.org/10.1371/journal.pcbi.1003963

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes Special Issue: Language acquisition and connectionism., 13*(2–3), 221–268.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120,* 235–253. http://dx.doi.org/10.1037/0096-3445.120.3.235

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 24–39. http://dx.doi.org/10.1037/0278-7393.31.1.24

Crick, F. (1989). The recent excitement about neural networks. *Nature, 337,* 129–132. http://dx.doi.org/10.1038/337129a0

Dale, R., Duran, N. D., & Morehead, J. R. (2012). Prediction during statistical learning, and implications for the implicit/explicit divide. *Advances in Cognitive Psychology, 8,* 196–209. http://dx.doi.org/10.5709/acp-0115-z

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences, 19,* 92–99. http://dx.doi.org/10.1016/j.tics.2014.12.004

Dominey, P. F., & Ramus, F. (2000). Neural network processing of natural language: I. Sensitivity to serial, temporal, and abstract structure of language in the infant. *Language and Cognitive Processes, 15,* 87–127. http://dx.doi.org/10.1080/016909600386129

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211. http://dx.doi.org/10.1207/s15516709cog1402_1

Emberson, L. L., & Rubinstein, D. Y. (2016). Statistical learning is constrained to less abstract patterns in complex sensory input (but not the least). *Cognition, 153,* 63–78. http://dx.doi.org/10.1016/j.cognition.2016.04.010

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44,* 491–505. http://dx.doi.org/10.1111/j.1469-8986.2007.00531.x

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 18014–18019. http://dx.doi.org/10.1073/pnas.0608811103

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504. http://dx.doi.org/10.1111/1467-9280.00392

Fiser, J., & Aslin, R. N. (2005). Encoding multi-element scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General, 134,* 521–537. http://dx.doi.org/10.1037/0096-3445.134.4.521

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences, 14,* 119–130. http://dx.doi.org/10.1016/j.tics.2010.01.003

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117,* 107–125. http://dx.doi.org/10.1016/j.cognition.2010.07.005

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review, 118,* 614–636. http://dx.doi.org/10.1037/a0025255

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19,* 117–125. http://dx.doi.org/10.1016/j.tics.2014.12.010

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

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202. http://dx.doi.org/10.1007/BF00344251

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59,* 167–192. http://dx.doi.org/10.1146/annurev.psych.58.110405.085632

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science, 33,* 260–272. http://dx.doi.org/10.1111/j.1551-6709.2009.01012.x

Gonnerman, L. M., Seidenberg, M. S., & Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General, 136,* 323–345. http://dx.doi.org/10.1037/0096-3445.136.2.323

Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language, 20,* 229–252. http://dx.doi.org/10.1017/S0305000900008266

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18,* 254–260. http://dx.doi.org/10.1111/j.1467-9280.2007.01885.x

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14,* 357–364. http://dx.doi.org/10.1016/j.tics.2010.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17,* 767–773. http://dx.doi.org/10.1111/j.1467-9280.2006.01780.x

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21,* 263–268. http://dx.doi.org/10.1177/0963721412447619

Hay, J. F., Pelucchi, B., Graf Estes, K., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology, 63,* 93–106. http://dx.doi.org/10.1016/j.cogpsych.2011.06.002

Hinton, G. E. (Ed.), (1991). *Connectionist symbol processing*. Cambridge, MA: MIT Press.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18,* 1527–1554. http://dx.doi.org/10.1162/neco.2006.18.7.1527

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313,* 504–507. http://dx.doi.org/10.1126/science.1127647

Hinton, G. E., & Sejnowski, T. J. (1983, June). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC: IEEE.

Hoffman, D. D., & Singh, M. (1997). Salience of visual parts. *Cognition, 63,* 29–78. http://dx.doi.org/10.1016/S0010-0277(96)00791-3

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34,* 169–188, 188–231. http://dx.doi.org/10.1017/S0140525X10003134

Jun, J., & Chong, S. C. (2016). Visual statistical learning of temporal structures at different hierarchical levels. *Attention, Perception & Psychophysics, 78,* 1308–1323. http://dx.doi.org/10.3758/s13414-016-1104-9

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review, 119,* 685–722. http://dx.doi.org/10.1037/a0029347

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116,* 20–58. http://dx.doi.org/10.1037/a0014282

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10,* e1003915. http://dx.doi.org/10.1371/journal.pcbi.1003915

Kibbe, M. M., & Feigenson, L. (2016). Infants use temporal regularities to chunk objects in memory. *Cognition, 146,* 251–263. http://dx.doi.org/10.1016/j.cognition.2015.09.022

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83,* B35–B42. http://dx.doi.org/10.1016/S0010-0277(02)00004-5

Koehne, J., & Crocker, M. W. (2015). The interplay of cross-situational word learning and sentence-level constraints. *Cognitive Science, 39,* 849–889. http://dx.doi.org/10.1111/cogs.12178

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex, 20,* 2916–2925. http://dx.doi.org/10.1093/cercor/bhq042

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, 12,* 114–122. http://dx.doi.org/10.1016/j.tics.2007.12.006

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1,* 417–446. http://dx.doi.org/10.1146/annurev-vision-082114-035447

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31,* 32–59. http://dx.doi.org/10.1080/23273798.2015.1102299

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521,* 436–444. http://dx.doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1,* 541–551. http://dx.doi.org/10.1162/neco.1989.1.4.541

Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology, 36,* 635–645. http://dx.doi.org/10.1037/0012-1649.36.5.635

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101,* 3–33. http://dx.doi.org/10.1037/0033-295X.101.1.3

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101–B111. http://dx.doi.org/10.1016/S0010-0277(01)00157-3

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology, 4,* 503.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14,* 348–356. http://dx.doi.org/10.1016/j.tics.2010.06.002

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457. http://dx.doi.org/10.1037/0033-295X.102.3.419

Mel, B. W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation, 9,* 777–804. http://dx.doi.org/10.1162/neco.1997.9.4.777

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Mirman, D., Graf-Estes, K., & Magnus, J. S. (2010). Computational modeling of statistical learning: Effects of transitional probability versus frequency and links to word learning. *Infancy, 15,* 471–486. http://dx.doi.org/10.1111/j.1532-7078.2009.00023.x

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron, 92,* 530–543. http://dx.doi.org/10.1016/j.neuron.2016.09.038

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 2745–2750. http://dx.doi.org/10.1073/pnas.0708424105

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8,* 895–938. http://dx.doi.org/10.1162/neco.1996.8.5.895

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. *Frontiers in Psychology, 4,* 124.

Otsuka, S., Koch, C., & Saiki, J. (2016). Visual statistical learning produces implicit and explicit knowledge about temporal order information and scene chunks: Evidence from direct and indirect measures. *Visual Cognition, 24,* 155–172. http://dx.doi.org/10.1080/13506285.2016.1211209

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36,* 1299–1305. http://dx.doi.org/10.3758/MC.36.7.1299

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10,* 233–238. http://dx.doi.org/10.1016/j.tics.2006.03.006

Perruchet, P., & Vinter, A. (1998). PARSER: A model of word segmentation. *Journal of Memory and Language, 39,* 246–263. http://dx.doi.org/10.1006/jmla.1998.2576

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes, 15*(4–5), 445–485. http://dx.doi.org/10.1080/01690960050119661

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence, 46,* 77–106. http://dx.doi.org/10.1016/0004-3702(90)90005-K

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience, 16,* 1170–1178. http://dx.doi.org/10.1038/nn.3495

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes, 23,* 942–971. http://dx.doi.org/10.1080/01690960802069730

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81,* 275–280. http://dx.doi.org/10.1037/h0027768

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025. http://dx.doi.org/10.1038/14819

Rogers, T. T., & McClelland, J. L. (2014). Parallel Distributed Processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science, 38,* 1024–1077. http://dx.doi.org/10.1111/cogs.12148

Roser, M. E., Fiser, J., Aslin, R. N., & Gazzaniga, M. S. (2011). Right hemisphere dominance in visual statistical learning. *Journal of Cognitive Neuroscience, 23,* 1088–1099. http://dx.doi.org/10.1162/jocn.2010.21508

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. E. Rumelhart (Eds.), *Back-propagation: Theory, architectures, and applications* (pp. 1–34). Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533–536. http://dx.doi.org/10.1038/323533a0

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928. http://dx.doi.org/10.1126/science.274.5294.1926

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology, 37,* 74–85. http://dx.doi.org/10.1037/0012-1649.37.1.74

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience, 26,* 1736–1747. http://dx.doi.org/10.1162/jocn_a_00578

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from

temporal community structure. *Nature Neuroscience, 16,* 486–492. http://dx.doi.org/10.1038/nn.3331

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus, 26,* 3–8. http://dx.doi.org/10.1002/hipo.22523

Schreuder, R., & Baayen, H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 131–154). Hillsdale, NJ: Lawrence Erlbaum.

Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences, 4,* 353–361. http://dx.doi.org/10.1016/S1364-6613(00)01515-1

Sirois, S., Buckingham, D., & Shultz, T. R. (2000). Artificial grammar learning by infants: An autoassociator perspective. *Developmental Psychology, 3,* 442–456.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences, 18,* 251–258. http://dx.doi.org/10.1016/j.tics.2014.02.007

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50,* 86–132. http://dx.doi.org/10.1016/j.cogpsych.2004.06.001

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. In D. Sandra & M. Taft (Eds.), *Morphological structure, lexical representation and lexical access*. Hillsdale, NJ: Lawrence Erlbaum. http://dx.doi.org/10.1080/01690969408402120

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Review, 139,* 792–814. http://dx.doi.org/10.1037/a0030801

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 399–407. http://dx.doi.org/10.1037/0278-7393.34.2.399

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience, 21,* 1934–1945. http://dx.doi.org/10.1162/jocn.2009.21131

Wheeler, D. (1970). Processes in word recognition. *Cognitive Psychology, 1,* 59–85.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 8619–8624. http://dx.doi.org/10.1073/pnas.1403112111

Yermolayeva, Y., & Rakison, D. H. (2014). Connectionist modeling of developmental changes in infancy: Approaches, challenges, and contributions. *Psychological Bulletin, 140,* 224–255. http://dx.doi.org/10.1037/a0032150

Zhao, J., Ngo, N., McKendrick, R., & Turk-Browne, N. B. (2011). Mutual interference between statistical summary perception and statistical learning. *Psychological Science, 22,* 1212–1219. http://dx.doi.org/10.1177/0956797611419304

Zhao, L., Cosman, J. D., Vatterott, D. B., Gupta, P., & Vecera, S. P. (2014). Visual statistical learning can drive object-based attentional selection. *Attention, Perception, & Psychophysics, 76,* 2240–2248. http://dx.doi.org/10.3758/s13414-014-0708-1

Received September 6, 2016
Revision received November 7, 2016
Accepted November 9, 2016 ■