PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Thiessen ED. 2017 What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Phil. Trans. R. Soc. B* **372**: 20160056.
http://dx.doi.org/10.1098/rstb.2016.0056

Accepted: 28 July 2016

One contribution of 13 to a theme issue 'New frontiers for statistical learning in the cognitive sciences'.

Subject Areas:
cognition

Keywords:
statistical learning, memory, modelling

Author for correspondence:
Erik D. Thiessen
e-mail: thiessen@andrew.cmu.edu

THE ROYAL SOCIETY
PUBLISHING

What's statistical about learning? Insights from modelling statistical learning as a set of memory processes

Erik D. Thiessen

Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

EDT, 0000-0002-2563-032X

Statistical learning has been studied in a variety of different tasks, including word segmentation, object identification, category learning, artificial grammar learning and serial reaction time tasks (e.g. Saffran *et al.* 1996 *Science* **274**, 1926–1928; Orban *et al.* 2008 *Proceedings of the National Academy of Sciences* **105**, 2745–2750; Thiessen & Yee 2010 *Child Development* **81**, 1287–1303; Saffran 2002 *Journal of Memory and Language* **47**, 172–196; Misyak & Christiansen 2012 *Language Learning* **62**, 302–331). The difference among these tasks raises questions about whether they all depend on the same kinds of underlying processes and computations, or whether they are tapping into different underlying mechanisms. Prior theoretical approaches to statistical learning have often tried to explain or model learning in a single task. However, in many cases these approaches appear inadequate to explain performance in multiple tasks. For example, explaining word segmentation via the computation of sequential statistics (such as transitional probability) provides little insight into the nature of sensitivity to regularities among simultaneously presented features. In this article, we will present a formal computational approach that we believe is a good candidate to provide a unifying framework to explore and explain learning in a wide variety of statistical learning tasks. This framework suggests that statistical learning arises from a set of processes that are inherent in memory systems, including activation, interference, integration of information and forgetting (e.g. Perruchet & Vinter 1998 *Journal of Memory and Language* **39**, 246–263; Thiessen *et al.* 2013 *Psychological Bulletin* **139**, 792–814). From this perspective, statistical learning does not involve explicit computation of statistics, but rather the extraction of elements of the input into memory traces, and subsequent integration across those memory traces that emphasize consistent information (Thiessen and Pavlik 2013 *Cognitive Science* **37**, 310–343).

This article is part of the themed issue 'New frontiers for statistical learning in the cognitive sciences'.

1. Introduction

The term 'statistical learning' was initially used to describe the fact that infants are sensitive to the probability with which syllables co-occur, and can use this property to segment words from fluent speech [1,2]. In this original set of experiments, the probability of co-occurrence between syllables was described in terms of 'transitional probability', which is defined as the frequency with which syllables X and Y co-occur, relative to the frequency with which X occurs. That is, if the syllable X occurs 100 times, and the conjunction of XY occurs 80 times, the transitional probability between X and Y is 80%. These experiments demonstrated that infants preferentially segment syllable groupings with high transitional probabilities compared with syllable groupings with low transitional probabilities. The fact that infants could use the statistical structure of the input to learn the identity of words provided support to theoretical accounts suggesting that learning plays an important role in language acquisition, and sparked tremendous interest in the nature and extent of statistical learning in infants and adults, and other species.

Subsequent research has demonstrated the breadth of statistical learning along three general dimensions. First, it is available over multiple stimulus types. In addition to linguistic stimuli, statistical learning has been demonstrated for pure tones (e.g. [3]), action sequences [4], visual stimuli such as scenes and shapes [5,6], and cross-modal relations (e.g. [7]). Second, statistical learning is available across the lifespan from neonates to older adults (e.g. [8,9]), and present in many species other than humans, including primates [10,11] and rats [12]. Third, learners are sensitive to more statistical relations than the sequential conditional relations described in terms of transitional probabilities in the original Saffran *et al.* [1] experiments. In addition, learners are sensitive to simultaneous co-occurrence probability [13], and the frequency and variability of exemplars (e.g. [14,15]).

The breadth of statistical learning phenomena demonstrates that this is a powerful learning mechanism, consistent with theories of development that place a more pronounced emphasis on learning as an important causal factor in developmental change (e.g. [16,17]). At the same time, however, these phenomena present a challenge to theories of statistical learning themselves: whatever mechanism or process underlies statistical learning must be able to account for learning in a wide variety of stimuli, tasks and statistical structure. In response to this variation, recent theories have suggested that statistical learning is in fact an umbrella term that describes the action of multiple independent mechanisms (e.g. [18]). Some of these theories suggest that there are different statistical learning mechanisms operating in different modalities, such that audio and visual statistical learning are accomplished by separate learning mechanisms (e.g. [19]). Others have suggested that different statistical structures are learned via different mechanisms, for example proposing a distinction between conditional statistics such as transitional probabilities, and distributional statistics such as frequency and variability (e.g. [20]). The goal of the remainder of this article is to provide a brief descriptive overview of the range of statistical learning tasks, and describe a modelling framework that can potentially capture all of these phenomena. This modelling framework is rooted in the assumption that statistical learning arises from processes endemic to memory such as activation, interference and decay.

2. Statistical learning: one mechanism or many?

Statistical learning has been studied in a variety of different tasks, including (but not limited to) word segmentation, visual scene segmentation, serial reaction time and category learning (e.g. [6,14,21,22]). In broad terms, the statistical regularities that learners detect in these kinds of tasks can be placed into two groups: conditional regularities and distributional regularities [20]. Conditional regularities refer to the likelihood of two or more elements co-occurring in the input, and can serve as a cue to grouping elements of the input into larger structures (e.g. grouping syllables into words). Transitional probabilities are often used to describe these kinds of statistical regularities. When statistical learning is studied in tasks where the learning outcome is segmenting units from the input (e.g. segmenting words from fluent speech), conditional statistical regularities are typically the statistical structure manipulated in the input. One reason that conditional statistics are such a useful cue to grouping elements into units is that they are more sensitive than simple frequency of co-occurrence. To take an example from language, while ‘the dog’ is a frequent co-occurrence, the probability of co-occurrence between these two words is quite low, as ‘the’ occurs in combination with many other words.

The other kind of statistical structure to which learners are sensitive is distributional statistical regularities. Distributional statistics refer to the frequency and variability of exemplars in the input, and can serve as a cue to group exemplars into categories. Maye *et al.*’s [14] experiments on phonemic category learning provide a paradigmatic example of the use of distributional statistics. Maye *et al.* found that infants’ grouping of exemplars along a continuum of voice onset time from /d/ to /t/ was influenced by the distribution of exemplars in the input. When exposed to a unimodal distribution, such that exemplars in the middle of the continuum occurred most frequently, infants treated the endpoints of the continuum as though they were exemplars of the same category. However, when exposed to a bimodal distribution where exemplars near the endpoints occurred most frequently, and exemplars in the middle of the distribution were relatively rare, infants treated the endpoints of the continuum (/d/ and /t/) as though they were members of different categories. As this example indicates, the distribution of exemplars in the input provides a useful cue to category membership beyond what the perceptual similarity of those exemplars indicates. Note that, as with conditional statistical learning, this sensitivity to the distribution of exemplars is not limited to linguistic input, and can be observed with many other kinds of stimuli, including visual stimuli (e.g. [23]). However, the ability to detect distributions of features may have a somewhat more different developmental time course than conditional statistical learning (e.g. [24,25]), consistent with suggestions that conditional and distributional statistical learning arise from at least partially independent processes.

Indeed, on the surface, conditional and distributional statistical regularities (and the tasks used to measure sensitivity to them) are quite distinct. Recently, these different types of tasks have been suggested to arise from separable, independent processes. For example, Endress & Bonatti [26] have demonstrated that the kinds of models that are adept at identifying word boundaries via conditional regularities are incapable of learning from distributional statistical regularities (for additional discussion, see [27]). Similarly, Thiessen *et al.* [20] argued that computational models of statistical learning focused on conditional regularities struggle to discover distributional regularities, and vice versa.

Both behavioural and neurological work support the claim that conditional and distributional statistical learning are independent processes. Behaviourally, these two types of statistical learning appear to operate under different constraints (e.g. [26,28,29]), and to mutually interfere with each other such that detecting one kind of structure impairs detection of the other kind [30]. Perhaps related to or emerging from this interference, the time course of the detection of conditional and distributional regularities also differs. After a brief exposure, adults prefer distributional regularities over conditional when the two are placed in conflict; this pattern reverses with a lengthier familiarization [27]. Neurologically, sensitivity to conditional and distributional regularities invokes different ERP responses [31,32].

3. Statistical learning as memory

Despite the evidence that statistical learning is not a unitary construct, recent work suggests that conditional and distributional learning may share a common basis in processes that are an inherent part of human memory: activation, decay, interference and prototype formation. For conditional statistical learning, one source of support for this assertion is recent neuroimaging work suggesting a role for the hippocampus in conditional statistical learning tasks [33,34]. This should not be taken to mean that the hippocampus is solely responsible for statistical learning. Both the fact that amnesiacs with hippocampal damage are capable of some forms of statistical learning (e.g. [35]), and the fact that multiple brain regions have been linked to statistical learning (e.g. [36,37]), indicate that statistical learning cannot be attributed to any single neural region or structure. Further, the activity of the hippocampus itself is complex, with different anatomical pathways that appear to represent different kinds of information [38]. Nevertheless, the involvement of the hippocampus—which has been so clearly implicated in memory formation—in statistical learning tasks suggests a relationship between memory and conditional statistical learning.

A second source of support for the linkage between conditional statistical learning and memory processes comes from modelling work demonstrating that these processes can give rise to sensitivity to conditional statistical structure. This assertion has now been modelled several times using different computational architectures (e.g. [39–41]). While these models differ somewhat in their details, they share a central assertion: the detection of statistical regularities can emerge simply from the extraction of chunks from the input. To explain this, we will focus on the model PARSER, a relatively simple exemplar memory model [42]. When exposed to a sequence of elements, PARSER randomly groups them into chunks, which it stores in memory. Over time, the activation of these chunks decays, unless they are experienced again, in which case its activation is increased. If an element within a chunk occurs within a different chunk, the previously stored chunk experiences interference and loses a degree of activation. In most cases, chunks that represent statistically coherent elements within a sequence (e.g. syllables that go together to form a word) will be experienced more frequently than elements that occur together spuriously (e.g. syllables that co-occur across word boundaries), so that the effect of interference will have a suppressing effect on the representation of these spurious groupings. Over exposure to a sequence characterized by conditional statistical relations, the model will be more likely to represent those chunks that are statistically coherent, and less likely to represent chunks whose elements have a lower probability of co-occurrence.

As such, PARSER [42] illustrates how activation, decay and interference can give rise to sensitivity to conditional statistical structure, often characterized in terms of transitional probabilities, without explicitly or implicitly calculating transitional probabilities (cf. [39]). Further, this kind of memory-based model makes a set of predictions about human performance in segmentation tasks that have largely been supported. One especially compelling demonstration of this relates to knowledge of subcomponents of words (e.g. ‘eleph’ in ‘elephant’). In a typical word segmentation task, the transitional probabilities between all of the syllables in a word are high. If learners are calculating (or representing) transitional probabilities, they should be able to differentiate these subcomponents from items with low transitional probabilities. However, when asked to differentiate between subcomponent items and items with low transitional probabilities, participants perform poorly [43,44]. This is consistent with memory-based accounts that argue that learners are extracting chunked representations, rather than calculating transitional probabilities between syllables.

Just as with conditional statistical learning, recent work has attempted to explain distributional statistical learning in terms of memory-based processes. Distributional statistical learning involves sensitivity to the central tendency of a set of exemplars, modulated by the frequency and variability of those exemplars. To model distributional statistical learning, my co-workers and I have relied on exemplar memory models of learning [20,45]. In these models, learners store prior exemplars in memory. Sensitivity to central tendency occurs because learners integrate information over these prior exemplars, such that features that are consistent across them are strengthened, and features that are inconsistent across them are weakened (e.g. [46]; though see [47] for a discussion of alternative approaches to solving this problem of discovering the central tendency of a set of exemplars via memory processes).

Our model iMinerva [45] provides a concrete example of how these memory-based processes can yield sensitivity to structure in a distributional statistical learning task. In iMinerva, a probe to memory (such as a stimulus presented during learning or test) activates prior exemplars as function of their similarity; more similar exemplars in memory are more strongly activated than less similar exemplars. If no similar prior exemplar exists, the probe to memory is stored as a new exemplar. If one or more similar exemplars exist, the probe is integrated with the prior exemplar with the greatest activation (activation varies as a function of similarity and the feature strength of the prior exemplars). This integration stores a new item in memory, one in which the features that are consistent across the two exemplars are strengthened, and the features that are inconsistent are weakened. In this way, the model eventually comes to represent a set of exemplars that are prototypical in nature—that is, reflecting the central tendency of the distributions to which it has been exposed.

To illustrate how the set of processes—activation of similar exemplars, and integration across them—can be applied to a distributional learning task, consider the experiments of Maye *et al.* [14]. Recall that in those experiments, infants discriminated to exemplars of /d/ and /t/ when exposed to a bimodal distribution, and failed to respond differentially to /d/ and /t/ when exposed to a unimodal distribution. When exposed to these distributions of exemplars, iMinerva produces a similar pattern [45]. Exposure to the unimodal distribution produces a single prototypical representation, one that is located (in similarity space) intermediate between /d/ and /t/, such that both probes to memory activate this single representation. When exposed to a bimodal distribution, the model forms two prototypes, one closer to /d/ and one closer to /t/, such that these test items activate different representations, and the model is capable of differentiating between the test items.

As is the case for conditional statistical learning, the processes invoked in this explanation of distributional statistical learning have long been thought to play a role in memory. Similarity-based activation is a feature of theories of not only prototype formation, but also priming and representation in long-term memory (e.g. [48,49]). Indeed, prototype formation is also linked to hippocampal activity (e.g. [50,51]). Of course, regions in addition to the hippocampus have been implicated in the formation and representation of prototypicality information (e.g. [52,53]). Memory is undoubtedly a distributed system, and our goal here is not to reduce the processes underlying conditional and distributional statistical learning to any single 'statistical learning box' in the brain. Rather, it is to demonstrate that both conditional and distributional statistical learning can be viewed as natural extensions of processes deeply rooted in a more general memory system.

4. Implications of a memory-based perspective

Treating statistical learning as a set of phenomena that arise from more general characteristics of memory processes has several clear and relatively novel theoretical implications (cf. [39]). First, it suggests a clear connection between statistical learning and other forms of learning. This is especially true of implicit learning, which has largely been studied in isolation from statistical learning despite exploring many of the same kinds of statistical structures (for a more extensive discussion, see [54]). This connection is strengthened by recent work demonstrating that many of the same factors that influence memory—such as massed versus spaced practice—also influence children in statistical learning tasks such as discovering categories (e.g. [55,56]). To the extent that language processing is accomplished by the same kinds of mechanisms studied in more traditional memory paradigms, it may be the case that many aspects of language learning and comprehension are constrained by the way in which stimuli are encoded, stored and accessed in memory (e.g. [57]).

Additionally, exploring the connection between statistical learning and memory provides novel insights into explaining developmental change. One conundrum faced by theoretical positions which suggest that statistical learning plays an important (though certainly not the only) role in language development (e.g. [16,58]) is that the ability to learn a new language declines with age (e.g. [59–61]). However, statistical learning can be observed across the lifespan, from infancy to adulthood (e.g. [1,62]). If statistical learning is constant across the lifespan, whereas language learning outcomes are not, then statistical learning cannot possibly help to explain age-related declines in language learning ability. Drawing a connection between memory and statistical learning may help to elucidate how statistical learning can be involved in developmental changes in learning outcomes. Though memory is present across the lifespan, it clearly undergoes developmental change; statistical learning may undergo the same kinds of developmental changes. For example, Arciuli [63] illustrates how changes to factors such as attention and processing speed may give rise to different statistical learning outcomes across developmental time, and Gomez [64] suggests that the process of memory consolidation changes with age. There are two additional kinds of age-related changes that are particularly likely to alter the function of memory, and thus of statistical learning: changes in the precision with which input is represented, and changes in the nature of the representation itself.

First, we suggest that infants represent the input in a noisier manner than do adults, such that representations of similar or identical events are more likely to differ than those of adults [65]. These noisy representations have a series of implications for infant learning. The first of these relates to generalization: when the features of the current input match a high percentage of the features of older information stored in memory, the old information is activated and influences processing of the new information (which we model via the process of integration in iMinerva). This process of integration is crucial for generalization, as it reinforces those features that are consistent across category members, and deemphasizes those features that are not. Encoding a greater number of idiosyncratic features reduces the likelihood that two exemplars related to the same category or central concept will activate each other. This, in turn, decreases a learner's ability to identify features that are common across the exemplars and generalize that commonality to novel exemplars and contexts [45]. Taken to the extreme, such a tendency would make it impossible to learn. Fortunately, the processes of consolidation illustrated by Gomez [64] help to ensure that commonalities across exemplars are reinforced over time.

Second, we suggest that compared with adults, infants are more likely to encode—or weight more heavily—features of the input that are irrelevant [65]. For example, when learning words, infants are more likely to represent irrelevant information such as indexical characteristics of the speaker in addition to, or perhaps at the expense of, linguistically relevant features such as phonemic identity (e.g. [66–68]). In part, this is due to the fact that infants are not yet familiar with the structure of the environment; much like adult novices in a domain, they lack the knowledge to focus on the relevant features of the input (e.g. [69]). Additionally, some part of this encoding of irrelevant features is due to the fact that infants are less able to control the focus of their attention than are older children and adults (e.g. [70]). This developmental change has been linked to the development of the pre-frontal cortex (e.g. [71]). The ability to selectively attend to an object or element of the environment, and maintain attention on the target, develops dramatically across the first 5 years of life, and clearly shapes many forms of learning, including statistical learning (e.g. [72,73]). As with infants' lack of knowledge, their relative inability to control attention means that their experiences will generate representations that are highly idiosyncratic and lower in similarity across similar instances than would be expected of older learners.

As discussed above, these kinds of idiosyncratic representations are likely to slow learning. But they may also have a positive effect. Representations containing a greater weight—relative to adult learners—on idiosyncratic features may also help infants with the process of discovering features that are relevant to the statistical structure of the environment. Across any set of N exemplars characterized by both common features and idiosyncratic features, common features are more likely to 'survive' being encoded in the presence of noise, and these common features are more likely to be strengthened by integration with other exemplars [46,74]. This is because idiosyncratic features are present across fewer members of the set, so are more likely to be erased or altered by noisy, inaccurate encoding. As such, immature encoding may actually accentuate the commonalities across a set of exemplars, leading to greater likelihood of detecting the consistent features that characterize them (cf. [75]). By contrast, encodings that are weighted toward a particular set of features—as is the case for adults who have discovered the regularities of a domain such as language—are less likely to identify novel commonalities across a set of exemplars, especially commonalities that contradict the regularities they have learned. Thus, while early learning is slower, it is perhaps better adapted to discover structure in novel environments. Conversely, once infants have identified a set of relevant features that characterize the input, they are less likely to generalize over features of the input that have been irrelevant in their prior experience (e.g. [76,77]).

From this perspective, the developmental change in learning can be characterized as a transition from an early state of learning that is slow but flexible enough to adapt to many environments, to a later state that is more efficient but also more constrained by the regularities of the environment. Language development provides a paradigmatic example of this shift. Early in life, infants have the ability to perceive most, if not all, of the phonemic contrasts used in the world's languages (e.g. [78]). As infants acquire experience with their native language, their representations tune to the phonemic contrasts that are used in that language, causing a loss of sensitivity to non-native phonemic contrasts (e.g. [79]). This adaptation is a double-edged sword. The loss of sensitivity to non-native contrasts is associated with an increase in sensitivity to contrasts used in the native language, and which facilitates subsequent learning of the language [80,81], but which impairs learning in languages where perception of the non-native contrasts would be useful. This pattern of early, flexible giving way to more efficient and specialized learning can be seen in many aspects of language acquisition beyond phonemic perception, including phonotactics, syntax and phonology (e.g. [82–84]).

Perhaps, the most important implication of a memory-based perspective on statistical learning, however, is the suggestion that both conditional and distributional statistical learning share at least some underlying processes in common. That is, from this memory-based perspective, many of the same mechanisms—such as similarity-based activation, integration, decay and interference—play a role in learning both conditional and distributional regularities. That is, while statistical learning has been studied in an incredible variety of tasks, including (but not limited to) word segmentation, serial reaction time, category learning, phonotactic learning, and visual scene segmentation, and learning about the social world, it may be possible to explain all of these tasks by appealing to a relatively limited set of underlying processes. Similarly, it may be possible to model all of these tasks using a common computational framework. Our model iMinerva is a step in this direction, as it provides an existence proof that several different statistical learning tasks can be modelled using the same memory-based computational framework [45].

5. How a memory-based approach differs from a probability-based approach

My co-workers and I have suggested that both conditional and distributional learning arise from a limited set of underlying memory process [20,45,58]. Conditional statistical learning arises from extracting exemplars from the input, and the processes of activation, decay and interference. Distributional statistical learning arises from the integration of information across these stored exemplars. We suggest that conditional and distributional statistical learning are deeply linked, both in terms of sharing at least a partially overlapping set of common underlying processes, and in their interaction in learning. The output of conditional statistical learning (a set of stored exemplars) provides the input to distributional statistical learning, and the output of distributional statistical learning (learned regularities about the environment) influences the exemplars that are subsequently extracted from the input (e.g. [84]).

Note that from this perspective, while learners are sensitive to the statistical structure of the environment, they are not learning by explicitly or implicitly calculating probability. In this regard, a memory-based perspective differs profoundly from accounts that describe statistical learning in terms of 'sensitivity to' or 'calculation of' transitional probabilities. Transitional probabilities, as described previously, are a popular metric used to describe the statistical structure in many conditional statistical learning tasks. However, from a memory-based perspective, these probabilities lack psychological reality.

There are both theoretical and empirical reasons to doubt the psychological reality of transitional probabilities. Theoretically, transitional probabilities provide a very specialized account, capable of explaining a small proportion of the wide variety of statistical learning tasks in the literature. They provide no explanation for distributional statistical learning (for a more extensive discussion, see [20]). They are not even useful for explaining all forms of conditional statistical learning; both simultaneously presented conditional statistics and non-adjacent conditional statistics present difficulties for a transitional-probability approach. Transitional probabilities are explicitly about the transition from one element to a subsequent element. As such, they are not informative about conditional relations among simultaneously presented elements. For example, in Fiser & Aslin's [6,13] visual scene segmentation experiments, learners are presented with complex shapes such that multiple elements are presented simultaneously. Learners detect that some of these elements are likely to co-occur, but this cannot be due to a transition from element X to element Y, as they are always co-present. Sensitivity to non-adjacent transitional probabilities presents different problems for perspectives suggesting that learners calculate transitional probabilities. Learners are sensitive not just to conditional relations among adjacent elements, but also to relations among non-adjacent elements (e.g. [85,86]). That is, when presented with sequences like AXB, AYB and AZB, learners detect the relationship between A and B even though these items are not directly adjacent. If learners are computing all possible transitional probabilities, including non-adjacent ones, the computational demands on learning become quite high. A similar problem is presented by the demonstration that learners are sensitive to both backward and forward transitional probability [87]. Chunk-based accounts elegantly solve this problem because coherent items are preserved whether that coherence arises from forward-going or backward-going regularities [88].

In addition to these theoretical concerns, empirical data suggest that the calculation of transitional probability does not provide a good fit to human learning. First, transitional probabilities are insensitive to frequency. That is, if XY occurs twice, and X occurs twice, the transitional probability between X and Y is 100%. If XY occurs 100 times, and X occurs 100 times, the transitional probability between X and Y is 100%. If learners' sensitivity to conditional probability arises from their computation of transitional probability, their confidence in grouping X and Y should be equivalent in these two cases (as the transitional probability between X and Y is identical in both cases). But in fact, learners have a stronger expectation that X and Y will co-occur as they see the grouping more frequently [89]. This sensitivity to frequency falls naturally out of a memory-based perspective, given the important role of repetition in memory (e.g. [90]).

Second, transitional probabilities are insensitive to the order in which information is presented. That is, among a stream with elements A, B, C and D, if there are high transitional probabilities between A and B and C and D, learners should learn to group AB and CD regardless of the order in which these elements are presented. By contrast, order effects are one of the signature empirical phenomena associated with the memory literature. The temporal order of presentation has an influence both in terms of primacy and recency effects, and in terms of massed and spaced practice. A memory-based approach to learning suggests that these temporal order phenomena should also influence performance in statistical learning tasks. This prediction falls out of both the wider literature on memory, and from the characteristics of memory-based models such as iMinerva, which is sensitive to the order in which information is presented because early representations constrain the way in which the learner responds to subsequent information [74].

A fuller review of the literature on the effects of spacing, frequency and order in both the statistical learning literature (for discussion, see [20,89]) and the memory literature (for discussion, see [91,92]) is beyond the scope of the current discussion. However, the central point is straightforward: probability metrics provide a useful way of *describing* the statistical structure of the input. But theoretical accounts that posit some psychological reality to these computations, either explicitly or implicitly, fail to account for profoundly important considerations, frequency and order effects that fall naturally out of a memory-based account. Proponents of such probability-based accounts might argue that memory effects can be built 'on top of' probability calculations (e.g. [89]). But this approach seems excessively ornate; if it is the case that memory processes themselves can account for sensitivity to statistical structure, then positing that humans can calculate such statistics is unnecessary.

Regardless of the differences in these accounts, both suggest that exploring statistical learning in the context of memory is likely to be a fruitful endeavour. This is consistent with modelling work, which—as discussed above—has largely suggested that a memory-based approach to statistical learning accounts for a wider variety of data than models based on transitional probability (e.g. [40,45,93]). Nevertheless, none of these models provide a precise fit to human behavioural data [89]. As we will discuss below, this suggests that theoretical accounts drawing connections between statistical learning and memory are still in need of improvement.

6. Challenges for a memory-based perspective

While recent research has provided compelling evidence of the link between statistical learning and memory, several challenges remain before this link can be considered fully substantiated. One of the most important of these challenges is to connect statistical learning to process models of memory. Like statistical learning, memory is not a monolithic construct; it is accomplished by a number of processes and neurological substrates. Drawing connections between aspects of statistical learning and aspects of memory can serve to improve our understanding of both. In particular, the proposed distinction between conditional and distributional statistical learning would be more firmly substantiated if it could be tied to distinctions in the human memory system.

One possibility, originally proposed by McClelland *et al.* [52] is that the distinction between conditional and distributional statistical learning maps onto the distinction between hippocampal and cortical memory systems. In this view, the hippocampus is responsible for extracting items (exemplar memory) while the cortex is responsible for integrating across these items (via prototype formation). One criticism of this approach is that the cortical memory system is usually thought to act more slowly than the hippocampal system, which would predict that distributional statistical learning would be slower than conditional statistical learning. Behavioural evidence, however, suggests that both forms of statistical learning are quite rapid, and that at least in some cases distributional statistical learning proceeds more quickly (e.g. [26]). One intriguing response to this mismatch, suggested by Schapiro [38] is that the hippocampus may be more prominently involved in some aspects of distributional statistical learning.

Another possibility is that conditional and distributional statistical learning arise from different potential encodings of the input [27]. Sensitivity to transitional probabilities arises, from this perspective, when participants encode the input in terms of chains of elements (e.g. A-B-C). This kind of encoding is insensitive to the position at which each element occurs, and instead represents only their serial order (the fact that A preceded B, irrespective of which absolute position that occurred in). Sensitivity to distributional information may arise when participants encode the absolute position of items in the sequence (especially first and last position), an encoding that is easier when there are breaks or pauses in the input that provide cues to element boundaries. Together, McClelland *et al.*'s [52] and Endress & Bonatti's [27] approaches highlight two important distinctions within the memory system—neurological substrate and encoding—that may explain distinctions within statistical learning. But note that not only are there several possible variations on neurological substrate (e.g. differential function within the hippocampus; e.g. [94]) and on possible encodings, there are several additional distinctions within the memory system, such as the difference between implicit and explicit memory, that may additionally provide some explanation for differences across statistical learning tasks.

If the premise that memory processes can explain statistical learning is correct, then models that instantiate these processes should be able to simulate human performance in statistical learning tasks. To an extent, this argument has already been supported by the existence of memory-based models such as PARSER [42], TRACX [39], TRACX2 [40] and iMinerva [45], and various connectionist models such as TRACE [95,96] that simulate sensitivity to statistical structure. However, most of these models have only been applied to a small number of statistical learning tasks, or even simply attempted to model a single task. Given the overlap in shared processes underlying conditional and distributional statistical learning (and, in parallel, the overlap in process and neurology across different aspects of memory), it should be possible to develop a model that simulates performance in the whole variety of statistical learning tasks that can be characterized as requiring sensitivity to either conditional or distributional structure.

iMinerva [45,65] provides one route toward doing this. This model is an extension of classic exemplar memory models such as Hintzman's MINERVA [97], and has been used to successfully simulate statistical learning in a variety of distributional learning tasks such as category learning and acquired distinctiveness training (e.g. [14,15]). But the processes simulated by iMinerva—similarity-based activation, interference and decay—are also, according to a memory-based framework, at least partially responsible for conditional statistical learning. If this is the case, it should be possible for the iMinerva architecture to simulate performance in conditional statistical learning tasks. This extension is not simply a formal nicety, or an important step in formalizing a theory and generating novel predictions; it is also a necessary challenge for any theoretical perspective arguing that conditional and distributional learning share some (or all) underlying processes. As such, the modelling effort to formalize this memory-based perspective serves as a necessary existence proof of the validity of this perspective.

A related challenge for this perspective is to understand individual variability in performance in statistical learning tasks. For a memory-based account of statistical learning to be successful, it must be able to account for the pattern of individual differences in statistical learning tasks, both within individuals and over developmental time. In general, individual performance in one statistical learning task is only slightly correlated (and in some cases uncorrelated) with performance in another statistical learning task (e.g. [98]). On its surface, this pattern of results is inconsistent with theories which suggest that performance in different statistical learning tasks is determined by partially shared underlying memory processes.

However, there are two caveats worth noting. The first is that the psychometric characteristics of many statistical learning tasks may not be especially useful for explorations of individual differences [99]. Many of these tasks were designed to probe learning at the group level, rather than to assess individual differences in learning outcomes. For example, as Siegelman & Frost [98] point out, many statistical word segmentation tasks use two-alternative forced choice tasks rather than three-alternative forced choice tasks, resulting in a smaller range of performance between no learning (50% in 2AFC, 33% in 3AFC) and learning ranges of performance. Because of design decisions favouring group-level over individual-level analyses, these tasks may have low validity as measures of individual difference. As such, some portion of the low correlation observed between performance in different statistical learning tasks may be due to the reliability of the measures themselves. More sensitive measures may uncover different patterns of performance across tasks; at the very least, they will provide more informative data.

Second, as discussed above, the human memory system is not a monolithic construct. Even primarily cognitive models of memory suggest important divisions in the memory system, such as between short- and long-term memory, or between implicit and explicit memory (e.g. [100,101]). Neurological work has demonstrated that these cognitive divisions, if anything, underrepresent the complexity of memory (e.g. [102]). Consider as an example the demonstration that rate of presentation has differential effects in auditory and visual statistical learning [103]. One possibility is that this differential effect arises because there are different processes underlying visual and audio statistical learning. Another possibility is that the same processes are at work in learning both kinds of material, but operating under a different set of constraints as a function of the input (for discussion, see Siegelman [99]). Visual information activates different representations than does audio information, representations that are supported by different regions of cortex and defined by different peripheral processing constraints. Rate of presentation has different effects for memory for visual and auditory objects (e.g. [104]), but there are doubtless important commonalities in the way audio and visual information are stored in memory; the same may be true of learning statistical regularities in these perceptual domains. That is, the low correlation between statistical learning tasks may arise not only due to the poor psychometric properties of these tasks, but also reflect important distinctions between statistical learning tasks either as a function of which aspect of the memory system they draw on, or as a function of the characteristics of the salience, encoding or distinctiveness of the stimuli (cf. [105]). This discussion highlights the importance of modelling efforts. Given differing, sometimes underspecified definitions of the nature of statistical learning, it is difficult to satisfactorily resolve disputes about when differences reflect qualitative differences in underlying processes, and when they reflect distinctions in more peripheral systems supporting statistical learning, such as attention and perception.

7. Conclusion

Statistical learning has been suggested to play an important role in many aspects of development and cognition, including language learning, maths learning, decision-making and social interaction. Across each of these domains, and many others, authors have advanced convincing demonstrations that humans, and other animal species, are sensitive to the statistical structure of the environment. While these demonstrations of the breadth of statistical learning have been impressive, they bring into sharp relief a set of questions about the nature of the processes underlying statistical learning: how can the same mechanism be responsible for so many different kinds of learning? How is this sensitivity accomplished over so many different stimulus types, tasks and time frames? The fact that we have no settled answers for these questions suggests that our understanding of statistical learning is less clear than our ability to demonstrate its existence.

Nevertheless, recent modelling and neuroscience work suggests a compelling avenue for answering these questions in terms of explicating the relationship between statistical learning and memory. Modelling research has demonstrated that processes thought to be integral aspects of memory are also capable of giving rise to sensitivity to statistical structure (e.g. [42,45]). Neuroscience work has demonstrated that regions thought to be critical for memory, such as the hippocampus, also play a role in statistical learning (e.g. [33,34,50,51]). Exploring this connection has the potential to define the mechanisms or processes underlying statistical learning, as well as reveal connections between statistical learning and many other forms of learning thought to be rooted in the characteristics of memory, such as state-dependent learning, paired-associate learning and implicit learning.

Competing interests. I have no competing interests.

Funding. This work was supported by NSF grant BCS-1451706.

Acknowledgements. Thank you to Morten Christiansen and three anonymous reviewers for helpful comments on a prior version of this manuscript, as well as to Lucy Erickson, Sandrine Girard, Nick Turk-Browne and Blair Armstrong for talking through some of the early ideas that worked their way into this paper, and to Kristen Boyle for copyediting.

References

1. Saffran JR, Aslin RN, Newport EL. 1996 Statistical learning of 8-month-old infants. *Science* **274**, 1926–1928. (doi:10.1126/science.274.5294.1926)
2. Aslin RN, Saffran JR, Newport EL. 1998 Computation of conditional probability statistics by 8-month-old infants. *Psychol. Sci.* **9**, 321–324. (doi:10.1111/1467-9280.00063)
3. Saffran JR, Johnson EK, Aslin RN, Newport EL. 1999 Statistical learning of tone sequences by human infants and adults. *Cognition* **70**, 27–52. (doi:10.1016/S0010-0277(98)00075-4)
4. Baldwin D, Andersson A, Saffran J, Meyer M. 2008 Segmenting dynamic human action via statistical structure. *Cognition* **106**, 1382–1407. (doi:10.1016/j.cognition.2007.07.005)
5. Brady TF, Oliva A. 2008 Statistical learning using real-world scenes extracting categorical regularities without conscious intent.
6. Fiser J, Aslin RN. 2002 Statistical learning of new visual feature combinations by infants. *Proc. Natl Acad. Sci. USA* **99**, 15 822–15 826. (doi:10.1073/pnas.232472899)
7. Thiessen ED. 2010 Effects of visual information on adults' and infants' auditory statistical learning. *Cogn. Sci.* **34**, 1093–1106. (doi:10.1111/j.1551-6709.2010.01118.x)
8. Teinonen T, Fellman V, Naatanen R, Alku P, Huotilainen M. 2009 Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience* **10**, 21. (doi:10.1186/1471-2202-10-21)
9. Cherry KE, Stadler MA. 1995 Implicit learning of a nonverbal sequence in younger and older adults. *Psychol. Aging* **10**, 379–394. (doi:10.1037/0882-7974.10.3.379)
10. Conway CM, Christiansen MH. 2001 Sequential learning in non-human primates. *Trends Cogn. Sci.* **12**, 539–546. (doi:10.1016/S1364-6613(00)01800-3)
11. Hauser MD, Newport E, Aslin RN. 2001 Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition* **78**, B53–B64. (doi:10.1016/S0010-0277(00)00132-3)
12. Toro JM, Trobalón JB. 2005 Statistical computations over a speech stream in a rodent. *Percept. Psychophys.* **67**, 867–875. (doi:10.3758/BF03193539)
13. Fiser J, Aslin RN. 2001 Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **28**, 458–467. (doi:10.1037/0278-7393.28.3.458)
14. Maye J, Werker JF, Gerken L. 2002 Infant sensitivity to distri- butional information can affect phonetic discrimination. *Cognition* **82**, B101–B111. (doi:10.1016/S0010-0277(01)00157-3)
15. Thiessen ED. 2007 The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language* **56**, 16–34. (doi:10.1016/j.jml.2006.07.002)
16. Romberg AR, Saffran JR. 2010 Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* **6**, 906–914. (doi:10.1002/wcs.78)
17. Thiessen ED, Girard S, Erickson LC. 2016 Statistical learning and the critical period: how a continuous learning mechanism can give rise to discontinuous learning. *Wiley Interdiscip. Rev. Cogn. Sci.* **7**, 276–288. (doi:10.1002/wcs.1394)
18. Conway CM, Christiansen MH. 2006 Statistical learning within and between modalities. *Psychol. Sci.* **17**, 905–912. (doi:10.1111/j.1467-9280.2006.01801.x)
19. Emberson LL, Conway CM, Christiansen MH. 2011 Timing is everything: changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Q. J. Exp. Psychol.* **64**, 1021–1040. (doi:10.1080/17470218.2010.538972)
20. Thiessen ED, Kronstein AT, Hufnagle DG. 2013 The extraction and integration framework: a two-process account of statistical learning. *Psychological Bulletin* **139**, 792–814. (doi:10.1037/a0030801)
21. Thiessen ED, Saffran JR. 2003 When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Dev. Psychol.* **39**, 706–716. (doi:10.1037/0012-1649.39.4.706)
22. Hunt RH, Aslin RN. 2001 Statistical learning in a serial reaction time task: access to separable statistical cues by individual learners. *J. Exp. Psychol. Gen.* **130**, 658–680. (doi:10.1037/0096-3445.130.4.658)
23. French RM, Mareschal D, Quinn PC. 2004 The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: simulations and data. *Journal of Experimental Psychology: General* **133**, 382–397. (doi:10.1037/0096-3445.133.3.382)
24. Marchetto E, Bonatti LL. 2013 Words and possible words in early language acquisition. *Cognitive Psychology* **67**, 130–150. (doi:10.1016/j.cogpsych.2013.08.001)
25. Younger BA, Hollich G, Furrer SD. 2004 An emerging consensus: Younger and Cohen revisited. *Infancy* **5**, 209–216. (doi:10.1207/s15327078in0502_6)
26. Endress AD, Bonatti LL. 2007 Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition* **105**, 247–299. (doi:10.1016/j.cognition.2006.09.010)
27. Endress AD, Bonatti LL. 2016 Words, rules, and mechanisms of language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* **7**, 19–35. (doi:10.1002/wcs.1376)
28. Endress AD, Mehler J. 2009 Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology* **62**, 2187–2209. (doi:10.1080/17470210902783646)
29. Pena M, Bonatti LL, Nespor M, Mehler J. 2002 Signal-driven computations in speech processing. *Science* **298**, 604–607. (doi:10.1126/science.1072901)
30. Zhao J, Ngo N, McKendrick R, Turk-Browne NB. 2011 Mutual interference between statistical summary perception and statistical learning. *Psychological Science* **22**, 1212–1219. (doi:10.1177/0956797611419304)
31. de Diego Balagauer R, Toro JM, Rodríguez-Fornells A, Bachoud-Levi AC. 2007 Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS ONE* **14**, e1175. (doi:10.1371/journal.pone.0001175)
32. Mueller JL, Bahlmann J, Friederici AD. 2008 The role of pause cues in language learning: the emergence of event-related potentials related to sequence processing. *Journal of Cognitive Neuroscience* **20**, 892–905. (doi:10.1162/jocn.2008.20511)
33. Hindy NC, Ng FY, Turk-Browne NB. 2016 Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience* **19**, 665–667. (doi:10.1038/nn.4284)
34. Schapiro AC, Gregory E, Landau B, McCloskey M, Turk-Browne NB. 2014 The necessity of medial temporal lobe for statistical learning. *J. Cogn. Neurosci.* **26**, 1736–1747. (doi:10.1162/jocn_a_00578)
35. Cohen NJ, Eichenbaum H. 1993 *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.
36. Abla D, Okinoya K. 2008 Statistical segmentation of tone sequences activates the left inferior frontal cortex: a near-infrared spectroscopy study. *Neuropsychologia* **46**, 2787–2795. (doi:10.1016/j.neuropsychologia.2008.05.012)
37. Karuza EA, Newport EL, Aslin RN, Starling SJ, Tivarus ME, Bavelier D. 2013 The neural correlates of statistical learning in a word segmentation task: an fMRI study. *Brain and Language* **127**, 46–54. (doi:10.1016/j.bandl.2012.11.007)
38. Schapiro AC, Turk-Browne NB, Botvinick MM, Norman KA. 2017 Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Phil. Trans. R. Soc. B* **372**, 20160049. (doi:10.1098/rstb.2016.0049)
39. French RM, Addyman C, Mareschal D. 2011 TRACX: a recognition-based connectionist framework for the sequence segmentation and chunk extraction. *Psychol. Rev.* **118**, 614–636. (doi:10.1037/a0025255)
40. Mareschal D, French RM. 2017 TRACX2: a connectionist autoencoder using graded chunks to

model infant visual statistical learning. *Phil. Trans. R. Soc. B* **372**, 20160057. (doi:10.1098/rstb.2016.0057)

41. Shi L, Griffiths T, Feldman NH, Sanborn AN. 2010 Exemplar models as a mechanism for performing Bayesian inferenc. *Psychonomic Bulletin and Review* **17**, 443–464. (doi:10.3758/PBR.17.4.443)
42. Perruchet P, Vinter A. 1998 PARSER: a model for word segmentation. *J. Memory Lang*. **39**, 246–263. (doi:10.1006/jmla.1998.2576)
43. Fiser J, Aslin RN. 2005 Encoding multielement scenes: statistical learning of visual feature hierarchies. *J. Exp. Psychol. Gen*. **134**, 521–537. (doi:10.1037/0096-3445.134.4.521)
44. Giroux I, Rey A. 2009 Lexical and sublexical units in speech perception. *Cogn. Sci*. **33**, 260–272. (doi:10.1111/j.1551-6709.2009.01012.x)
45. Thiessen ED, Pavlik PI. 2013 iMinerva: a mathematical model of distributional statistical learning. *Cogn. Sci*. **37**, 310–343. (doi:10.1111/cogs.12011)
46. Hintzman DL. 1984 MINERVA 2: a simulation model of human memory. *Behav. Res. Methods* **16**, 96–101. (doi:10.3758/BF03202365)
47. Altmann GTM. 2017 Abstraction and generalization in statistical learning: implications for the relationship between semantic types and episodic tokens. *Phil. Trans. R. Soc. B* **372**, 20160060. (doi:10.1098/rstb.2016.0060)
48. Collins AM, Quillian MR. 1969 Retrieval time from semantic memory. *J. Verbal Learning Verbal Behav*. **8**, 240–247. (doi:10.1016/S0022-5371(69)80069-1)
49. Meyer DE, Schvaneveldt RW. 1971 Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *J. Exp. Psychol*. **90**, 227–234. (doi:10.1037/h0031564)
50. Cabeza R, Rao SM, Wagner AD, Mayer AR, Schacter DL. 2001 Can medial temporal lobe distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proceedings of the National Academies of Science* **98**, 4805–4810. (doi:10.1073/pnas.081082698)
51. Zeithamova D, Maddox WT, Schnyer DM. 2008 Dissociable prototype learning systems: evidence from brain imaging and behavior. *J. Neurosci*. **28**, 13 194–13 201. (doi:10.1523/JNEUROSCI.2915-08.2008)
52. McClelland JL, McNaughton BL, O'Reilly RC. 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. *Psychol. Rev*. **102**, 419–457. (doi:10.1037/0033-295X.102.3.419)
53. O'Reilly RC, Bhattacharyya R, Howard MD, Ketz N. 2014 Complementary learning systems. *Cogn. Sci*. **38**, 1229–1248. (doi:10.1111/j.1551-6709.2011.01214.x)
54. Perruchet P, Pacton S. 2006 Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Science* **10**, 233–238. (doi:10.1016/j.tics.2006.03.006)
55. Vlach HA. 2014 The spacing effect in children's generalization of knowledge: allowing children time to forget promotes their ability to learn. *Child Dev. Perspect*. **8**, 163–168. (doi:10.1111/cdep.12079)
56. Vlach HA, Sandhofer CM, Bjork RA. 2014 Equal spacing and expanding schedules in children's categorization and generalization. *J. Exp. Child Psychol*. **123**, 129–137. (doi:10.1016/j.jecp.2014.01.004)
57. Christiansen MH, Chater N. 2016 The Now-or-Never bottleneck: a fundamental constraint on language. *Behav. Brain Sci*. **39**, e62.
58. Erickson LC, Thiessen ED. 2015 Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Dev. Rev*. **37**, 66–108. (doi:10.1016/j.dr.2015.05.002)
59. Best CT. 1994 The emergence of native-language phonological influences in infants: a perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words* **167**, 224.
60. Flege JE. 1995 Second language speech learning: theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language* **Research**, 233–277.
61. Johnson JS, Newport EL. 1989 Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* **21**, 60–99. (doi:10.1016/0010-0285(89)90003-0)
62. Hayes JR, Clark HH. 1970 Experiments on the segmentation of an artificial speech analogue. In JR Hayes (Ed.) *Cognition and the development of language* (pp. 221–234). New York, NY: Wiley.
63. Arciuli J. 2017 The multi-component nature of statistical learning. *Phil. Trans. R. Soc. B* **372**, 20160058. (doi:10.1098/rstb.2016.0058)
64. Gómez RL. 2017 Do infants retain the statistics of a statistical learning experience? Insights from a developmental cognitive neuroscience perspective. *Phil. Trans. R. Soc. B* **372**, 20160054. (doi:10.1098/rstb.2016.0054)
65. Thiessen ED, Pavlik PI. 2016 Modeling the role of distributional information in children's use of phonemic contrasts. *Journal of Memory and Language* **88**, 117–132. (doi:10.1016/j.jml.2016.01.003)
66. Houston DM, Jusczyk PW. 2000 The role of talker-specific information in word segmentation by infants. *J. Exp. Psychol. Hum. Percept. Perform*. **26**, 1570. (doi:10.1037/0096-1523.26.5.1570)
67. Newman RS. 2008 The level of detail in infants' word learning. *Curr. Direct. Psychol. Sci*. **17**, 229–232. (doi:10.1111/j.1467-8721.2008.00580.x)
68. Singh L. 2008 Influences of high and low variability on infant word recognition. *Cognition* **106**, 833–870. (doi:10.1016/j.cognition.2007.05.002)
69. Chi MT, Feltovich PJ, Glaser R. 1981 Categorization and representation of physics problems by experts and novices. *Cognitive Science* **5**, 121–152. (doi:10.1207/s15516709cog0502_2)
70. Oakes LM, Kannass KN, Shaddy DJ. 2002 Developmental changes in endogenous control of attention: the role of target familiarity on infants' distraction latency. *Child Development* **73**, 1644–1655. (doi:10.1111/1467-8624.00496)
71. Diamond A. 2000 Toward an understanding of the human frontal lobes. *Contemporary Psychology: APA Review of Books* **45**, 564–565. (doi:10.1037/002307)
72. Baker CL, Olson CR, Behrmann M. 2004 Role of attention and perceptual grouping in visual statistical learning. *Psychological Science* **15**, 460–466. (doi:10.1111/j.0956-7976.2004.00702.x)
73. Erickson LC, Thiessen ED, Godwin KG, Dickerson JP, Fisher AV. 2015 Endogenously—but not exogenously—driven selective attention predicts classroom learning in kindergarten children. *Journal of Experimental Child Psychology* **114**, 275–294.
74. Thiessen ED, Pavlik P. 2012 iMinerva: a mathematical model of distributional statistical learning. *Cognitive Science* **37**, 310–343.
75. Elman JL. 1993 Learning and development in neural networks: the importance of starting small. *Cognition* **48**, 71–99. (doi:10.1016/0010-0277(93)90058-4)
76. Dawson C, Gerken L. 2009 From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition* **111**, 378–382. (doi:10.1016/j.cognition.2009.02.010)
77. Gerken L, Bollt A. 2008 Three exemplars allow at least some linguistic generalizations: implications for generalization mechanisms and constraints. *Lang. Learn. Dev*. **4**, 228–248. (doi:10.1080/15475440802143117)
78. Cheour M, Ceponiene R, Lehtokoski A, Luuk A, Allik J, Alho K, Näätänen R. 1998 Development of language-specific phoneme representations in the infant brain. *Nat. Neurosci*. **1**, 351–353. (doi:10.1038/1561)
79. Werker JF, Tees RC. 1984 Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev*. **7**, 49–63. (doi:10.1016/S0163-6383(84)80022-3)
80. Kuhl PK, Stevens E, Hayashi A, Deguchi T, Kiritani S, Iverson P. 2006 Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. Sci*. **9**, F13–F21. (doi:10.1111/j.1467-7687.2006.00468.x)
81. Tsao FM, Liu HM, Kuhl PK. 2004 Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Dev*. **75**, 1067–1084. (doi:10.1111/j.1467-8624.2004.00726.x)
82. Estes KG, Edwards J, Saffran JR. 2011 Phonotactic constraints on infant word learning. *Infancy* **16**, 180–197. (doi:10.1111/j.1532-7078.2010.00046.x)
83. Onnis L, Thiessen ED. 2014 Language experience changes subsequent statistical learning. *Cognition* **126**, 268–284. (doi:10.1016/j.cognition.2012.10.008)
84. Thiessen ED, Saffran JR. 2007 Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Lang. Learn. Dev*. **3**, 73–100.
85. Gómez RL. 2002 Variability and detection of invariant structure. *Psychol. Sci*. **13**, 431–436. (doi:10.1111/1467-9280.00476)
86. Newport EL, Aslin RN. 2004 Learning at distance I. Statistical learning of non-adjacent dependencies. *Cognit. Psychol*. **48**, 127–162. (doi:10.1016/S0010-0285(03)00128-2)

87. Pelluchi B, Hay JF, Saffran JR. 2009 Learning in reverse: eight-month-old infants track backward transitional probabilities. *Cognition* **113**, 244–247. (doi:10.1016/j.cognition.2009.07.011)

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

89. Frank MC, Goldwater S, Griffiths T, Tenenbaum JB. 2010 Modeling human performance in statistical word segmentation. *Cognition* **117**, 107–125. (doi:10.1016/j.cognition.2010.07.005)

90. Atkinson RC, Shiffrin RM. 1968 Human memory: a proposed system and its control processes. In *The Psychology of Learning and Motivation* (eds KW Spence, JT Spence), pp. 89–195. New York, NY: Academic Press.

91. Burgess N, Hitch GJ. 1999 Memory for serial order: a network model of the phonological loop and its timing. *Psychol. Rev*. **106**, 551–581. (doi:10.1037/0033-295X.106.3.551)

92. Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. 2006 Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychol. Bull*. **132**, 354–380. (doi:10.1037/0033-2909.132.3.354)

93. Orban G, Fiser J, Aslin RN, Lengyel M. 2008 Bayesian learning of visual chunks by human observers. *Proc. Natl Acad. Sci. USA* **105**, 2745–2750. (doi:10.1073/pnas.0708424105)

94. Ji J, Maren S. 2008 Differential roles for hippocampal areas CA1 and CA3 in encoding and retrieval of extinguished fear. *Learn. Memory* **15**, 244–251. (doi:10.1101/lm.794808)

95. McClelland JL, Elman JL. 1986 The TRACE model of speech perception. *Cognitive Psychology* **18**, 1–86.

96. McClelland JL, Mirman D, Holt LL. 2006 Are there interactive processes in speech perception? *Trends in Cognitive Science* **10**, 363–369.

97. Hintzman DL. 1984 MINERVA 2: a simulation model of human memory. *Behavior Research Methods, Instruments, and Computers* **16**, 96–101.

98. Siegelman N, Frost R. 2015 Statistical learning as an individual ability: theoretical perspectives and empirical evidence. *J. Memory Lang*. **81**, 105–120. (doi:10.1016/j.jml.2015.02.001)

99. Siegelman N, Bogaerts L, Christiansen MH, Frost R. 2017 Towards a theory of individual differences in statistical learning. *Phil. Trans. R. Soc. B* **372**, 20160059. (doi:10.1098/rstb.2016.0059)

100. Baddeley, Hitch. 1974.

101. Squire. 1987.

102. Eichenbaum. 1997.

103. Conway CM, Christiansen MH. 2009 Seeing and hearing in space and time: effects of modality and presentation rate on implicit statistical learning. *Eur. J. Cogn. Psychol*. **21**, 561–580. (doi:10.1080/09541440802097951)

104. Frick RW. 1985 Testing visual short-term memory: simultaneous and sequential presentations. *Memory Cogn*. **13**, 346–356. (doi:10.3758/BF03202502)

105. Frost R, Armstrong BC, Siegelman N, Christiansen MH. 2015 Domain generality versus modality specificity: the paradox of statistical learning. *Trends Cogn. Sci*. **19**, 117–125. (doi:10.1016/j.tics.2014.12.010)