Published in final edited form as:
Neurobiol Learn Mem. 2014 March ; 0: 193–206. doi:10.1016/j.nlm.2013.09.012.

Combining fMRI and Behavioral Measures to Examine the Process of Human Learning

Elisabeth A. Karuza[1,*], **Lauren L. Emberson**[1,*], and **Richard N. Aslin**[1]
[1]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA

Abstract

Prior to the advent of fMRI, the primary means of examining the mechanisms underlying learning were restricted to studying human behavior and non-human neural systems. However, recent advances in neuroimaging technology have enabled the concurrent study of human behavior and neural activity. We propose that the integration of behavioral response with brain activity provides a powerful method of investigating the process through which internal representations are formed or changed. Nevertheless, a review of the literature reveals that many fMRI studies of learning either (1) focus on outcome rather than process or (2) are built on the untested assumption that learning unfolds uniformly over time. We discuss here various challenges faced by the field and highlight studies that have begun to address them. In doing so, we aim to encourage more research that examines the process of learning by considering the interrelation of behavioral measures and fMRI recording during learning.

Keywords

fMRI; cognitive neuroscience; learning; process; time-course; outcome measures; behavioral measures

1. Introduction

Relatively recent advances in neuroimaging technology, specifically functional magnetic resonance imaging (fMRI), have made possible the large-scale study of neural systems underlying learning in the human brain. Learning, or the experience-based process by which we form representations of the world around us, is a topic particularly suited to investigation using neuroimaging. Crucially, fMRI has the potential to reveal fluctuations in neural activity as learning unfolds over time, thereby allowing researchers to tap into the *process* through which internal representations undergo change.

Despite the natural fit between fMRI and the study of learning, a critical review of the relevant literature reveals that studies tend to address the question of which brain areas subserve retrieval or recognition of *already learned* items, not the process which generates

Corresponding Author(s): Elisabeth Karuza, Lauren Emberson, Department of Brain & Cognitive Sciences, Meliora Hall, RC 270268, University of Rochester, Rochester, NY 14627, USA, ekaruza@bcs.rochester.edu, lemberson@bcs.rochester.edu.
[*]These authors contributed equally to this manuscript

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

changes in representation in the first place. While an outcome-focused approach is certainly valuable and germane to the study of learning, we suggest that the time is ripe for the field to focus instead on the *process* of acquisition rather than its outcomes. In this review, we will discuss some corresponding challenges, both methodological and theoretical in nature, and offer suggestions for improved experimental design and analysis. Specifically, we will consider the benefits of incorporating on-line behavioral testing and the use of computational models to predict the time-course of learning. Along these lines, we will discuss those studies that have begun to surmount these challenges to begin to uncover the neural systems engaged over the time-course of learning. This review also explores the critical yet open question of how to interpret the neural changes that occur before behavioral evidence of learning emerges. We begin by considering the unique niche of fMRI in the study of learning and the ways in which this methodology has already shaped and been shaped by the field of cognitive neuroscience.

1.1 What fMRI has Already Offered the Study of Learning

For over a century, there has been a voluminous and fruitful tradition of research aimed at the general study of learning in both animals and humans (Rescorla, 1988; Shanks & St. John, 1994; Skinner, 1938; Thorndike, 1931; Tolman, 1951). Until very recently, most of our understanding of this process has relied on studies of either human behavior (e.g., through learning tasks and behavioral manipulations) or non-human neural systems (e.g., through electrophysiological recordings). While research on the neural mechanisms of human learning has benefitted from examining the effects of brain lesions, this case study approach has limited power in revealing the neural systems supporting cognitive mechanisms (Zurif, Swinney & Fodor, 1991; but see also Caramazza & Badecker, 1991).

While our review certainly intersects with the study of learning in general, we will focus on the interrelated and overlapping research areas of (1) incidental learning, or acquisition in the absence of specific intention to learn; (2) statistical learning, or acquisition of structural representations via distributional regularities in sensory input; and (3) sequence learning, or acquisition of sequential information across perceptual modalities using both motor measures (motor sequence learning) and measures not specifically focused on motor responses[1].

In the past decade or so, the emerging use of neuroimaging methods, particularly fMRI, has provided the unique opportunity to study the relationship between neural systems and behavior in human participants on a large scale. While other neuroimaging modalities such as electroencephalography (EEG) and positron emission tomography (PET) have been employed to study learning, fMRI offers a combination of unambiguous spatial location of signals (as opposed to EEG), while being less invasive and safer than PET with high functional and anatomical image resolution (not available using either EEG or PET). The significant technological advances of this method have resulted in a veritable explosion in fMRI studies of human learning.

In addition to opening up the door to the concurrent study of behavior and neural activity in humans, fMRI allows for the (virtually) simultaneous recording of activity across the entire brain and, for some tasks, across the entire time-course of learning. Indeed, the use of fMRI has already enabled investigations into learning that had been tortuous or impossible when

[1] The majority of studies falling into one or more of these three categories purportedly involve implicit learning or learning without conscious awareness. While the distinction between explicit and implicit forms of learning (both neurally and behaviorally) remains a major area of active debate, we elect not to make any strong claim as to the extent to which the learning studies covered here are wholly implicit or wholly explicit. In order to focus on the broader points laid out above, we will neither weigh in on this debate nor discuss any differences between the studies reviewed along this dimension.

using historically employed methods. For example, the theory that the basal ganglia and hippocampus comprise multiple, dissociable learning and memory systems has been investigated and supported using lesion studies in both human and non-human animals and electrophysiology (see Eichenbaum & Cohen, 2001 for an excellent historical review). However, the ability to record activity across the entire human brain with fMRI has enabled the *in vivo* investigation of the activity of both of these systems, allowing researchers to ask questions such as, are the basal ganglia and hippocampus simultaneously active during a single learning task (Poldrack, Prabhakaran, Seger, & Gabrieli, 1999)? If not, then do they directly inhibit each other (Poldrack et al., 2001)? Are there some tasks where these systems complement each other (Shohamy & Wagner, 2008)? While the answers to such questions remain elusive, this example serves to illustrate the way in which the goal-directed use of fMRI has fueled productive discussion and advanced our understanding of learning. Thus, fMRI has already provided new avenues to consider the interrelationship between functional neural activity and human learning.

While fMRI continues to be a popular and powerful method for answering a variety of empirical questions, no single method can fully delineate a system as complex as the human brain. FMRI is no exception, in part because it is an indirect measure of the neural activity in the brain that results from changes in blood oxygenation (the blood-oxygen-level-dependent or BOLD response). It has been well established that the BOLD response can be stimulus-evoked (e.g., Belliveau et al., 1991; Ogawa et al., 1992) and, by extension, sensitive to functional neural activity. However, the specific aspects of the neural signal producing the BOLD response are still not entirely clear. Logothetis and Wandell (2004) propose that the BOLD response best corresponds to local field potentials (LFPs) rather than spiking activity directly. Of course, these two aspects of the neural signal are interrelated, but LFPs and spiking pick up on separable aspects of the neural signal; LFPs reflect sub-threshold integrative processes or computations on the input of neural signals, while spiking reflects the output of this computation. If the BOLD response does reflect LFPs more directly than spiking, fMRI can then be considered complementary to the spiking activity typically gathered using electrophysiological methods. It therefore follows that fMRI and electrophysiology can be seen as distinct but highly compatible methods capable of probing neural computations.

1.2 The Current Use of FMRI to Study Learning

Given the current impact and future potential of fMRI as a method of investigating human learning, it is perhaps surprising that a significant number of fMRI studies dealing with this topic have elected to focus on the outcome of learning (Forkstam, Hagoort, Fernandez, Ingvar, & Petersson, 2006; Lieberman, Chang, Chiao, Bookheimer, & Knowlton, 2004; Petersson, Folia, & Hagoort, 2012; Petersson, Forkstam, & Ingvar, 2004; Seger, Prabhakaran, Poldrack, & Gabrieli, 2000; Skosnik et al., 2002; Yang & Li, 2012). That is, fMRI recordings are typically collected during tests of already acquired information, not during the initial processing of structured stimuli (henceforth referred to as the *exposure/acquisition* phase). In such cases, the extent of learning is often measured using post-acquisition tasks that typically involve novelty detection and/or accuracy judgments. Implicit in this approach is that learning is a relatively uniform, time-invariant process that can be adequately examined using post-acquisition outcome measures. Importantly, it has been largely unstudied whether the neural systems involved in recognition or retrieval overlap with the systems involved in initial acquisition. Thus, it is best to be cautious in interpreting activation during posttest as reflecting learning related activity. There is, however, emerging but indirect evidence that outcome measures and novelty detection may tap into systems that are not involved in acquisition. Thus, the widespread use of these methods to study the neural bases of *learning* may actually tap into different, if related, cognitive processes. In a critical review below, we will compare and contrast studies examining neural activity *after* acquisition and those examining neural activity *during* acquisition.

Given the challenges arising from scanning during acquisition, we consider a number of different ways in which researchers can leverage well-designed behavioral measures to tap into changes in neural signals that correlate with the time-course of learning. Finally, we will consider how to investigate neural changes that necessarily must occur during acquisition but before any behavioral evidence of learning can be obtained.

2. Processes and Stages of Learning

While learning is often referred to as a single cognitive process, it undoubtedly involves several neural mechanisms (e.g., those associated with neural plasticity such as long term potentiation vs. depression) encompassing several neural regions (e.g., the hippocampus, basal ganglia, frontal cortex). Moreover, there might be multiple, separable neural or psychological processes that are engaged as learning, defined in the broadest sense, takes place. It remains an open question which one or combination of these processes constitutes “learning.” Here we present a simple, generic architecture of the basic cognitive processes that are likely involved in learning (Figure 2). Our goal is not to provide a detailed or unifying theory of learning, but rather to illustrate the immense complexity and interconnectedness of the cognitive components involved in learning tasks, and the challenges involved in isolating, and subsequently mapping each of these component processes to specific brain areas, particularly when time-course data are not available.

As alluded to above, there are likely a number of dissociable processes that are necessary for learning to take place. For example, many tasks might first involve detection of structure or **pattern extraction** from the sensory input (e.g., by drawing attention to relevant aspects of the input or calculating statistical or associative information). Then, subsequent processes might come on-line, allowing participants to capitalize on these initial processes to change behavior. These latter processes have been conceptualized in a number of ways, but in general could be framed as **model building**. For example, after extracting a component or feature of the pattern from sensory input, a participant might rely on *prediction* and *prediction error* to support additional pattern extraction and behavioral change (Pavlov, 1927; Rescorla & Wagner, 1972; Schultz, Dayan, & Montague, 1997; Thorndike, 1911; Waelti, Dickinson, & Schultz, 2001). Alternatively, in a Bayesian framework, a participant might search for *latent causes* to explain the pattern of data and engage in *belief updating* when the input fails to match their current model (Gerschman & Niv, 2010). In both of these cases, a participant must build some knowledge about the structure of the environment (e.g., a prediction or representation of an association, a latent cause or structure in the environment) using processes that then likely feedback to the pattern extraction mechanism (e.g., via prediction error or belief updating). Finally, a retrieval or recognition process is likely necessary for participants to demonstrate knowledge acquisition and produce an outcome measure via some decision process.

While it is beyond the scope of this review to provide a unifying account of the (potentially numerous) processes necessary for learning, Figure 2 presents a generic architecture containing four essential components: 1) sensory or **input encoding** which involves transmission of sensory input to the cortex; 2) **pattern extraction**, 3) **model building** and 4) **retrieval/recognition.** These processes likely all have different time-courses of operation during a given learning task. For example, processes involved in pattern extraction are likely engaged quite early in learning but then taper off, while model building might be engaged for the majority of the task. Model building processes might interact in large part with pattern extraction processes early in learning and then more directly with retrieval or recognition processes later in learning. The time-course of involvement of each of these systems is an open empirical question, but it is likely that these processes are engaged at different times during learning.

While all of these processes are likely necessary for a participant to ultimately display behavioral evidence of learning in a posttest, it is unclear whether a single process in this architecture independently reflects "learning." To illustrate with an extreme example, while vision is necessary for visual statistical learning, not all aspects of the visual pathway are considered part of the learning process. That is, while the ability to perceive visual stimuli is necessary, it is not sufficient for learning in this task. However, in the generic architecture of Figure 2, pattern extraction and model building are likely the processes that are most directly associated with learning, compared to input encoding on one end of the architecture and retrieval on the other. We return to these points in Section 5 of the paper.

In sum, we will refer to all four of these processes in general terms in order to simplify the current discussion, but they are not meant to indicate a particular theoretical commitment. Moreover, it is important to note that the studies discussed throughout this review likely tap into more than one of the processes outlined in this learning framework.

3. Current Methods Aimed at Capturing Learning Using fMRI

In this section, we consider some standard methods of studying learning using fMRI and examine their effectiveness at capturing the process of learning. In general, fMRI studies of learning take one of three forms: (1) functional imaging data are collected during the entirety of the exposure phase; (2) the entirety of the exposure phase takes place outside the scanner and imaging data are collected only at test; (3) some pre-exposure or training takes place outside the scanner and imaging data are collected for only a portion of the exposure phase. Each of these design types has the potential to inform a step in the progression from acquisition to application/retrieval of knowledge for recognition (see Figure 2), but we argue here that the latter two designs may neither directly nor fully reveal the neural substrates underlying learning. For example, it is possible that scanning after the exposure phase or when the structure of the task has already been learned may not tap into areas supporting learning (e.g., pattern extraction and model building). Rather, these post-acquisition assessments may reflect the outcome of a recognition or retrieval process. As a result, studies employing these designs may have unintended and often unrecognized limitations on the conclusions that can be drawn from them with regard to understanding the processes underlying learning, at least until the assumptions behind these conclusions have been investigated directly.

We begin by exploring a body of literature that examines learning by focusing on outcome measures (i.e., the relationship between behavioral performance and activation during a posttest). We examine neural evidence supporting distinct learning and knowledge application systems and probe the potential overlap between expectancy violation on outcome measures and the formation of predictions during learning (see Textbox 2). We then consider studies that acquire imaging data during a *portion* of the exposure phase but also involve some learning outside of the scanner (study type 3 above). The implications of employing a design that confines fMRI data collection to only a part of the acquisition phase are discussed. In a subsequent section, we consider designs that focus exclusively on the complete exposure phase, the period of time in which the neural changes associated with learning are most likely to be observed.

Textbox 2

The Relationship between Expectancy Violation and Learning

One possible implementation of the architecture illustrated in Fig. 2 is that separable neural regions are active during exposure and outcome (i.e., post-test) phases. Activation in these distinct regions could reflect two types of learning systems, with the system engaged during test being responsive to negative evidence present in incorrect or inconsistent stimuli (e.g., the partwords or nonwords from McNealy et al., 2006). While McNealy et al. (2006) specifically observed greater activation for words relative to partwords and nonwords in frontal cortex (i.e., greater activation for test items *consistent* with the syllable statistics from the exposure phase), this result stands in contrast to a robust finding that emerges from the AGL literature – namely, the involvement of specific frontal areas in response to the *violation* of previously learned grammatical rules at test (e.g., Forkstam et al., 2006; Petersson et al., 2004; Petersson et al., 2012). Along these lines, Petersson et al. (2004) proposed that prefrontal activation during the classification of ungrammatical strings could be framed in terms of a model that learns through negative evidence (i.e., the difference between input and prediction; see Elman, 1990 and Haykin, 1998). The violation-based approach to pattern/rule acquisition has also been applied to the processing of probabilistic pure-tone sequences (Furl et al., 2011). MEG results showed that both learning and post-test classification were supported by error-driven activation in the temporoparietal junction (TPJ), though at different stages in time (200ms and 150 ms, respectively). That is, violations of predictions led to sudden increases in neural activity in this region. These findings suggest a possible convergence between areas involved in learning (the acquisition phase) and areas activated by expectancy violations (the testing phase), albeit with different neural systems likely supporting (1) initial learning when negative evidence is absent and (2) learning when negative evidence (via feedback) is present.

However, this possibility must contend with evidence showing similar patterns of activation in change-detection tasks that are not learning related. Zevin, Yang, Skipper & McCandliss (2010) argue that the TPJ, the region implicated in learning from the violation of expectations in Furl et al., (2011), is actually a domain-general change-detection region. Indeed, Zevin et al., (2010) obtained greater activity in the TPJ when simply contrasting repeating vs. alternating syllable conditions in participants engaged in a passive listening task. TPJ activation was also modulated by speaker identity shifts, providing further evidence that this region supports domain-general change detection outside the realm of learning. Prior work has also shown temporal cortex involvement in similar tasks involving the detection of deviant tones (i.e., those mismatched in frequency) in a lengthy stimulus train (Opitz, Rinne, Mecklinger, von Cramon, & Schröger, 2002). The key point here is that the learning of any structured sequence of elements (e.g., abbccc) is typically assessed by presenting test sequences that violate some aspect of that structure. If the violation involves a novel element (e.g., *abbccd*) then neural activations may reflect *both* a response to the low-level change (i.e., the novelty of element-d) and a response to the higher-level change (i.e., violation of the 1-2-3 pattern). Importantly, the former change does not necessarily involve any learning (only novelty detection), while the latter clearly involves the formation (and then violation) of a pattern. While these studies cannot rule out the possibility that change detection supports, rather than results from, learning, the relationship between these related processes remains an open question – one with implications for the study of learning using functional neuroimaging.

3.1 Outcome vs. Process: Do the Neural Systems Engaged at Test Support Learning?

As introduced in Section 1.2, many fMRI studies of pattern learning within the AGL literature examine the neural response at test rather than during exposure/acquisition (Forkstam et al., 2006; Lieberman et al., 2004; Petersson et al., 2012; Petersson et al., 2004; Seger et al., 2000; Skosnik et al., 2002; Yang & Li, 2012). Classic AGL study-test designs begin with exposure to a structured set of artificial grammar strings. During the presentation of the strings, participants may be asked to perform a cover task such as reproducing it from memory after a brief interval of time (Lieberman et al., 2004). Following the exposure phase, subjects typically participate in a testing phase while undergoing an fMRI scan. During this post-acquisition test phase, subjects often perform a behavioral task (e.g., recognition, familiarity judgment, grammaticality rating). Standardly, fMRI contrast analyses are then performed: activations for grammatical strings relative to ungrammatical strings are compared in addition to activations for the familiarity judgment task relative to a recognition control task (e.g., Seger et al., 2000).

This task design and analysis enables the investigation of neural substrates underlying recognition or retrieval. However, it is yet unclear whether or not these substrates additionally support the process of learning. To be clear, we are not claiming that the investigation of learning outcomes is without independent merit. Indeed, it is crucial that we examine how the human brain taps into acquired representations during test. We simply argue here that an important avenue of research involves validating the assumption that the process of learning and the result of learning are essentially interchangeable. This step is necessary before the regions involved in discrimination at test can be considered to also reflect the same processes involved in the initial learning.

In fact, some evidence suggests that brain areas engaged during pattern learning are distinct from those areas engaged at test. In reviewing the diverse set of paradigms found in studies of learning, it is apparent that the results of studies focusing on the process of acquisition diverge from the results of studies focusing on the outcomes of acquisition. Dolan & Fletcher (1999), for example, found that encoding during exposure to an artificial grammar relied on anterior hippocampus, but retrieval of learned sequences relied on posterior hippocampal areas[2]. Using a word segmentation task in the statistical learning framework, McNealy, Mazziota and Dapretto (2006, see also 2010) provide an additional opportunity to compare neural activation underlying both phases, as the authors scanned during the exposure phase as well as at posttest. After exposure to a continuous stream of both statistically regular and random syllables, a subset of participants was presented with shorter sequences that varied in their syllable-to-syllable predictability. Increases in activation were observed throughout the exposure phase in bilateral temporal cortices and left-lateralized parietal areas, yet contrasting activity at test for structured (words) and unstructured sequences (partwords and nonwords) revealed prefrontal activation. Thus, results from McNealy et al. (2006) suggest that neural activity involved in the acquisition and outcome phases appear to be at least partially separable. While, it is difficult to determine the full extent of overlap given the paucity of studies that directly compare the learning process and its outcomes, there is little evidence that the neural regions involved in recognition or outcome of learning are the same regions involved in the initial acquisition.

[2] See Ross, Brown, and Stern (2009) for evidence that both learning and retrieval of sequences of faces rely on the medial temporal lobe, but that the hippocampus shows greater activation during retrieval. While this suggests that knowledge acquisition and knowledge application do engage some overlapping areas, data from these and other learning studies indicate that neural involvement for each task is not identical.

3.2 Considerations for Early vs. Late Stages in the Process of Learning

Unlike AGL tasks, which often conduct the full exposure phase outside the scanner, many motor and perceptual sequence learning studies present a portion of the exposure phase prior to or between fMRI scans (Daniel & Pollmann, 2012; Gheysen, Van Opstal, Roggeman, Van Waelvelde, & Fias, 2011; van der Graaf, Maguire, Leenders, & de Jong, 2006; Willingham, Salidis, & Gabrieli, 2002). Though this experimental design clearly involves scanning during some part of acquisition, it may only reveal patterns of activation involved in later learning stages, especially in the case of a pre-exposure phase. Notably, there is evidence that initial acquisition of a sequence (i.e., in the first stages of exposure) engages different neural regions than the processing of the same sequence after a period of short-term exposure (i.e., later in the exposure phase) and long-term consolidation (i.e., after sleep). Shadmehr and Holcomb (1997), for example, found that though behavioral performance on a motor sequence task was maintained 6 hours after initial practice, early exposure relied more on prefrontal cortex while later exposure involved more posterior regions such as parietal and cerebellar cortex. Across the literature, activation patterns have been shown to differ in terms of spatial extent (such as increases in motor cortex activation as shown by Karni et al., 1995) and/or anatomical location. Doyon et al. (2009) contrasted performance on a commonly practiced stitch sequence vs. a novel stitch sequence in a group of expert knitters (experience range 14–58 years). Expert knitters showed predicted patterns of activation in the striatum and motor cortex when performing the standard stitch but the cerebellum was engaged when completing the novel stitch pattern. Furthermore, Seger and Cincotta (2006) examined trial-by-trial pattern learning in the visual modality, comparing stages they termed “rule learning” (i.e., early learning that involved many mistakes) and “rule application” (i.e., when performance indicated a participant had ceased to make errors). They demonstrated that the early process of learning relied on a widespread network including the cerebellum, occipito parietal areas, prefrontal areas, and the striatum. In contrast, the ability to apply learned knowledge later in the exposure phase[3] was associated with activation in right hippocampus and bilateral insula. Findings such as these suggest that different stages of learning engage different brain regions and provide evidence against the implicit assumption that neural regions revealed in latter stages of learning or during post-test tasks may be indicative of the neural regions involved in acquisition of these representations or the process of learning.

4. The Importance of Time-course Information in Learning

In this section we focus on studies that have collected functional imaging data during exposure, but lacked (or did not capitalize on) behavioral indications of the time-course of learning. We also offer some possible solutions to methodological challenges (i.e., between-subject variability) associated with examining the time-course of learning. In Textbox 3, we briefly place this discussion in historical context.

Textbox 3

Brief Historical Perspective on the Study of Individual and Developmental Variability in Implicit Learning

There has been little study of individual variability and developmental change in implicit learning tasks. In fact, it has been argued that this type of learning is exceptionally devoid of such variability. In an effort to examine individual differences in implicit learning, Reber, Walkenfeld, and Hernstadt (1991) showed that the distribution of performance

[3] They did not include a clearly indicated posttest phase but rather elected to divide one acquisition phase into “learning” and “application” stages based on trial-by-trial accuracy data.

scores on an explicit series-judgment task had significantly greater variance than scores on an implicit AGL task. Moreover, they found no significant correlation between intelligence quotient and artificial grammar learning ability. These results were interpreted as lending support to the hypothesis that,

> …implicit processes, owing to their phylogenetic antiquity, will show less individual-to-individual variation than comparable explicit processes and, given the nature of standard psychometric techniques for measuring intelligence, will show lower correlations with IQ (p. 894).

This historical claim that implicit forms of learning rely on a system that is evolutionarily older and therefore highly consistent across the population may have had the unintended consequence of shaping the study of learning using modern imaging methods. Though many of the issues we bring up here are not specific to implicit AGL studies, the discounting of individual learning differences appears to recur across a variety of study types. While performance on largely implicit learning tasks might be less variable across subjects, participants may still show meaningful differences in learning rates. Indeed, there has been some suggestion that individual differences might actually be heightened in neuroimaging studies of learning due to the stress-inducing nature of the scanner environment (see Muehlhan, Lueken, Wittchen, & Kirschbaum, 2011). Embracing these differences may be beneficial to revealing the complex neural systems that support the learning process, enabling researchers to separate early learning, later learning, and the retrieval of pattern or rule knowledge. Rather than impeding the fMRI study of learning, we propose that individual differences in the accuracy and time-course of learning may actually provide uniquely valuable insight into its process.

Related to the historical claim that implicit learning arose from an evolutionarily primitive neural substrate and has relatively little variability across subjects is the claim that implicit learning is developmentally invariant (i.e., available from birth without a marked developmental improvement, Reber, 1993). While such a view has resulted in an academic disinterest in studying how differences and changes in implicit learning could relate to the development of learning systems, this claim has limited empirical support. There have been only a small number of studies that have specifically compared implicit learning abilities over developmental time and even fewer that have compared the neural systems engaged during implicit learning through development. The few studies that have been conducted to this end argue against a developmentally invariant account of implicit learning. Notably, Thomas et al., (2004) had children (7 to 11 year olds) and adults perform a SRT task in the scanner. Age-related differences were found in numerous cortical regions including the putamen and hippocampus. Taken together with studies comparing behavioral outcomes of implicit learning across development (see review in Forkstam and Petersson, 2005), there is an emerging view that implicit learning has a substantial developmental trajectory, arguing against a commonly assumed developmental invariance model.

In sum, there are historical roots to the view that implicit learning is an evolutionarily primitive learning system that is relatively homogenous across individuals and invariant over development. This view, we argue, has dissuaded researchers from using individual differences and changes in implicit learning across development to undercover the neural systems involved in learning. However, these claims have relatively little empirical support and indeed, there are differences in implicit learning outcomes and their neural substrates within and between individuals.

4.1 Collecting Temporally Sensitive Measures of Learning throughout Exposure

As made clear in the previous description of McNealy et al. (2006), researchers have recently begun to examine the brain areas specifically subserving statistical learning during passive viewing/listening tasks (see also Cunillera et al., 2009; Turk-Browne, Scholl, Chun, & Johnson, 2009). However, the results of such studies can be challenging to interpret without (1) strong behavioral evidence of learning at test (i.e., participants showing that they could consistently discriminate statistically structured from unstructured sequences) and (2) behavioral measures across the time-course of learning (but see Turk-Browne, Scholl, Johnson, & Chun, 2010). In this section, we review evidence that points to the importance of the inclusion of temporally-sensitive learning measures throughout the acquisition phase.

A key feature of our perspective is that correlating outcome scores with activation patterns during learning can only provide, at best, an approximate answer to the question of which brain areas underlie the learning process. To illustrate this point conceptually, we present trajectories from data gathered in a replication of the learning paradigm used by Shohamy and Wagner (2008). In this task, participants are presented with two scenes (out of 32 possible scenes) and a face (out of 32 possible faces) and asked to determine which of the two scenes goes with the face. Initially, there is no basis to answer this question, and participants must render a guess. But eventually, through exposure to these 3-element trials, participants learn which particular scene is associated with a given face. Their selection accuracy is collected on each trial, thereby providing a running estimate of each participant’s learning trajectory. In a sample of 17 participants, we found clear examples where the learning accuracy in the final block does not represent that participant’s learning trajectory. Fig. 3 presents three of the 17 learning trajectories. Two participants (both in blue) have similar outcomes (note that the error bars in the final block overlap) but divergent learning trajectories. Specifically, the participant in dark blue shows a greater increase in performance in the second block, and their performance appears to plateau. This participant might be in the latter stages of learning after having reached their maximum performance. The participant represented in light blue has a lower level of performance in the first three blocks and exhibits an increase in performance in the last block. This person may be in the earlier stages of learning but with an overall slower trajectory. However, these differences in rate of learning are not well represented by their performance in the final block. Comparing the participants presented in dark blue and pink, we see that they have nearly identical learning trajectories in the first three blocks but diverge in the final block (note that the error bars in the final block are not overlapping). Thus, it is clear that individual learning trajectories and outcomes of learning vary such that the final learning outcome does not well represent these different learning trajectories. If we used outcome measures as a proxy for learning trajectories, these participants would be treated quite differently (e.g., a strong vs. intermediate learner) despite the striking similarity of their trajectories for the majority of the task. These are examples where a simple outcome measure lacking any temporal information would not tease apart participants’ actual learning trajectories. The correlation of outcome scores with activation patterns during learning is a common method employed in the use of fMRI to study learning. However, in most cases, a single outcome measure does not neatly represent the learning trajectories of individual participants and thus, at best, using outcome measures can only act as an approximation for a particular point in the process of learning.

Another common analysis method involves contrasting different epochs within the exposure phase (e.g., Schendan, Searl, Melrose, & Stern, 2003). Contrast analyses are performed for each run during the acquisition phase relative to the other runs, presumably with the goal of discriminating between early, mid and late stages in learning. However, contrasting arbitrary chunks of exposure might mask important differences in individual learning rates. A comparison of the first and last runs is a weak test if one subject learns best during the first run and another learns best during the last. Crucially, time-point contrast analyses rest on the assumption that participants share common learning trajectories. Using ERPs, Alba, Katahira and Okanoya (2008) present evidence that this assumption can be unwarranted, finding that variations in learning outcomes in a statistical learning task were correlated with temporal variations in changes in the neural signal. They examined the neural changes associated with statistical learning across three sessions of training. Overall test performance was significantly above chance after training, but participants exhibited high, medium, and low levels of learning in the post-test. The three groups demonstrated qualitatively different changes in neural activity, including an increase in the N400 component in session 1 for high learners and session 2 for medium learners. Moreover, low learners, despite achieving above chance behavioral performance, did not show evidence of the same, qualitative neural changes in any of the three sessions. This study provides evidence that differences in learning outcomes map onto different time-courses of neural change and highlights the significant degree of inter- and intra-subject variation present across the acquisition phase, which can be related to variability in patterns of neural activity.

Generally, we see that taking a single outcome measure of learning as representative of the entire time-course of learning or assuming that all participants learn at the same rate (e.g., by using an average learning trajectory) is at best introducing substantial noise into the resulting analyses, and at worst producing results that do not reflect learning but other cognitive processes. There are a number of methods that can be employed to incorporate time-course information into fMRI analyses to circumvent this problem, thereby more completely capturing the neural systems underlying the process of learning. One method would be to employ the average learning time-course, calculated across all participants, and to use this change in behavior to uncover correlated neural changes representative of the entire group. However, while employing an average time-course of learning could prove beneficial, we argue that it may fail to resolve the discrepancy between an individual’s performance on an outcome measure and their learning time-course. Thus, the remainder of this section will discuss a complementary approach: applying *individual* behavioral measures of learning (i.e., as regressors at the single subject level) to uncover areas of the brain that are related to learning across *all participants* (i.e., results at the group level). The logic of this individual differences approach lies in the fact that it provides a more fine-grained indicator of changes in general learning systems active across a population. Moreover, it offers greater fidelity than 1) an average time-course and 2) a simple group contrast of high-performing vs. low-performing learners based on a simple outcome measure. It is important to note that this approach rests on the assumption that differences in learning across participants are not simply a result of task-irrelevant processes or noise but reflect important differences in the engagement of relevant learning systems.

Intermittent testing provides one possible method for exploiting individual variation across subjects to uncover learning-related neural changes across participants. Ideally, one would want to gather trial-by-trial accuracy or reaction time data, but this approach cannot be applied to certain experimental designs (e.g. those involving the presentation of continuous auditory stimuli as in a word segmentation task). Intermittent testing provides useful “snapshots” of the extent of learning (i.e., the current output of learning) throughout the entire exposure phase. For example, Karuza et al. (2013) made use of intermittent testing when collecting functional imaging data across the course of four separate exposure phases. The exposure phases consisted of streams of continuous syllables in which the primary statistical cues to word boundaries were the non-adjacent dependencies between consonants (Newport & Aslin, 2004). The extent to which participants had successfully segmented the speech stream was assessed after each of the four, 2-min exposure phases. Learning achieved on each test was quantified as the difference between word and partword ratings and then the time-course in learning was determined by comparing the word-partword difference score form one test to the next (a “delta score”). Such an approach was intended to capture the individualized learning trajectories of subjects over the course of the four exposure phases. Figure 4 presents the change in word-partword ratings over time, thus illustrating inter-subject variability in learning trajectories. While some participants, (e.g., the trend line highlighted in pink) showed large fluctuations in performance across the 4 tests, others (e.g., the trend line in purple) showed consistently high performance. The latter pattern is indicative of an early spike in learning during the first exposure phase followed by successful maintenance of acquired knowledge in the later stages of the experiment. In contrast, roughly linear *increases* in learning were observed in some subjects (e.g., teal, who peaked in performance on test 3 and subsequently showed fatigue effects by test 4), and others showed little to no learning at all (e.g., orange). These vast differences in learning rates might mask the statistical reliability of activations in a simple contrast analysis. In Karuza et al. (2013), individual delta scores were entered as explanatory variables into a GLM in order to reveal areas of the brain that covaried with learning over time for each subject. A subsequent group-level analysis revealed which voxels exhibited this variation across participants. Indeed, exploiting these differences in learning rates revealed learning-related neural activity that was not present in simple contrast analyses.

Although this method represents an improvement over other methods, reviewed above, that do not take both temporal and individual variability in performance into account, it is subject to certain limitations. Though participants in Karuza et al. (2013) were informed prior to the first exposure phase that they would be tested repeatedly, it is possible that learning after the first testing phase was altered in some way (i.e., after the first test, perhaps participants began to cue into triplet structure in subsequent exposure phases). This possibility would be consistent with findings from studies of declarative memory showing improved performance with repeated testing (e.g., Vaughn & Rawson, 2011). However, Orban, Aslin, Fiser, and Lengyel (in prep.) and Reeder, Aslin, Newport, and Bavelier (in prep.) have gathered substantial evidence, across several different implicit statistical learning tasks, that intermittent tests have little or no effect on subsequent test performance after further learning exposure. Thus, intermittent testing remains one of the few methods able to investigate the time-course of learning during exposure to continuous stimuli.

While intermittent testing can be used to get a better view of the learning time-course, more sensitive measures are possible if the task involves discrete trials in an event-related design. For example, Turk-Browne et al. (2010) were able to make use of reaction time during a cover task in order to probe visual SL of face/scene pairings. Participants were asked to classify a series of images as either a face or a scene. Unbeknownst to them, the ordering of the images was governed by an underlying probabilistic structure. Interestingly, the authors found evidence of learning even on the unrelated classification task; participants showed a reaction time increase for the first item of an ordered pair and a reaction time decrease on the second item of a pair. Evidence that unrelated cover tasks can reveal implicit pattern learning has exciting implications for fMRI studies, as these measures are largely incidental and collected at consistent intervals throughout the exposure phase. Such reaction time measures, like those obtained during the MSL task described in Section 3.2, provide an excellent means of capturing individual variation over time, as they can be entered as predictors of neural activation at the single subject-level. Outside of the learning literature, reaction time has been successfully incorporated as a trial-by-trial regressor in general linear models for individual subjects within an fMRI data set. The correlation of BOLD response with changes in reaction time across the course of an experiment has revealed brain activation in widespread regions for a variety of experiment types (Yarkoni, Barch, Gray, Conturo & Braver, 2009). If time-sensitive measures can be used to reveal task-general neural systems engaged during working-memory, decision-making, and emotion-rating experiments, then it follows that this method can also be applied to learning studies.

4.2 Using Computational Models to Constrain Predictions about the Time-course of Learning

In addition to repeated behavioral testing (either on each trial or in intermittent blocks), computational modeling provides another emerging method to tap into the fMRI activations correlated with the time-course of learning. This approach does not employ behavioral testing directly, but indirectly through using behaviorally-validated computational models. These models can be used to make predictions about the underlying learning processes evident in the average time-course of learning based on the input that the learner receives in the MRI scanner. Changes in these parameters of the model can be used to generate predictors that in turn can be used as regressors for the neural signal.

There has been a rich history of computational models in the field of learning and memory, beginning with the success of the Rescorla-Wagner (R-W) model in the 1970s (Rescorla & Wagner, 1972). Indeed, the direct descendants of the R-W model continue to shape the field of reinforcement learning (e.g., Temporal Difference or TD learning, Sutton, 1988; for a critical review of contemporary models, their neural correlates in relation to temporal sequence learning see Wörgötter and Porr, 2005). It is commonplace to employ these computational models to generate testable behavioral predictions (see Miller, Barnet & Grahame, 1995 for a review of the successes of the R-W model in this regard; for a more recent prediction from a neurobiological model of the MTL and semantic learning by O'Reilly and colleagues, see Norman & O'Reilly, 2003; Bayley, O'Reilly, Curran & Squire, 2008).

The current section proposes a different use of computational models, specifically as a method for uncovering the neural systems involved in learning. In addition to providing behavioral predictions and elegantly capturing many behavioral phenomena, many models can provide predictive information about the time-course of learning. For example, the R-W model was formulated to provide a "trial-by-trial description of how the associative status of a conditioned stimulus (CS) changes when a stimulus is trained" (Miller, Barnet & Grahame, 1995, pp. 363). Indeed, this incremental or trial-by-trial aspect is shared with many other computational models already populating the literature (e.g., TD learning, and to an increasing extent in Bayesian models, see Kruschke, 2006; however, also see Sakamoto, Jones & Love, 2008.). Thus, the parameters of these various models can be calculated for each trial given a certain set of inputs (for example, associative strength from the R-W model, model-estimates of probabilities for a given stimulus from Bayesian models). Given that all of these models have been subject to extensive behavioral-validation, they can provide an excellent estimate of the average time-course of learning given specific input, and this estimate of learning can then be used as a regressor to examine which regions of the brain correlate with or predict this estimated learning time-course.

Indeed, a number of papers have already successfully used parameter estimates from computational models to uncover neural regions exhibiting online changes related to learning. While certainly the combination of computational models and fMRI recordings can be used to validate models and potentially to uncover neural correlates of model predictions, these two measures can also be combined to probe broader questions. An excellent example of this approach is den Ouden, Friston, Daw, McIntosh and Stephan (2009). While den Ouden et al. (2009) employed the R-W model estimate of associative strength, the authors are very clear that the goal of the study was to examine the consequences of incidental learning on cortical connectivity between regions of occipital and temporal lobes and not to add validation to the R-W model. den Ouden et al. (2009) focus on the R-W model as "a generic and well-established model of associative learning that has been successful in modeling a wide range of learning processes " (pp. 1181). To this end, the authors employed the R-W model's estimate of associative strength for each trial as part of their hierarchical model of the fMRI data. See the top panel of Figure 5 for the parameter estimates of associative strength. Note: the parameter estimates vary with stimulus condition, thereby mapping out separate learning time-courses but also trial-by-trial variability depending upon recent trials. The bottom panel of Figure 5 illustrates a subset of the regions of interest (ROIs) that showed modulation of activity that is significantly predicted by experimental manipulations (e.g., the presence or absence of the conditioned stimulus) as well as the model parameter estimates (associative strength) for that trial. These ROIs were interpreted to be involved in incidental associative learning of the modeled stimuli and were used in additional analyses to examine the effects of learning on connectivity across cortical regions (see also Behrens, Hunt, Woolrich & Rushworth, 2008).

This approach of using model predictions as fMRI regressors can certainly employ estimates from other models as well. Temporal difference models have been employed extensively to uncover regions involved in prediction error and model-updating by O'Doherty, Daw, Dayan and colleagues (see Dayan and Daw, 2008 for a review). Similarly, recent examples employing Bayesian model can also be found in Behrens, Woolrich, Walton, & Rushworth (2007) and den Ouden, Daunizeau, Roiser, Friston, & Stephan (2010).

Of course, when selecting learning models and interpreting the results of a combined computational and fMRI analysis, it is important to consider whether different models arrive at the same results behaviorally or neurally. It is possible that not all learning models will predict learning outcomes equally well or provide valid estimates of learning trajectories for a given task. Even models that have comparable validity might have parameters that are correlated with different neural regions (e.g. a R-W model might correlate well with regions of the striatum in a given task, while a Bayesian model might correlate with a region in the frontal cortex). It is important to consider these differences when using parameters from computational models to uncover the neural systems correlated with the average time-course of learning.

Despite the foregoing concerns, there are certain clear advantages to employing computational models to uncover neural systems associated with the time-course of learning. First, these computational models provide a trial-by-trial estimate of learning, which can be used as part of the analysis of the fMRI data. In the case of den Ouden et al. (2009), model parameters were employed to determine ROIs associated with learning, which then supported further analyses (e.g., connectivity analyses and dynamic causal modeling). Second, depending on the model and/or parameter chosen, different aspects of learning can be probed. For example, if a given model has different parameters mapping onto different computational aspects of the learning process, such as associative strength, prediction, or prediction error, it is possible that trial-by-trial changes in these parameters can each be used as regressors to examine whether dissociable neural regions are involved in these different computational aspects of learning. Thus, employing computational models allows for the possibility of disassociating different processes involved in learning (see Section 2 and Figure 2 for a discussion on this topic). Third, as reviewed above, it is difficult to gather and use trial-by-trial measures of learning to undercover the neural changes correlated with this learning time-course. This is the case even if behavioral measures are captured on every trial. Moreover, employing computational models allows participants to simply participate in the task without significant behavioral intervention. As reviewed above, one of the disadvantages of significant behavioral intervention is that it could potentially affect learning, especially in incidental or implicit learning tasks. In sum, there are a number of major strengths to using trial-by-trial estimates of learning provided by behaviorally validated computational models to uncover the time-course of learning available in fMRI recordings.

There are, however, a number of limitations to employing computational models in this way, most notably the lack of direct behavioral evidence of learning. As noted earlier in the case of den Ouden et al. (2009), no behavioral measures of learning were gathered at all, not even after scanning was complete. While computational models like the R-W model have received extensive behavioral validation, certainly not collecting any measures of learning is problematic for a number of reasons that are the topic of this review. For example, there are substantial individual differences in learning outcomes, especially with incidental learning tasks, and unless this variability is directly a result of differences in the input, this will not be reflected in a model like the R-W model. Relatedly, this approach assumes that the computational model provides a good description of the average time-course of learning at the group level, an assumption that must be validated in additional behavioral studies. However, even with such behavioral validation, this approach does not attempt to account for individual differences that the model was not designed to predict. One possibility for future work is to combine intermittent behavioral testing (as described in the previous section) with computational models to have trial-by-trial estimates of learning constrained by an individual's learning trajectory.

In sum, in this section we have proposed that one emerging method for uncovering the time-course of neural activity associated with learning consists of combining computational models with fMRI recordings. Specifically, computational models can provide trial-by-trial estimates of learning. By applying these models to the input that the learner receives in the scanner, the changing parameters of the model (e.g., associative strength, prediction error, Bayesian priors) can be used to generate predictors that are mapped onto the neural signal, and the neural activity associated with the average time-course of learning can be elucidated. While this is a relatively new method with its own trade-offs and limitations, some of which are discussed above, this is a promising avenue for leveraging fMRI to uncover the process of learning rather than focusing solely on learning outcomes.

5. Examining the Earliest Stages of Learning: Before Behavioral Evidence

A major focus of this article has been the use of behavioral measures to uncover the neural systems supporting the process of learning. However, there are likely changes in neural activity *before* there can be behavioral evidence of learning, presumably during initial exposure to structured stimuli. It is difficult to examine the time-course of these earliest stages of learning as they necessarily occur before corresponding behavioral changes are evident. For example, in the absence of behavioral change, it would be difficult to differentiate participants with a slower than average learning time-course (but who are successfully engaging in the earliest stages of learning) from those who fail to learn no matter how much training they receive. However, if the field continues to examine the neural correlates of the online process, rather than the outcomes, of learning, it will become increasingly important to address the issue of how to examine learning before it gives rise to behavioral changes.[4]

In general, this question is most pressing for forms of learning with a protracted time-course. Some forms of learning occur on a comparatively shorter time scale: for example, priming and episodic memory result in behavioral change after a single trial or very few trials. For types of learning like these, it would be exceedingly difficult to gather time-course information from designs that rely on a slow physiological measure like the hemodynamic

[4] Of course, one reason why participants might fail to show behavioral evidence of learning in a given timeframe is that they fail to pay attention to the stimuli (e.g., fall asleep) or do not follow task instructions. This is a trivial case where behavioral evidence of learning will not arise. Experimenters should take care to ensure that participants are engaged with the task and obeying task instructions. This can be ensured through employing a cover task that is not immediately related to the learning task or "catch" trials to ensure that all participants are properly engaged (Turk-Browne et al., 2010).

response. Instead, learning outcomes for individual items are often compared (e.g., successfully recalled vs. false memory). However, certain types of learning require a lengthy exposure phase before participants show behavioral evidence of having learned. In these circumstances, a participant with slower than average learning time-course could very well not show evidence of learning during the experiment even though they are in fact engaged in the earliest stages of learning.

The relevance of early learning is highlighted by a number of recent studies that failed to show strong behavioral evidence of learning, yet obtained patterns of neural activity suggesting some form of learning-related processing. Turk-Browne et al. (2009) scanned during the presentation of two different visual sequences: One structured sequence whose visual elements were grouped into triplets and thus had very high transitional probabilities (i.e., $p(Y|X)$ where X and Y are two elements of the sequence), and one random sequence where elements were presented in random order and thus all elements are linked by much lower transitional probabilities. As noted earlier, a rich behavioral literature shows that participants can use these transitional probabilities to passively learn which elements cohere within a sequence (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002; Saffran et al., 1996; Turk-Browne, Junge, & Scholl, 2005). However, in an overall post-test measure employed by Turk-Browne et al. (2009)[5], participants failed to differentiate the triplets with high transitional probabilities from triplets with low transitional probabilities, a finding suggesting that participants failed to learn robustly in this task. Nevertheless, comparisons of the neural systems engaged during viewing of the structured vs. random sequences revealed that participants engaged multiple areas of the brain that are robustly associated with learning and memory (e.g., the hippocampus, basal ganglia) as well as a number of cortical regions (e.g. fusiform gyrus, lateral occipital cortex or LOC). Thus, it seems likely that participants were engaged in learning at some level during the structured condition. However, lacking robust behavioral evidence of learning in this particular study means that it is not clear whether participants are at early stages of learning or whether other processes not associated with learning are being differentially engaged across conditions and result in engagement of these neural regions. See also McNealy et al. (2006; 2010) for a parallel example to Turk-Browne et al., (2009) in the area of language learning or auditory statistical learning.

One possibility is that while participants might not show differences in behavioral outcome measures related to the task, they might show more indirect behavioral evidence that learning is taking place. Recent work by Emberson and Amso (2012) points to such a possibility. They recorded both neural activity and eye movements while participants learned to perceive a novel visual object. The study employed both a structured condition (presentation of the novel object in different orientations in different visual scenes) and a random control condition (control scenes that did not contain the novel object). While the behavioral outcome measures did indicate overall learning in the structured condition, a substantial subset of these participants failed to show evidence of learning. Examination of the eye movement patterns revealed that Non-learners in the structured condition more closely resembled Learners (again in the structured condition) compared to Non-learners in the random condition (see Figure 6). Indeed, comparing eye movements between Learners and Non-Learners in the structured condition revealed only very subtle differences in their looking behavior. However, there were pronounced differences in eye movements between Non-Learners in the structured condition and Non-Learners in the random condition. By measuring eye movements (a form of information gathering previously found to be involved in a relevant task, Johnson, Slemmer & Amso, 2004), there was evidence that participants in

[5] Turk-Browne et al., (2009) did however find evidence of learning in the first half of the posttest.

the structured condition performed differently from participants in the random condition, regardless of whether they showed evidence of learning as assessed in behavioral outcome measures (also see Zhao, Ngo, McKendrick & Turk-Browne, 2011).

Thus, we see that even when outcome measures of learning are not reliable, participants exposed to conditions where learning can take place may engage different neural systems from participants exposed to conditions where no learning can take place, and indeed there may be changes in incidental behavioral measures such as eye movements. Such evidence suggests that one possible way to examine the earliest stages of learning (i.e., before behavioral evidence) is to compare neural activations between conditions where learning has been previously demonstrated versus where learning is not possible (e.g., a random condition). In the absence of incidental measures of behavior, the assumption of learnability of the structured condition needs to be supported by additional behavioral studies. However, employing a random condition can allow for an estimate of when and where neural activity diverges between these conditions (i.e., when neural systems presumably associated with learning are engaged, see Turk-Browne et al., 2009 for an implementation of this method). Indeed, these claims are strengthened if a more incidental behavioral measure (e.g., eye movements) can be relied upon to differentiate activity across these conditions without requiring overt behavioral outcome measures.

However, comparing across learnable or structured and non-learnable, control or random conditions has a number of limitations. First, it is not clear what computational aspect of learning is being investigated. If the differences between these conditions tap into the earliest stages of learning (i.e., before overt behavioral evidence), are these indicative of differences in the processes of pattern extraction or of model building (see Figure 2 and Section 2 for a discussion)? While it seems likely that some amount of pattern extraction would precede model building, how much do these processes temporally overlap during learning? As discussed above, these processes of pattern extraction and model-building may be exceedingly difficult to disentangle, especially at the earliest stages of learning, because it is likely that both involve pattern extraction based on structured stimuli as well as model-building based on these structured representations. However, given the clear theoretical and cognitive distinction between these processes and their likely separation at latter stages of learning, it is important to consider both of these processes (pattern extraction and model-building) even under circumstances when they will likely be difficult to disentangle. Returning to the data from Emberson and Amso (2012), it is possible that these differences in eye movements are reflective of differences in pattern extraction, which may be necessary for learning, but are not sufficient. These are open questions that may not be easily addressed using this method alone.

Second, while the comparison of structured and random sequences may be one way to examine learning before behavioral evidence is obtained, one should consider whether the differences between learnable and non-learnable conditions might be resulting in differential engagement of cognitive processes other than learning. For example, in temporally ordered statistical learning studies, random conditions often have very low transitional probabilities between every element (e.g., in a stream of 15 elements a random stream might have an average transitional probability of 0.07). In contrast, structured conditions have on average much higher transitional probabilities (e.g., 0.6–1.0) and a clearly bimodal distribution between high transitional probabilities (e.g., 1.0) and lower transitional probabilities (e.g., 0.25). While there certainly need to be systematic differences in statistical information between learnable and non-learnable conditions in a statistical learning study, these differences might be engaging additional unintended processes. Indeed, recent research has suggested that attentional mechanisms are differentially engaged depending on stimulus predictability, with the best engagement of attention at intermediate levels of predictability (i.e., not too low or random and not too high or entirely predictable; Kidd, Piantadosi, & Aslin, 2012). Thus, a learnable condition might be reflexively and unconsciously garnering more attention than a non-learnable condition, and this might be an important aspect of the design to control (e.g., by having a non-learnable condition with an additional attentionally-demanding cover task like detecting a certain stimulus). While differences in the statistical information across learnable and non-learnable conditions is simply one example of systematic differences that may affect learning during the task, a deeper examination of the effects of these systematic differences that lead to learning or not will aid the comparison of neural systems engaged across these conditions and provide insight into the areas of the brain involved in the earliest stages of learning.

In sum, this section considered the difficult question of how to examine the process of learning before there is behavioral evidence. This issue comes into play as researchers continue to use fMRI to examine the process as opposed to the outcome of learning. To this end, the field must attempt to differentiate between participants who are engaged in the processes which will ultimately support a positive learning outcome but who have not shown behavioral evidence of learning, and those who are not engaging learning processes at all. To foster discussion of this topic, we considered how the inclusion of a non-learnable or random condition can help to tap into processes that separate early from later stages of learning, and how the use of more indirect behavioral measures such as eye tracking can provide evidence for early learning vs. non-engagement of learning mechanisms. However, it is important to consider the possibility that uncovering differences associated with early vs. late learning may reflect other component processes of learning that are being engaged (e.g., pattern extraction vs. model building). This is because the relative temporal onsets of these processes are unknown and these processes may or may not be engaged when exposed to learnable vs. non-learnable conditions.

6. Conclusion

This review has both explored current fMRI analysis techniques employed to study the neural basis of learning and discussed promising avenues for methodological improvement. In describing the depth, complexity, and variability of the learning process over time, we fully acknowledge the immense challenges surrounding this field of study, particularly in light of the basic limitations of fMRI (i.e., fairly coarse temporal resolution). Nevertheless, we propose that the basic assumptions implicitly forming the foundation of many imaging methods and analysis techniques require behavioral validation. To be clear, our proposals for improvement have evolved from over a decade of valuable neuroimaging work dedicated to the topic of learning. Our current knowledge has been built on investigations into how the human brain taps into acquired rule-systems during test, how it reacts to the violation of expectations, and how it responds differentially to patterned and random stimuli[6]. The questions probed within this body of work are important and interesting ones, but they cannot advance our understanding of the *process* of learning without a focus on methods of studying learning from its earliest stages onward.

To conclude, we maintain that the fMRI studies which form the basis of our understanding of the neural systems supporting human learning have been shaped by two basic, untested assumptions: (1) the process of learning and the results of learning are mediated by largely overlapping neural substrates[7] and (2) the time-course of learning is relatively uniform both within and between subjects. Correspondingly, these assumptions must be either minimized through improved experimental designs or validated by direct hypothesis-testing. It is our

[6] Indeed, many of the studies reviewed here point to at least some overlap in frontal, striatal, and hippocampal systems engaged during a broad spectrum of learning-related tasks.

hope that this review will begin a dialogue within the field and encourage more research that examines the process of learning by considering the interrelation of behavioral measures and fMRI recording during acquisition in a learning task.

Text Box 1

Relevant Behavioral Paradigms

In studies of human learning, the most common experimental design consists of two phases: an exposure phase to a set of structured stimuli and a subsequent testing phase (posttest). The extent to which participants have extracted, maintained and retrieved information acquired during the exposure phase is evaluated during the posttest. To uncover supporting neural systems, studies employing fMRI have collected imaging data at various stages in this process, ranging from initial exposure to final posttest, and often after some level of acquisition has taken place outside the scanner. This experimental method has been extensively applied to various forms of **pattern learning** including, but not limited to, ***artificial grammar learning*** (AGL), visual and auditory ***statistical learning*** (SL), and ***motor sequence learning*** (MSL). Because learning paradigms within these different literatures comprise the majority of studies discussed in this review, we detail the basic methods and principles of these paradigms here (Figure 1).

Pioneered by Reber (1967), the classic **artificial grammar learning (AGL)** paradigm involves exposure to a set of letter strings generated by a finite state grammar. The rule-governed relationships between elements in the strings are intended to be so complex as to preclude explicit learning of the underlying grammar. Nevertheless, after an acquisition period, participants shown novel grammatical and ungrammatical letter sequences are able to discriminate the test items (strings or partial strings) significantly better than chance performance (e.g., Dienes, Broadbent, & Berry, 1991; Knowlton & Squire, 1996).

In a related vein, **statistical learning (SL)** of both visual (e.g., Fiser & Aslin, 2002, 2005; Fiser, Scholl & Aslin, 2007) and auditory stimuli (e.g., Newport & Aslin, 2004; Gebhart, Newport & Aslin, 2009) entails the extraction of regularities, such as the transitional probabilities between elements, in order to form structured representations of initially unfamiliar exposure stimuli. Following a period of passive exposure, participants discriminate statistically consistent combinations of items from combinations that violate these statistical regularities. Learning takes place without conscious intent or awareness; participants are not explicitly tallying up co-occurrence frequencies during the exposure and dividing them by individual element frequencies (p (Y|X)). Correspondingly, they cannot concretely verbalize why some combinations seem more familiar at test. Studies by Saffran, Newport, and Aslin (1996) with adults and Saffran, Aslin, and Newport (1996) with infants demonstrated that learners are able to use the statistical relationships between adjacent syllables in order to segment word-like units from a continuous stream of nonsense syllables.

Finally, **motor sequence learning (MSL)** encompasses a diverse set of tasks that typically measure learning by recording reaction time and/or accuracy on a sequence

[7] There are notable exceptions to this in the field of memory where there is a focus on the non-overlapping neural regions involved in encoding and retrieval of memories (e.g., the HERA model, Habib, Nyberg, & Tulving, 2003). However, as discussed above (e.g., Textbox 1, Section 5), the current review is focused on different behavioral phenomena than those typically examined in the field of memory where stimuli are seen and remembered through single or a very few identical experiences. The learning paradigms examined in this review unfold over a, comparatively, lengthy period of time and can incorporate variability in presentation to create relative probabilities (e.g., some items follow each other with a high probability but others are marked by variable or random order). Despite these obvious differences between explicit tasks and the types of implicit learning discussed in the current paper, any future steps may benefit from knowledge of other fields that are already making this important distinction.

generation or completion task. MSL tasks range in difficulty from simple skill learning such as touching the thumb to the fingers in a repeated 5-item sequence (e.g., Karni et al., 1995) to executing a timed motor response to visual elements that occur in a complex ordering (e.g., Rauch et al., 1997; also known as a serial reaction time task). MSL tasks vary widely in the extent of their implicit learning component, with some studies involving explicit instruction to learn (Lehericy et al., 2005), others requiring the subject to learn by trial-and-error (Sakai et al., 1998), and still others in which subjects are unaware of the patterned nature governing sequence order (Hunt & Aslin, 2001; Nissen & Bullemer, 1987).

Acknowledgments

The authors would like to acknowledge Dr. Elissa Newport for helpful comments on an earlier draft of this article and to funding sources: NIH R01 HD-37082 to RA and EN.

References

Abla D, Katahira K, Okanoya K. Online assessment of statistical learning by event-related potentials. Journal of Cognitive Neuroscience. 2008; 20:952–964. [PubMed: 18211232]

Bayley P, O'Reilly R, Curran T, Squire L. New semantic learning in patients with large medial temporal lobe lesions. Hippocampus. 2008; 18:575–583. [PubMed: 18306299]

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF. Associative learning of social value. Nature. 2008; 456:245–249. [PubMed: 19005555]

Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nature Neuroscience. 2007; 10:1214–1221.

Belliveau J, Kennedy D, McKinstry R, Buchbinder B, Weisskoff R, Cohen M, et al. Functional mapping of the human visual cortex by magnetic resonance imaging. Science. 1991; 254:716–719. [PubMed: 1948051]

Caramazza A, Badecker W. Clinical syndromes are not God's gift to cognitive neuropsychology : A reply to a rebuttal to an answer to a response to the case against syndrome-based research. Brain & Cognition. 1991; 16:211–227. [PubMed: 1930976]

Cunillera T, Càmara E, Toro JM, Marco-Pallares J, Sebastián-Galles N, Oritz H, Pujol J, Rodríguez-Fornells A. Time course and functional neuroanatomy of speech segmentation in adults. Neuroimage. 2009; 48:541–553. [PubMed: 19580874]

Daniel R, Pollmann S. Striatal activations signal prediction errors on confidence in the absence of external feedback. Neuroimage. 2012; 59:3457–67. [PubMed: 22146752]

Dayan P, Daw ND. Decision theory, reinforcement learning and the brain. Cognitive, Affective, & Behavioral Neuroscience. 2008; 8:429–453.

Dienes Z, Broadbent D, Berry D. Implicit and explicit knowledge bases in artificial grammar learning. Journal of Experimental Psychology: Learning, Memory and Cognition. 1991; 17:875–887.

den Ouden HEM, Daunizeau J, Roiser J, Friston K, Stephan K. Striatal prediction error modulates cortical coupling. Journal of Neuroscience. 2010; 30:3210–3219. [PubMed: 20203180]

den Ouden HEM, Friston KJ, Daw N, McIntosh AR, Stephan KE. A dual-role for prediction error in associative learning. Cerebral Cortex. 2009; 19:1175–1851. [PubMed: 18820290]

Dolan RJ, Fletcher PF. Encoding and retrieval in the human medial temporal lobes: An empirical investigation using functional magnetic resonance imaging. Hippocampus. 1999; 9:25–34. [PubMed: 10088897]

Doyon J, Bellec P, Amsel R, Penhune VB, Monchi O, Benali H. Contributions of the basal ganglia and functionally related brain structures to motor learning. Behavioral and Brain Research. 2009; 199:61–75.

Eichenbaum, H.; Cohen, N. From conditioning to conscious recollection: Memory systems of the brain. USA: Oxford University Press; 2001.

Elman JL. Finding structure in time. Cognitive Science. 1990; 14:179–211.

Emberson LL, Amso D. Learning to sample: Eye tracking and fMRI indices of changes in object perception. Journal of Cognitive Neuroscience. 2012; 24:2030–2042. [PubMed: 22721373]

Fiser J, Aslin RN. Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences. 2002; 99:15822–15826.

Fiser J, Aslin RN. Encoding multi-element scenes: Statistical learning of visual feature hierarchies. Journal of Experimental Psychology: General. 2005; 134:521–537. [PubMed: 16316289]

Fiser J, Scholl BJ, Aslin RN. Perceived object trajectories during occlusion constrain visual statistical learning. Psychological Bulletin and Review. 2007; 14:173–178.

Forkstam C, Hagoort P, Fernandez G, Ingvar M, Petersson KM. Neural correlates of artificial syntactic structure classification. Neuroimage. 2006; 32:956–967. [PubMed: 16757182]

Forkstam C, Petersson KM. Towards an explicit account of implicit learning. Current Opinion in Neurobiology. 2005; 18:435–441.

Furl N, Kumar S, Alter K, Durrant S, Shawe-Taylor J, Griffiths T. Neural prediction of higher-order auditory sequence statistics. Neuroimage. 2011; 54:2267–2277. [PubMed: 20970510]

Gebhart AL, Newport EL, Aslin RN. Statistical learning of adjacent and non-adjacent dependencies among nonlinguistic sounds. Psychonomic Bulletin & Review. 2009; 16:486–490. [PubMed: 19451373]

Gershman SJ, Niv Y. Learning latent structure: Carving nature at its joints. Current Opinion in Neurobiology. 2010; 20:251–256. [PubMed: 20227271]

Gheysen F, Van Opstal F, Roggeman C, Van Waelvelde H, Fias W. The neural basis of implicit perceptual sequence learning. Frontiers in Human Neuroscience. 2011; 5:1–12. [PubMed: 21283556]

Habib R, Nyberg L, Tulving E. Hemispheric asymmetries of memory: the HERA model revised. Trends in Cognitive Sciences. 2003; 7:241–245. [PubMed: 12804689]

Haykin, S. Neural networks: A comprehensive foundation. 2. Upper Saddle River, NJ: Prentice-Hall; 1998.

Hunt RH, Aslin RN. Statistical learning in a serial reaction time task: Simultaneous extraction of multiple statistics. Journal of Experimental Psychology: General. 2001; 130:658–680. [PubMed: 11757874]

Johnson S, Slemmer J, Amso D. Where infants look determines how they see: Eye movements and object perception performance in 3-month-olds. Infancy. 2004; 6:185–201.

Karni A, Meyer G, Jezzard P, Adams MM, Turner R, Ungerleider LG. Functional MRI evidence for adult motor cortex plasticity during motor skill learning. Nature. 1995; 377:155–158. [PubMed: 7675082]

Karuza EA, Newport EL, Aslin RN, Starling SJ, Tivarus ME, Bavelier D. Neural correlates of statistical learning in a word segmentation task: An fMRI study. Brain and Language. 2013; 127:46–54. [PubMed: 23312790]

Kidd C, Piantadosi S, Aslin R. The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PloS One. 2012; 7:e36399. [PubMed: 22649492]

Kirkham N, Slemmer J, Johnson S. Visual statistical learning in infancy: Evidence for a domain general learning mechanism. Cognition. 2002; 83:B35–B42. [PubMed: 11869728]

Knowlton BJ, Squire LR. Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. Journal of Experimental Psychology: Learning and Memory Cognition. 1996; 22:169–181.

Kruschke J. Locally Bayesian learning with applications to retrospective revaluation and highlighting. Psychological Review. 2006; 113:677–698. [PubMed: 17014300]

Lehéricy S, Benali H, Van de Moortele PF, Pélégrini-Issac M, Waechter T, Ugurbil K, et al. Distinct basal ganglia territories are engaged in early and advanced motor sequence learning. Proceedings of the National Academy of Sciences. 2005; 102:12566–12571.

Lieberman MD, Chang GY, Chiao J, Bookheimer SY, Knowlton BJ. An event-related fMRI study of artificial grammar learning in a balanced chunk strength design. Journal of Cognitive Neuroscience. 2004; 16:427–438. [PubMed: 15072678]

Logothetis NK, Wandell BA. Interpreting the bold signal. Annual Review of Psychology. 2004; 66:735–769.

McNealy K, Mazziotta JC, Dapretto M. Cracking the language code: Neural mechanisms underlying speech parsing. Journal of Neuroscience. 2006; 26:7629–7639. [PubMed: 16855090]

McNealy K, Mazziotta JC, Dapretto M. The neural basis of speech parsing in children and adults. Developmental Science. 2010; 13:385–406. [PubMed: 20136936]

Miller R, Barnet RC, Grahame NJ. Assessment of the Rescorla-Wagner model. Psychological Bulletin. 1995; 117:363–386. [PubMed: 7777644]

Muehlhan M, Lueken U, Wittchen HU, Kirschbaum C. The scanner as a stressor: Evidence from subjective and neuroendocrine stress parameters in the time course of an functional magnetic resonance imaging session. International Journal of Psychophysiology. 2011; 79:118–126. [PubMed: 20875462]

Newport EL, Aslin RN. Learning at a distance I. Statistical learning of non-adjacent dependencies. Cognitive Psychology. 2004; 48:127–162. [PubMed: 14732409]

Nissen MJ, Bullemer P. Attentional requirements of learning: evidence from performance measures. Cognitive Psychology. 1987; 19:1–32.

Norman K, O’Reilly R. Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. Psychological Review. 2003; 110:611–645. [PubMed: 14599236]

Ogawa S, Tank DW, Menon R, Ellermann JM, Kim SG, Merkle H, et al. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proceedings of the National Academy of Sciences. 1992; 89:5951–5955.

Opitz B, Rinne T, Mecklinger A, von Cramon DY, Schröger E. Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. Neuroimage. 2002; 15:167–174. [PubMed: 11771985]

Orban G, Aslin RN, Fiser J, Lengyel M. Occam’s razor at work: The dynamics of visual chunk learning. in preparation.

Pavlov, IP. Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Anrep, GV., translator and editor. London: Oxford University Press; 1927.

Petersson KM, Folia V, Hagoort P. What artificial grammar learning reveals about the neurobiology of syntax. Brain and Language. 2012; 120:83–95. [PubMed: 20943261]

Petersson KM, Forkstam C, Ingvar M. Artificial syntactic violations activate Broca’s region. Cognitive Science. 2004; 28:383–407.

Poldrack R, Clark J, Pare-Blagoev E, Shohamy D, Moyano J, Myers C, et al. Interactive memory systems in the human brain. Nature. 2001; 414:546–550. [PubMed: 11734855]

Poldrack R, Prabhakaran V, Seger C, Gabrieli J. Striatal activation during acquisition of a cognitive skill. Neuropsychology. 1999; 13:564–574. [PubMed: 10527065]

Rauch SL, Whalen PJ, Savage CR, Curran T, Kendrick A, Brown HD, et al. Striatal recruitment during an implicit sequence learning task as measured by functional magnetic resonance imaging. Human Brain Mapping. 1997; 5:124–132. [PubMed: 10096417]

Reber AS. Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior. 1967; 6:855–863.

Reber, AS. Implicit learning and tacit knowledge: An essay on the cognitive unconscious. New York: Oxford University Press; 1993.

Reber AS, Walkenfeld FF, Henstad R. Implicit and explicit learning: individual differences and IQ. Journal of Experimental Psychology: Learning, Memory and Cognition. 1991; 17:888–896.

Reeder PA, Aslin RN, Newport EL, Bavelier D. Learning-rate differences in expert players of first-person-shooter video games. in preparation.

Rescorla, R.; Wagner, A. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.; Prokasy, W., editors. Classical conditioning II: Current research and theory. New York: Appleton-Century-Crofts; 1972. p. 64-99.

Rescorla RA. Behavioral studies of Pavlovian conditioning. Annual Review of Neuroscience. 1988; 11:329–352.

Ross R, Brown T, Stern C. The retrieval of learned sequences engages the hippocampus: Evidence from fmri. Hippocampus. 2009; 19:790–799. [PubMed: 19219919]

Saffran J, Aslin R, Newport E. Statistical learning by 8-month-old infants. Science. 1996; 274:1926–1928. [PubMed: 8943209]

Saffran JR, Newport EL, Aslin RN. Word segmentation: The role of distributional cues. Journal of Memory and Language. 1996; 35:606–621.

Sakai K, Hikosaka Ok, Miyauchi S, Takino R, Sasaki Y, Putz B. Transition of brain activation from frontal to parietal areas in visuomotor sequence learning. Journal of Neuroscience. 1998; 18:1827–1840. [PubMed: 9465007]

Sakamoto Y, Jones M, Love BC. Putting the psychology back into psychological models: Mechanistic versus rational approaches. Memory & Cognition. 2008; 36:1057–1065. [PubMed: 18927024]

Schendan HE, Searl MM, Melrose RJ, Stern CE. An fMRI study of the role of the medial temporal lobe in implicit and explicit sequence learning. Neuron. 2003; 37:1013–1025. [PubMed: 12670429]

Schultz W, Dayan P, Montague PR. A neural substrate for prediction and reward, Science. 1997; 275:1593–1599.

Seger CA, Cincotta CM. Dynamics of frontal, striatal, and hippocampal systems in rule learning. Cerebral Cortex. 2006; 16:1546–1555. [PubMed: 16373455]

Seger CA, Prabhakaran V, Poldrack RA, Gabrieli JDE. Neural activity differs between explicit and implicit learning of artificial grammar strings: An fMRI study. Psychobiology. 2000; 28:283–292.

Shadmehr R, Holcomb HH. Neural correlates of motor memory consolidation. Science. 1997; 277:821–825. [PubMed: 9242612]

Shanks DR, St John MF. Characteristics of dissociable human learning systems. Behavioral and Brain Sciences. 1994; 17:367–447.

Shohamy D, Wagner A. Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. Neuron. 2008; 60:378–389. [PubMed: 18957228]

Skinner, BF. The behavior of organisms: An experimental analysis. New York: Appleton Century; 1938.

Skosnik PD, Mirza F, Gitelman DR, Parrish TB, Mesulam MM, Reber PJ. Neural correlates of artificial grammar learning. Neuroimage. 2002; 17:1306–1314. [PubMed: 12414270]

Sutton RS. Learning to predict by the methods of temporal differences. Machine Learning. 1998; 3:9–44.

Thomas KM, Hunt RH, Vizueta N, Sommer T, Durston S, Yang Y, et al. Evidence of developmental differences in implicit sequence learning: an fMRI study of children and adults. Journal of Cognitive Neuroscience. 2004; 16:1339–1351. [PubMed: 15509382]

Thorndike, EL. Animal intelligence: Experimental studies. New York: Macmillan; 1911.

Thorndike, EL. Human learning. New York: Century Co; 1931.

Tolman, EC. Purposive behavior in animals and men. Berkeley, CA: University of California Press; 1951.

Turk-Browne N, Jungé J, Scholl B. The automaticity of visual statistical learning. Journal of Experimental Psychology: General. 2005; 134:552–563. [PubMed: 16316291]

Turk-Browne N, Scholl B, Chun M, Johnson MK. Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. Journal of Cognitive Neuroscience. 2009; 21:1934–1945. [PubMed: 18823241]

Turk-Browne NB, Scholl BJ, Johnson MK, Chun MM. Implicit perceptual anticipation triggered by statistical learning. Journal of Neuroscience. 2010; 30:11177–11187. [PubMed: 20720125]

van der Graaf FHCE, Maguire RP, Leenders KL, de Jong BM. Cerebral activation related to implicit sequence learning in a double serial reaction time task. Brain Research. 2006; 1081:179–190. [PubMed: 16533501]

Vaughn KE, Rawson KA. Diagnosing criterion level effects on memory: What aspects of memory are enhanced by repeated retrieval? Psychological Science. 2011; 22:1127–1131. [PubMed: 21813798]

Waelti P, Dickinson A, Schultz W. Dopamine responses comply with basic assumptions of formal learning theory. Nature. 2001; 412:43–48. [PubMed: 11452299]

Willingham DB, Salidis J, Gabrieli JD. Direct comparison of neural systems mediating conscious and unconscious skill learning. Journal of Neurophysiology. 2002; 88:1451–1460. [PubMed: 12205165]

Wörgötter F, Porr B. Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. Neural Computation. 2005; 17:245–319. [PubMed: 15720770]

Yang J, Li P. Brain Networks of Explicit and Implicit Learning. PLoS ONE. 2012; 7:e42993. [PubMed: 22952624]

Yarkoni T, Barch DM, Gray JR, Conturo TE, Braver TS. BOLD correlates of trial-by-trial response time variability in gray and white matter: A multi-study fMRI analysis. PLoS ONE. 2009; 4:e4257. [PubMed: 19165335]

Zevin J, Yang J, Skipper J, McCandliss B. Domain general change detection accounts for “dishabituation” effects in temporal–parietal regions in functional magnetic resonance imaging studies of speech perception. Journal of Neuroscience. 2010; 30:1110–1117. [PubMed: 20089919]

Zhao J, Ngo N, McKendrick R, Turk-Browne NB. Mutual interference between statistical summary perception and statistical learning. Psychological Science. 2011; 22:1212–1219. [PubMed: 21852450]

Zurif E, Swinney D, Fodor JA. An evaluation of assumptions underlying the single-patient-only position in neuropsychological research: A reply. Brain and Cognition. 1991; 16:198–210. [PubMed: 1789838]

Highlights

- We examine the current use of fMRI to study the neural basis of human learning.
- We describe challenges involved in studying the process of learning with fMRI.
- We discuss the value of relating neural response to behavior over time.

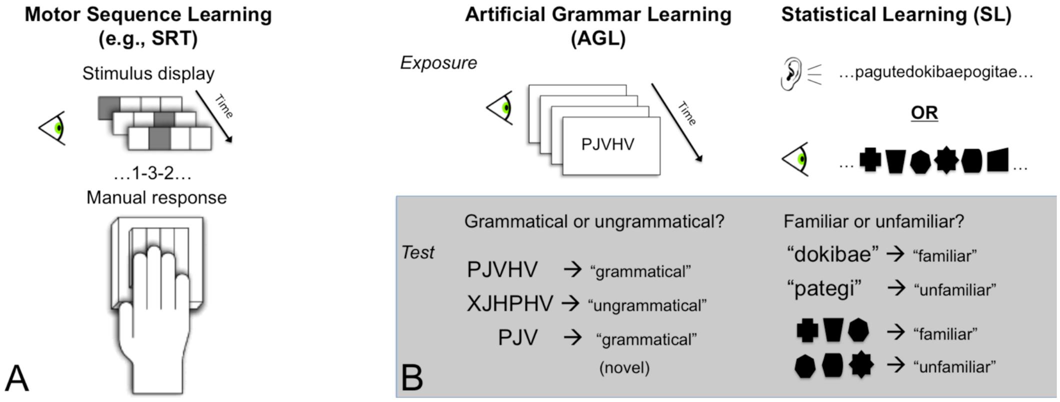

Figure 1.

Structure of various learning paradigms relevant to this review. (A) *Motor Sequence Learning:* Participants initiate a motor response to temporally-patterned visual stimuli. (B) *Artificial Grammar Learning* and *Statistical Learning:* Participants undergo an exposure phase during which they are presented with finite-state grammar sequences (AGL) or probabilistic auditory/visual patterns (SL). In a subsequent test phase, they make acceptability judgments on structured and unstructured test items.

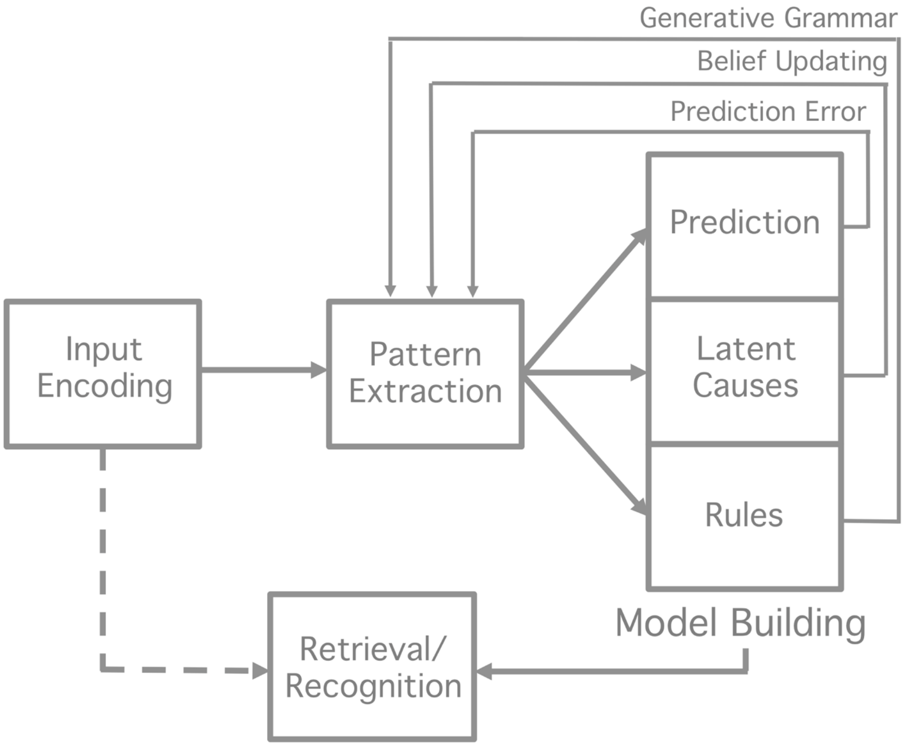

Figure 2.
A generic architecture of the processes involved in learning for purposes of the present discussion: 1) sensory or input encoding; 2) pattern extraction; 3) model building; 4) retrieval or recognition process. The latter process is denoted by a dashed line and involves a match process between a current sensory input and a stored representation resulting from learning and memory processes. This process also likely involves a decision component.

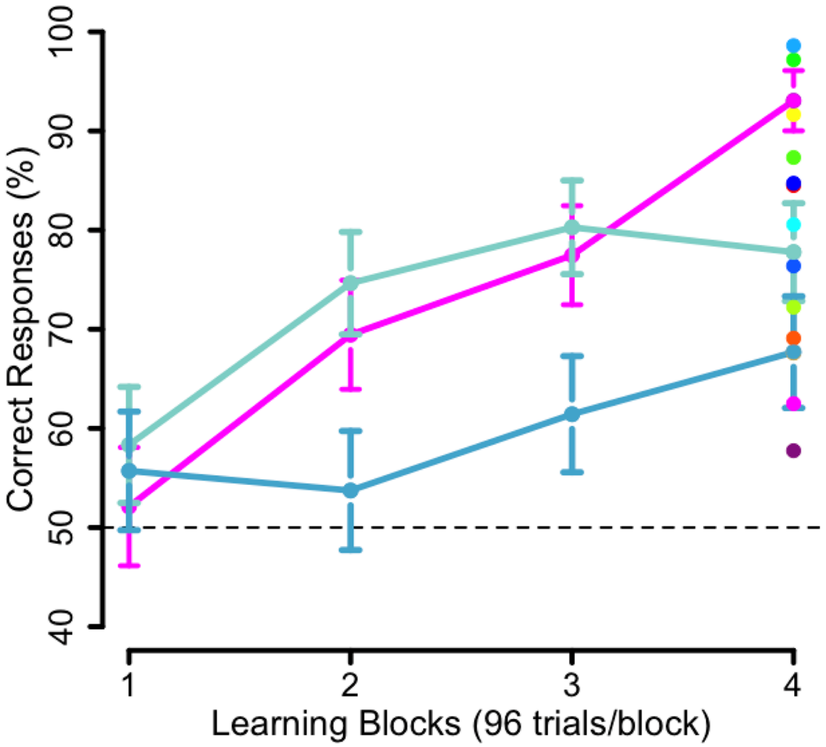

Figure 3.

Comparing learning outcomes (performance on the final block of a learning task) with overall learning trajectories in a replication of Shohamy and Wagner (2008). On the right of panel, dots represent individual performance on the final block of learning or learning outcomes for all participants'. The learning trajectories of three participants are presented. The two participants presented in blue have similar learning outcomes but divergent learning trajectories. The participants represented in dark blue and pink have very different learning outcomes despite having nearly identical learning trajectories for the first three blocks. Error bars represent the standard error of the mean.

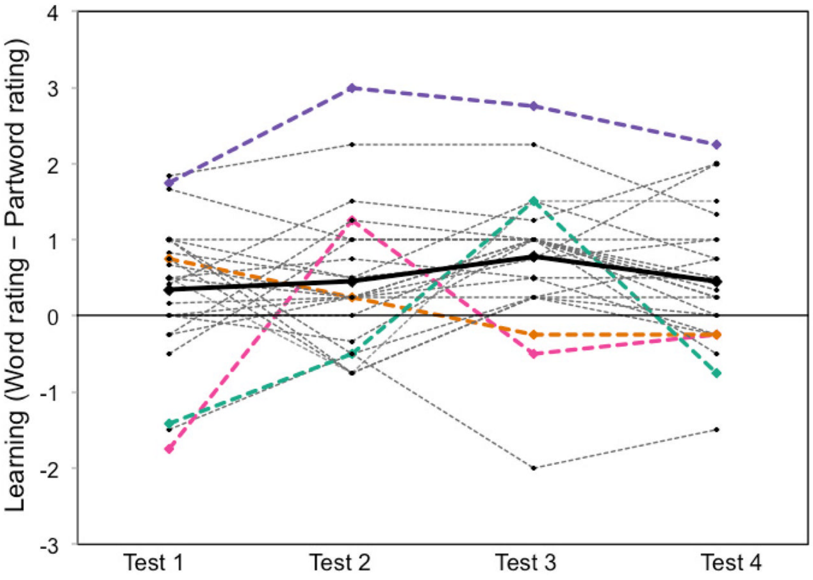

Figure 4.

Variability in individual learning trajectories (adapted from Karuza et al. (2013)). Data points plotted above were calculated as the word rating - partword rating (i.e., the extent of learning) for each test. A high score indicates that a participant was able to successfully discriminate words from partwords, whereas a score at or below zero indicates a failure to learn or to maintain previously acquired knowledge. A score below zero indicates that a participant rated partwords as more familiar than words. The bolded black line represents mean performance. Selected individual participants are presented in color to allow for examination of their individual trajectories.

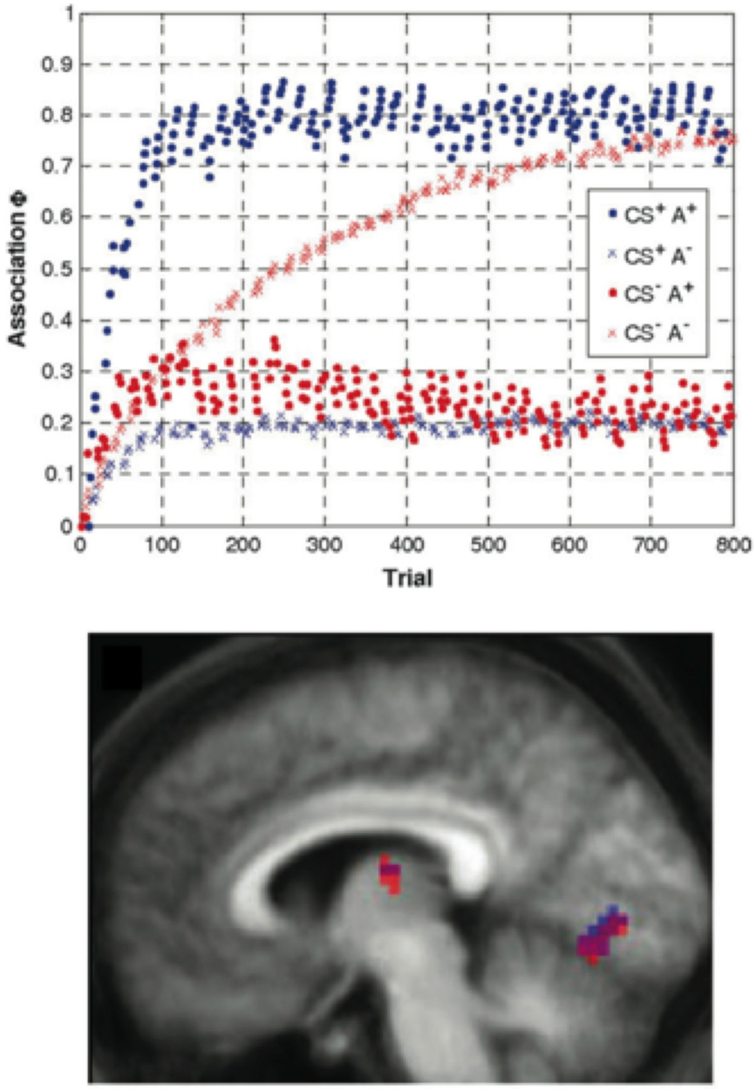

Figure 5.

Classic computational model (Rescorla-Wagner model) examining the effects of incidental, associative learning on cortical connectivity (from den Ouden et al. A dual-role for prediction error in associative learning. *Cerebral Cortex, 2009, 19*, pp. 178, 180, by permission of Oxford University Press). Top panel: parameter estimates of associative strength from the model for each stimulus condition in a 2x2 design (presence or absences of the conditioned visual stimulus, CS+ vs CS-, and presence or absence of the predictive auditory stimulus, A+ vs. A-). Bottom panel: a subset of the regions of interest that showed modulation of activity by experimental condition but also model-estimated associative strength.

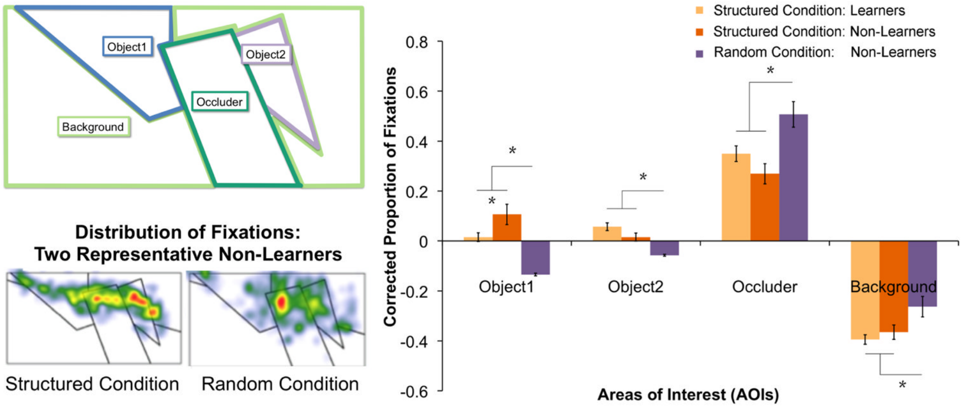

Figure 6.

Distribution of eye movements (fixations) for participants in learnable (Structured) and non-learnable (Random) conditions who showed evidence of learning in outcome measures (Learners) and those who did not (Non-Learners). Top left panel: Areas of Interest for the critical visual stimulus in Emberson and Amso (2012). Right Panel: Proportion of fixations for each AOI for Learners in the Structured condition, Non-learners in the Structured condition and Non-learners in the Random condition. Note: there were extremely few Learners in the Random Condition and thus they are not depicted. Since Areas of Interest (AOIs) were different sizes, proportions of looking were normalized relative to the AOI's proportion of the scene (corrected proportion of fixations) and thus zero can be considered baseline expected looking if eye movements were randomly distributed in the scene. There are significant differences in looking between Learners vs. Non-Learners in the Structured Condition for Object1 only. However, there are significant differences for all other AOIs between both outcome groups in the Structured Condition and the Non-Learners in the Random condition. Bottom left panel: depiction of the distribution of fixations for representative Non-learners in each exposure condition (Structured and Random).