Evaluating Differential Evidence of Multiple Hypothesis Tracking in Statistical Learning

Jordan B. Gunn

Princeton University

Author Note

This thesis was supported by the advice and counsel of Prof. Nicholas Turk-Browne and generated to satisfy the second part of the independent work requirement for fourth-year students concentrating in Princeton University's department of psychology.

Contents

Abstract

Statistical learning refers to describe the ability to detect and use statistical structure. The capacity is not just a feature of human intelligence; it is ubiquitous and effective across lifespans, domains, sensory modalities and levels of consciousness. I evaluate two competing models – the multiple hypotheses and propose-yet-verify models – of one variant of the process, as well as the possibility that neither sufficiently explain how statistical learning is carried out. First the literature on statistical learning is reviewed, beginning with an accounting of its generality and efficacy across situations as well constraints on that efficacy, and then focusing on theoretical issues surrounding its mechanism and how the process achieves its broad functionality. Different kinds of statistical learning are distinguished and the thesis is focused on the kind that involves quickly identifying perceptual cues to aspects of statistical structure of the input that are not directly perceptible, especially word meanings. The two models just mentioned are tested through experiments aimed at revealing traces of alternative hypothesis tracking in different ways and at different points along the course of statistical learning, providing substantial new evidence for the multiple hypotheses account of statistical learning. These results and their meaning for future theoretical and experimental work for understanding statistical learning more broadly are discussed and evaluated.

Introduction

Suppose a goal-oriented agent like you or me were plopped into an environment with no contingencies it could detect, an environment where no event could be discerned to have anything to do with what had already happened or might happen next. The agent might try selecting actions in pursuit of its goals, but the effort would be broadly pointless. Nothing it perceived would reliably indicate anything about what's going on in the outside world, so it would be impossible to discriminate how one action might impact progress toward the agent's goals compared to another action. If anything, all one could learn is that there's not much to learn.

The apparent futility of life in the world just described showcases just how fundamental statistical learning, the cognitive capacity to rapidly detect and exploit statistical structure in our environments, is to adaptive behavior. The capacity emerges in a variety of contexts across species. For example, the diversity of flowers is immense, but bees still exploit floral cues such as odor, shape and color to find and extract food from them (Sandoz, 2011). Similarly, by sensing a component of fish mucus, sea urchin larvae anticipate the presence of predators and bud off clones of themselves that are too small for the fish to see (Gilbert, 2012). The sort of statistical learning emphasized in most cognitive science research, though, is the kind first demonstrated in an experiment where infants around 8-months old demonstrated an ability to learn statistical relationships between the syllables of "pseudowords" after only two minutes of exposure to a speech stream containing them (Saffran, Aslin & Newport, 1996). This kind of statistical learning, mostly observed in humans, occurs over very short time scales (in this case, only two minutes). It also apparently happens automatically and effortlessly, without exercise of cognitive control; for example, statistical learning of mean size is not disrupted when cover tasks

require either distributed or global attention to an artificial environment, though focusing attention on specific objects within an environment could disrupt learning (Triesman & Chong, 2004). Finally, it enables performance in domains central to human behavior: along with syllables in language, learners have displayed sensitivity to sequential conditional statistics in tactile, visual and tonal stimuli (Thiessen, Kronstein, & Hufnagle, 2013). Furthermore, this kind of sensitivity has been found to facilitate visual search, contextual cuing, visuomotor learning, conditioning and general predictive behavior (Frost, Armstrong, Siegelman & Christiansen, 2014).

As statistical learning thus characterized is so powerful a feature of human cognition, much recent research has focused on understanding just *how* it happens – its mechanism. One locus of such study has centered on word-referent learning, the process whereby individuals learn the meaning of words through patterns of experience ambiguously pairing terms with their referents Medina, Snedeker, Trueswell, & Gleitman, 2011). The literature distinguish between and evaluates two widely discussed models of word-referent pair matching: the multiple hypotheses model of statistical learning, which asserts that numerous hypotheses are held in mind with changing weights that track co-occurrences within stimuli, and the propose-yet-verify model of statistical learning, which asserts that learners only hypothesize a single meaning based on their first encounter with a word – neither weighting nor storing back-up alternative meanings along the way.

This thesis is about evaluating these competing models, as well as the possibility that neither sufficiently explain how statistical learning is carried out. First the literature on statistical learning is reviewed, beginning with an accounting of its generality and efficacy across situations as well constraints on that efficacy, and then focusing on theoretical issues surrounding its

mechanism and how the process achieves its broad functionality. Different kinds of statistical learning are distinguished and the thesis is focused on the kind that involves quickly identifying perceptual cues to aspects of statistical structure of the input that are not directly perceptible, especially word meanings. The two models just mentioned are tested through experiments aimed at revealing traces of alternative hypothesis tracking in different ways and at different points along the course of statistical learning. Finally, the results of these experiments and their meaning for future theoretical and experimental work are discussed and evaluated.

The Power and Generality of Statistical Learning

Statistical learning is the most common term used to describe the ability to detect and use statistical structure (Alvarez, 2011). It seems appropriate to begin with an accounting of the features of the capacity because doing so seems to achieve two functions. First, it helps illustrate why and how much statistical learning matters as an object of study, justifying this thesis. Statistical learning is not just necessary; it is ubiquitous and adaptive across lifespans, domains, sensory modalities and levels of consciousness. Beyond that, though, an accounting of statistical learning's features – its apparent power and breadth of contribution to cognition – sets out the phenomena that the rest of this paper seeks to explore explanations for. Evidence outlining the power and generality of statistical learning are described here:

Statistical learning works from infancy

Some of the earliest work on statistical learning demonstrated its operation early in life. One experiment (Saffran, Aslin and Newport, 1996), for example, examined its role in the segmentation of words from fluent speech, an important marker of language acquisition. Eight-month-infants listened to a continuous, monotonic speech stream consisting of four three-syllable nonsense words repeated in random order. Every cue to word boundaries were removed, leaving

only the transitional probabilities between syllabus pairs. Despite this, infants indicated that they were familiar with the unique statistical structure of this stream by not listening as long to words derived from that structure as they would to novel non-words. Further research (Kirkham, Slemmer & Johnson, 2002) in fact found evidence of statistical learning from an even younger age – 2-month-old infants through an analogous research paradigm involving visual rather than auditory stimuli demonstrated discrimination between familiar and novel patterns of shapes.

Statistical learning occurs unconsciously, automatically and quickly

The infant studies just described do not merely illustrate that statistical learning is an ability available to humans from very early age; they show that statistical learning occurs automatically and apparently effortlessly (without exercising cognitive control), with only the presentation of a stimulus pattern necessary to trigger the process. Furthermore, this process of statistical learning proceeds rapidly; the words that babies became familiarized with in the Saffran et al (1996) study were presented within an only 2-minute-long stream. Studies involving adults demonstrate a similar pattern. For example, visual sequential statistical patterns shown to participants while completing an unrelated task still caused bidirectional associative shaping in brains indicative of rapid and incidental statistical learning even though more 87 percent of participants reported no awareness of any sequential pattern within stimuli (Turk-Browne, Scholl, Chun & Johnson, 2009). In a similar result, statistical learning of mean size was found undisrupted when cover tasks require either distributed or global attention to an artificial environment, though focusing attention on specific objects within an environment could disrupt learning (Triesman and Chong, 2008). In fact, participant awareness of statistical patterns within a stimulus set seems to relate little with either participant accuracy or confidence in their responses (Turk-Browne et al, 2009). All of this suggests that statistical learning is a highly

efficient process requiring no especial exertion of conscious effort or deliberate control of behavior, making its efficiency across situations compared to performance in other, more difficult cognitive tasks (such as those requiring attention) all the more curious.

Statistical learning enables apprehension of a broad variety of statistics

The variety of statistics that individuals are apparently able to extract from their environments is also noteworthy. As already noted, infants can extract transitional statistics – probabilities that a given unit will follow another specified unit in a sequence – from a stream of syllables. These can be thought of as examples of a broader class of statistics called conditional statistics, the predictive relationship between two events X and Y. A related sensitivity that humans exhibit is to cue-based statistics – relationships between perceptible attributes and attributes not directly perceptible, such as to emotion (Theissen, Kronstein, & Hufnagle, 2013). Individuals compute comparatively distributional statistics readily as well. For example, they evaluate the randomness of sequences in the same (somewhat biased) way across a variety of contexts, feature dimensions, sensory modalities, speed and manner of presentation (Yu, Gunn, Osherson & Zhao, Unpublished). Chong & Triesman (2004) found that judgment of the mean size of a set of circles are nearly as accurate as judgments of the size of single circles presented alone. Not only can humans judge the average emotion expressed by a crowd; they are also sensitive to the variance or heterogeneity of facial expressions within the same crowds (Haberman, Lee & Whitney, 2015). The list of statistics that humans seem to be able to extract through statistical learning is vast; it can only be partially delineated here. A key question in the literature considered later is whether a singular process we can call “statistical learning” underlies these achievements, or if wholly distinct processes drive the extraction of distinct statistics. Because this possibility casts doubt on the generality of any one potential explanation

of statistical learning, it demonstrates the challenge that a capacity as broadly powerful as statistical learning presents for theoretical work.

Statistical learning enables performance in a variety of tasks, domains, and modalities.

The variety of domains in which these described statistics can be extracted is similarly enormous. Individuals indeed can judge the mean size of circles efficiently as recounted earlier (Chong & Triesman, 2004), but they can also extract the average emotional expression, gender, identity, gaze direction, and ethnicity of members in a crowd (Haberman et al, 2015). In the same way, while individuals can from infancy extract transitional statistics from a sequence of heard verbal syllables (Saffran et al, 1996), sequence statistics can be extracted without much regard to the sensory modality or process in which said sequences are presented. Learners have displayed sensitivity to sequential conditional statistics in tactile, visual and tonal stimuli (Thiessen, Kronstein, & Hufnagle, 2013). And along with helping to segment continuous auditory input (Saffran et al, 1996), evidence for statistical learning has been found to facilitate visual search, contextual cuing, visuomotor learning, conditioning and generally any predictive behavior (Frost, Armstrong, Siegelman & Christiansen, 2014).

Factors Constraining Statistical Learning

Clearly, statistical learning is a powerful, broadly functional feature of human cognition. However, humans are of course not perfect learners, and the literature outlining constraints on statistical learning is around as substantial as the literature delineating its breadth and efficiency. Particularly speaking, stimulus presentation, sensory modality, task domain, extracted statistic, *and* experience have all been identified as profoundly impactful on the course and effectiveness of statistical learning. Specification of these constraints is useful for theoretical work because they demonstrate that statistical learning is not some universally operative catch-all learning process, but a concrete computational procedure (or set of computational procedures) implemented on a biophysically constrained neural substrate, thereby offering clues about how the process works. Key constraints on statistical learning well-substantiated within the literature are detailed here:

Statistical learning does not operate well across modalities and stimuli

Despite the apparent domain generality of statistical learning, evidence has emerged that the process is subject to modality and stimulus-specific constraints (Frost et al, 2015). In a review of evidence for transfer in artificial grammar learning (Redington & Chater, 1996), for example, only small and ambiguously meaningful magnitudes of artificial grammar learning effects upon modality transfer (e.g., visual to auditory) were observed. This separation of statistical learning between modalities may in fact have adaptive consequences: while transfer of learning between modalities may be limited, simultaneous learning of two sources of statistical structure can occur without mutual interference if implemented in distinct modalities (Conway & Christiansen, 2006). Other qualitative differences emerge across modalities in patterns of statistical learning, too. For example, compared to tactile and visual statistical learning, auditory

statistical learning is easier and better for the final part of input sequences than for earlier parts (Conway & Christiansen, 2005). Even within modalities, no statistical learning is transferred upon alteration of the stimuli by which a statistical structure is presented (Conway & Christiansen, 2006). Indeed, evidence suggests that as more time is taken to observe a structured sequence, knowledge of its statistical structure becomes more stimulus-specific rather than abstracted towards surface-independent representations (Johansson, 2009). These facts complicate the notion that statistical learning is a domain-general process. Though statistical learning works *within* many different modalities, in the strictest sense of the phrasing, it does not work much *across* those modalities. In fact, the same can similarly be stated at the stimulus rather than modality level of analysis. This finding recommends that a more precise conception of statistical learning must emerge beyond that of a unitary system indifferent to how statistics are conveyed to the learner.

Statistical learning across tasks/statistics/domains

In the same way, evidence has emerged that statistical learning operates differently depending on the statistics extracted to fulfill the demands of a task. For example, something about the process by which humans judge the randomness of sequences and other stimuli demonstrates a consistent bias in that form of statistical learning which is not evident in the way humans extract other statistics. In the case of randomness perception, humans generally exhibit negative recency, "expectations that a streak of events will end" (Oskarsson et al, 2009). They are more likely to judge sequences with short event streaks than truly random sequences as random. However, in certain domains, this pattern reverses: in domains involving wins and losses, especially in gambling scenarios, humans instead exhibit positive recency, expecting streaks of events to continue. The fact that evidence for similar patterns of bias have not emerged

in study of other sorts of statistical learning suggests that the computation of different statistics might involve different processes, complicating the notion that some unitary system does the work, or even that "statistical learning" refers to a single topic at all.

Other examples of this issue have been articulated in logical/conceptual terms. For example, it's asserted that learning about elements that have presented during exposure is an achievement conceptually distinct from that of learning rules that can be applied to novel elements and novel combinations (Aslin & Newport, 2016; Thiessen et al, 2013). Along a similar vein, a distinction has been repeatedly drawn between the extraction of discrete representations such as that of particular conditional statistics or words, and the extraction of distributional statistics that characterize the prototypical characteristics of an element set. Rooted in these distinctions is not a set of experimental results so much as a considered analysis of differences between the tasks learners perform well on. Generalizing learned statistics to apply to future stimuli or experiences seems to be a step that can *only* logically occur after and distinctly from actual learning of those statistics. Similarly, computation of the broad distributional statistics of a pattern seems to *require* integrating previously extracted low-level discrete statistics. Few theories of statistical learning in the literature manage these considerations through reference to only a single mechanism.

Statistical learning across experience

Along with previously described factors, the course of statistical learning also seems to be profoundly sensitive to past experience. Nine- and 10-month-old infants exposed to sequences of either disyllabic or trisyllabic words, for example, only successfully segment words within a subsequently presented sequence if the sequence is equally disyllabic or trisyllabic, suggesting that prior experiences of statistical structure equip infants with expectations about future

statistical structure (Lew-Williams & Saffran, 2012). Another experiment similarly presented adults with two sets of patterns – either first an unordered 'pattern' and then an ordered pattern, or first an ordered pattern and then an unordered pattern (Junge, Scholl, & Chun, 2007). Only participants in the order-first condition displayed learning of the ordered pattern later on, similarly suggesting that prior experiences of statistical structure impact expectations about future experiences about statistical structure. Results further suggested, too, that a revere effect does not exist, at least not as strongly: experiencing a novel statistical structure was not found to substantially impact learning of the statistics extracted immediately before the new experience. Interestingly, though, another study (Gebhart, Aslin, & Newport, 2009) replicated these findings but discovered that explicitly indicating the presence of two statistical structures (i.e. with some divider) or extending experience to the second statistical structure can eliminate these effects. The impact of experience on statistical learning may therefore be dependent on the extent to which prior experience can be judged to be relevant for ongoing performance.

Theoretical Issues - Do Different Processes Underlie Statistical Learning?

This nuanced portrait of statistical learning's power and constraints raises a few theoretical issues considered extensively within the literature; they are so integral to the subject that assumptions about them can seem to impact the course of much research on statistical learning, including the experimental work presented within this thesis. Reviewing them sets out explicitly the theoretical orientation of this thesis while also deepening its account of the state of research within the field.

The issue of whether statistical learning is one process turns on two issues. *First* (1), it is unclear if a single process performs statistical learning between stimuli or sensory modalities. It could be, for example, that statistical learning is unitary in the sense that a single module within the brain operates on experienced statistical patterns no matter which sensory modality or stimulus that the patterns are extracted from, or perhaps instead that distinct but similar learning processes operate in areas of the brain individually devoted to processing information from specific sense-modalities or concerning particular stimuli. *Second* (2), and separately, it is unclear if a single process can compute the broad variety of statistics that characterize statistical learning. It might be the case, for example, that some statistics are computed (whether this process is implemented individually in different parts of the brain depending on sense modality or stimuli or not) through one process while other statistics are computed in some other way.

Is statistical learning unitary across modalities?

The considerable similarities in our statistical learning capacities across modalities inclines the presumption that a single, stimulus-independent process carries all the learning out, but the limited transfer of learning between the modalities constitutes a serious challenge to that position (Frost et al, 2015). However, one line of argument might reframe that issue: failure to

transfer learning between sensory modalities might be a consequence of the same processes that govern how we determine if extracted statistics are likely to apply to new experiences. Since cross-sensory stimuli are so dissimilar, limited transfer of learning between sound and touch might be just an extreme example of the principles that prevent learning transfer between two same-modality structures separated only by a 30s pause (as demonstrated by Gebhart et al, 2009). Participants might not demonstrate transfer between modalities because no evidence is offered to indicate such a transfer is reasonable. Other challenges to the unitary model of statistical learning are more difficult to assuage. For example, some evidence (e.g. Frost et al, 2015) suggest that there is no significant correlation between individuals' performance on statistical learning tasks between sensory modalities as might be predicted if statistical learning were a unitary process; in other words, performance on a statistical learning task in one modality can't be predicted from performance on another that's similarly designed, suggesting that the same unitary process is not engaged after all.

Cognitive neuroscientific evidence, though, complicates this idea again. Participants observing statistically structured (rather than unstructured) sequences of shapes while doing something unrelated to the structure exhibited robust neural responses in the medial temporal lobe (Turk-Browne et al, 2009). In a stronger example, a patient named LSJ with "complete bilateral hippocampal loss and broader MTL damage" could exhibit no statistical learning under a procedure under which control group participants reliably could (Shapiro, Gregory, Landau, McCloskey and Turk-Browne, 2014). Since this damage was to a relatively focal area of the brain, this finding substantiates the argument that statistical learning "depends at least partially on a shared system (the MTL)". Still, input and output to the MTL at least are processed in

modality-specific areas, potentially explaining some modality-specificity within statistical learning, though perhaps unconvincingly.

Altogether, though, these results ambiguously suggest that statistical learning is best understood not as a single, unitary process but rather multiple, modality-constrained processes that share some computational 'machinery'. Does this mean that scientists should cease speaking of "statistical learning" as a unitary concept and instead talk only of "visual statistical learning" or "auditory statistical learning" and so forth? To a key extent, probably not: along with considerable evidence that statistical learning is not some modality-independent process, there is also a wealth of evidence that statistical learning operates similarly across modalities, with exceptions that seem to depend more on incidental aspects of the sense modality considered rather than fundamental difference in learning mechanism. Whatever underlies statistical learning seems to be rooted in a cognitive learning principle realized in different parts of the brain.

Is statistical learning unitary across computed statistics?

As already pointed out, the possibility that statistical learning is unitary across computed statistics is challenged by 1) the massive diversity of statistics that individuals seem capable of computing rapidly, and 2) biases evident in the computation of some statistics by individuals that are not apparent in that of others. Some authors (ex. Aslin & Newport, 2016; Thiesson et al, 2013) argue that learning about elements that are presented during exposure is an achievement distinct from that of learning rules that can be applied to novel elements and novel combinations. However, it is possible that additional rules do not have to be represented for learned structure to impact later experience. For example, extant grouping principles like temporal proximity, perceptual similarity and shared context might cause the individual to observe stimuli as

exemplary of past experience. Thiesson et al (2013), however, assert that the extraction of consistent clusters in the input cannot by itself enable the ability to generalize from prior experience. Furthermore, they assert that additional mechanisms are necessary to account for our sensitivity to distributional statistics like average size and variance. They thus account for statistical learning in terms of two major processes. First, individuals are thought to *extract* clusters of the input with processes and represent them discretely. These serve as exemplars in long-term memory. Second, individuals are thought to *integrate* across these clusters to represent the central tendency of exemplars stored in memory and benefit from their similarity and distribution, which then guides subsequent extraction. This distinction is rooted in a much deeper computational analysis of the tasks involved in statistical learning than that achieved by Aslin & Newport (2016). Because it casts statistical learning as dependent on two mutually dependent processes, too, it doesn't quite defeat the idea that statistical learning operates as a single system. Instead, different statistics are computed at different points along the extraction-and-integration process.

At the very least, this account fits with the most popular account of randomness perception, and the over-alternation bias more specifically. The most widespread theoretical account of these phenomena is the idea that judgments are made using a representativeness heuristic, such that people expect local regions of a random sequence to be representative of their prototype or schema for random processes (Kahneman & Tversky, 1972). For example, a streak of four heads in a row in coin tossing is judged unlikely because it violates the characteristic of random sequences that both outcomes occur equally often, leading to elevated expectations of tails for the next flip. Fitting with the extraction-and-integration framework, this account centers the selection of and integration over exemplary chunks for computation of the randomness

statistic. An over-alternation bias emerges because chunks are overly small or because integration over extracted chunks occurs prematurely. As a lot of evidence has been marshalled to support this account of randomness perception – for example, expected randomness judgments based on chunk size have been shown to closely track the judgment accuracy of participants with analogous short term memory capacities (Kareev, 1992) – the body of literature surrounding randomness perception substantiates the extraction-and-integration framework, though more evidence seems necessary before the framework can be confidently embraced.

Why does experience impact learning the way it does?

An additional key theoretical issue in research on statistical learning is in explaining the impact of prior experience on statistical learning. The literature generally agrees on a relatively simple mechanism for this impact. First, individuals experience and extract from some set of stimuli a statistical structure. Then, because these experiences have set hypotheses or expectations for future experience, they treat further experiences/stimuli as having the same (or at least similar) statistical structure, and only stop if extended experience makes evident that this treatment is inappropriate (Lew-Williams & Saffran, 2012; Gebhart et al, 2009; Junge et al, 2007). This account doesn't just explain why participants anchor on the statistical structure of early experiences even as new experience deviates from it; it also explains phenomena like the capacity of pauses and other dividers between ordered sequences to prevent such anchoring.

However, it is not obvious why it should take longer for learners to extract statistics after having extracted other statistics than after having extracted no statistics at all; the existence of these expectations alone cannot explain the impact that past experience of structure has on future learning of different structure. Gebhart et al (2009) argue that our cognitive architectures make an organizational trade-off between the inefficiency involved in waiting for an input to be

extracted and that in having to recover from errors driven by biasing learning in favor of early input. Here, the primacy effect is a deliberate heuristic that enables reaction to ordered stimuli faster than the time it takes for an entire body of stimuli to be extracted. The fact that this trade-off is moot when explicit cues for distinction between statistical structures are evident may limit the harmful effects of the bias to fewer contexts.

The impact of experience on statistical learning may therefore be dependent on the extent to which prior experience can be judged to be relevant for ongoing performance. Since this kind of judgment must not be based exclusively on the statistics participants may be choosing to disregard, these findings suggest that in practice statistical learning involves decision-making (at least of the implicit kind) that recruits knowledge gained in different circumstances and perhaps also through different processes. The nature of these knowledge interactions has not been studied much in detail, but may prove fundamental to comprehending the cognitive basis of statistical learning.

How Does Statistical Learning Happen?

At the core of all these theoretical issues so far discussed is ambiguity concerning the actual mechanism underlying statistical learning, especially that involved in extracting specific instances/clusters rather than integrating to compute distributional statistics. Though the project of thoroughly characterizing the nature of this kind of statistical learning is too ambitious for this thesis, we focus on a debate within the literature on a kind of statistical learning (word-referent pair mapping) comparing two different ideas about how the process is enacted. Though highly specific, the experimental paradigm surrounding this kind of research offers advantages of control and measurability that research surrounding more continuous stimulus sets such as transitional statistical learning do not. Since it's likely that all extractive statistical learning follow the same general cognitive principles, we can generalize from the results of this body of research to make inferences about statistical learning more broadly.

Literature review

Much of the cognitive neuroscientific work concerning statistical learning has been built upon the implicit assumption that statistical learning involves integration over all or most past learning instances (for example, Schapiro et al, 2017). However, precisely this point has been contested in recent behavioral examinations of statistical learning. One pattern of studies examines a manifestation of statistical learning where nonsense words are paired with one of several object categories such as ball, bear, hand, or shoe. For each trial, the experiments display several image referents representing object categories accompanied by a labelling utterance, perhaps of the form "Oh look, a ___", ending with a nonsense word that has been paired with one of the object categories represented. In these experiments, since the object category paired with its displayed nonsense word is always present in a trial while the presence of other object

categories is randomly varied, participants gradually learn that “zud”, for example, refers to “bear”.

One result (Yu & Smith, 2007) found that when individuals are tested afterward on their knowledge of these word-referent pairs, they chose the correct item more often than would happen by chance; its researchers concluded from this that individuals track co-occurrence frequencies across training trials. It’s complained repeatedly in the literature, though, that studies like this one cannot draw such strong conclusions about the mechanism of statistical learning since they don’t measure the course of learning throughout an experiment. In other studies (Medina et al, 2011; Trueswell et al, 2013), participants are immediately tested on their knowledge of statistical relationships after every stimulus presentation by being prompted to select the image referent they believe matches the nonsense word presented. A critical finding about the course of statistical learning during these experiments contradicts the claim that the likelihoods of multiple hypotheses about possible word-referent pairs are tracked over the course of multiple learning experiences. These experiments found that performance in this task improved as participants were presented with more stimuli. More importantly, though, when participants selected a wrong image referent in a trial to match a given nonsense word (and learned as much when the image referent did not occur as an option on a succeeding presentation of the nonsense word), they failed to apply memory of other image referents in the failed trial to perform better than chance on the succeeding trial. In contrast, when participants selected the correct image referent in a trial to match a given nonsense word, they performed much better than chance on succeeding trials when presented again with the same nonsense trial. In other words, no evidence was found of memory for the likelihood of alternative hypotheses. As an alternative mechanism for learning statistical patterns that accounts for these findings, the

propose-yet-verify model of statistical learning has been proposed. Instead of asserting that individuals “accrue a best final hypothesis by comparing multiple episodic memories of prior contexts or multiple semantic hypotheses”, the propose-yet-verify model asserts that “learners hypothesize a single meaning based on their first encounter with a word and on later encounters attempt to retrieve this hypothesis from memory and test it against a new context, updating it only if it is disconfirmed” (Medina et al, 2011).

Confoundingly, however, despite this support for the propose-yet-verify model, a comparison of computational models that simulate the two competing models just discussed found that the multiple hypotheses model more closely tracks human performance on statistical learning tasks than its competitor (Kachergis et al, 2012). Importantly, though, despite posing as a replication of experiments supporting the ‘propose-yet-verify’ model of statistical learning, the testing regime used to measure statistical learning is notably different from them: instead of interleaving test trials with stimulus presentation within blocks, the experiment tracks learning over time by alternating between blocks of stimuli and tests of learning.

A growing neuroscientific literature on statistical learning also offers conflicting evidence for these competing models of statistical learning. Neural network models suggest rapid statistical learning is most efficiently achieved by overlapping representations of experiences within the same neural substrate so that generalities between stimuli can be found (McClelland et al, 1995). This mechanism would seem to implement an abstraction of the multiple-hypotheses rather than a propose-yet-verify process for statistical learning, as regularities are found not by entertaining hypotheses and testing them but by tracking similarities across a set of experiences. Perhaps problematically for this account of statistical learning, a lot of work suggest that the hippocampus has a central role in the operation of statistical learning. For example, a patient

with “complete bilateral hippocampal loss and broader MTL damage” could exhibit no statistical learning under a procedure under which control group participants reliably could (Shapiro et al, 2014). Participants observing statistically structured (rather than unstructured) sequences of shapes while doing something unrelated to the structure exhibited robust neural responses in the medial temporal lobe (Turk-Browne et al, 2009). Though the hippocampus notably supports rapid learning, this is problematic for the described mechanism for statistical learning because the structure seems specifically specialized in avoiding rather than pursuing overlap of representations so that interference between memories does not happen. If the hippocampus is a substrate of statistical learning, it could be instead that the structure doesn’t integrate across overlapping representations of experiences, but instead stores fewer, more particular memories of hypothesis proposition and rejection to rapid regularity extraction. Seeming to resolve this tension, very recent work in neural network modeling of the hippocampus suggests that there may exist complementary learning systems in different parts of the structure that variously specialize in storing discrete representations and in generalizing between overlapping experiences (Schapiro et al, 2017). However, confirmation of this hypothesis would only leave ambiguous how statistical learning happens until the process is localized within one of the complementary systems. Even still, it could be that statistical learning engages both complementary systems or engages one over the other depending on the circumstances under which learning takes place.

Rationale

This kind of statistical learning is useful for study because stimuli representations and their underlying statistical structure of their ordering are discrete, straightforward both to manipulate for experimental purposes and to measure learning thereof compared to designs

emphasizing transitional statistics. For example, while word-referent pair matching tasks can convey 4 potential pairings of a word with a meaning in a single experimental trial and statistical structure over relatively few trials, experiments representing statistical structure with transitional probability-constrained sequences must convey 4 potential pairings between stimuli and what they transition to using at least 4 distinct trials; statistical structure necessarily involves proportionately more trials. For reasons like these, some hypotheses about statistical learning are easier to evaluate from the word-referent pair matching paradigm; after establishing principles within the paradigm it may then be useful to test their generality across other domains of statistical learning.

A defining difference between the multiple hypotheses and propose-yet-verify models is how learners are thought to track the likelihood of statistical pairings alternative to the ones they consider the most plausible. Per the multiple hypotheses model, humans engaged in statistical learning track over an experiment changes in the likelihood of multiple possible statistical pairings (i.e., hypotheses). According to the propose-yet-verify model, learners "hypothesize a single meaning based on their first encounter with a word" and "neither weight nor even store back-up alternative meanings/hypotheses" (Kachergis, Yu, & Shiffrin, 2012). To an important extent, this difference perhaps ought to be applied to fundamental characterize of the propose-yet-verify and multiple-hypotheses models, rather than any emphasis on specific hypothesis proposals or even probability tracking. The reason such an abstraction is important is because research in other parts of the literature, including the neuroscientific evidence just discussed and the extraction-and-integration framework considered earlier to account for all statistical learning either reject or disregard the possibility that the result of statistical learning is represented in these ways except perhaps to an analogical extent. For example, neural network models suggest

rapid statistical learning is most efficiently achieved by overlapping representations of experiences within the same neural substrate so that generalities between stimuli can be found (McClelland et al, 1995). Cognitive models of statistical learning such as the extraction-and-integration framework (Thiesson et al, 2013) on the other hand reject boundary-finding models of statistical learning (as most neural networking models of statistical learning are) in favor of clustering models, which aver that learners store discrete representations (e.g., words) from continuous input but not observed probabilities; instead, they assert that representations compete in the process of extraction. Both these kinds of models suggest to varying extents that experiences substantiating alternative hypotheses would be remembered and applied for learning the main 'hypothesis' but don't include the multiple hypotheses model's emphasis on hypotheses or probability tracking. To enable integration of results across literatures, from here out we now only distinguish between the propose-yet-verify and multiple hypotheses models in terms of their differing hypotheses about how alternative possible statistical extractions are remembered during and shortly after learning – either not at all in the former case, or to a significant extent in the latter. Similarly, references to alternative hypotheses, hypothesis tracking and so forth should be considered shorthand for underlying interactions between main hypothesis and alternative hypothesis *substantiating experiences* within the MTL and the rest of the brain, if not avoided altogether. Despite these abstractions, an emphasis on the competing hypotheses framework for understanding statistical learning is useful because it allows evaluation of a body of literature (that on word-referent pair matching) key to clarifying the mechanism of statistical learning. Conclusive evidence supporting the multiple-hypotheses or propose-yet-verify models of statistical learning would have broad-reaching implications about the nature of statistical learning. While the multiple-hypotheses model might conform to some extent with well-

articulated cognitive and neuroscientific theories of statistical learning, the propose-yet-verify model of statistical learning seems to roundly contradict both their findings and the assumptions underlying their research; its success is definitely of interest for the course of statistical learning research.

The Experiment

The conflicts between studies comparing potential mechanisms for statistical learning deserves accounting for. After all, the assumption that statistical learning integrates information over many trials colors the interpretation and design of much research in the literature at both the neural and cognitive levels. For this reason, the observation that the findings of studies about mechanism are closely tracked by the way learning is measured must be considered closely. A defining difference between the multiple hypotheses and propose-yet-verify models is how learners are thought to track the likelihood of statistical pairings alternative to the ones they consider the most plausible. Per the multiple hypotheses model, humans engaged in statistical learning track over an experiment changes in the likelihood of multiple possible statistical pairings (i.e., hypotheses). According to the propose-yet-verify model, learners "hypothesize a single meaning based on their first encounter with a word" and "neither weight nor even store back-up alternative meanings/hypotheses" (Kachergis, Yu, & Shiffrin, 2012). One way to determine *after* a set of learning experiences if multiple hypotheses about statistical pairings have been tracked or not is to modify midstream the set of statistical pairings and observe how quickly and effectively individuals adapt their behavior to the changes.

For example, suppose that in the first half of a word-referent mapping task, the nonsense word "brazz" is presented with the A object category 100% of the time, the B object category 50% of the time, and two other object categories picked randomly from a pool of four (C, D, E,

F). According to the multiple hypotheses model of statistical learning, learners monitor the likelihood of the brazz-A, brazz-B and similar possible pairings; according to the propose-yet-verify model, after learning pairings, learners tend to simply hold the brazz-A hypothesis in mind and check it on each succeeding trial. If, midstream, pairings change such that brazz is paired with B 100% time and A is as likely as C, D, E, or F, the two models predict different courses of adjustment, which can be measured by simultaneously beginning a testing regime periodically interleaved with presenting the new pattern of stimuli. The propose-yet-verify model predicts that learners will initially perform at chance as they adapt to the new sequence, as they haven't been tracking co-occurrences that make brazz-B more plausible than other possible pairings. The multiple hypotheses model, however, predicts no such fall to chance-level performance on the task, instead predicting gradual alternation from a preference for brazz-A to brazz-B without a phase of chance performance.

The possibility of this kind of research suggests a design centered on measuring memory of alternative hypotheses in different ways and at different points throughout an experiment. To test the hypothesis the multiple hypotheses model works more like human statistical learning than the propose-yet-verify model, participants could be presented with a standard word-referent pair matching task, where they are presented with a series of learning episodes (as already summarized earlier) and then tested on their learning of a word-referent pair. In both conditions during the first section of blocks in the experiment, words occur with a 'main' referent 100% of the time and also with a "secondary" referent more often that it is with all other possible referents, but not as often as the word is paired with its main referent. During the final section of the experiment, half of the words are selected, and their main referents changed (i.e. a new referent begins to occur with the word 100% of the time), with other potential referents reduced

to appearing randomly like other non-main and non-secondary referents. In the control condition, a random referent is selected for each modified word-pairing, with the previous main and secondary referents becoming randomly appearing. In the experimental condition, the secondary referent is always selected for each modified word-pairing (appearing 100% of the time with the word instead of a fraction of the time), with only the previous main referent becoming randomly appearing.

Possible Outcomes

This design enables testing of propose-yet-verify and multiple hypotheses and related models in at least two ways. First, the course of learning throughout the experiment can be analyzed. For example, if performance upon successive appearance of some word does not improve after selecting the wrong referent in a previous experience where that wrong referent is now absent, participants are not using their memories of alternative hypotheses (including the correct hypothesis that co-occurred with their incorrect selection) to improve their learning. If the learning curve is smoother and incorporates information from failed trials, then the multiple hypotheses model is supported. Second, and a more novel contribution to the literature, learning at the end of the experiment can be analyzed by comparing performance in the second section of the experiment between the two conditions. If participants have maintained memory of alternative hypotheses through the end of the first section of the experiment, they will perform significantly better in the condition on modified terms where secondary referents that have always had a weak matching with their word become main referents than in the condition where random referents that previous had no pairing with their word become main referents. If participants have not maintained memory of alternative hypotheses through the end of the first section of the experiment, their performance on modified terms in the second block of the

experiment will be indistinguishable. It's highly unlikely that performance would be better in the control condition than the experimental condition, so that possibility isn't considered.

An interesting possibility is that results analyzed over the course of learning differ from results obtained during the final block when learning is considered finished and the experimental method is implemented. For example, it could be found that performance improves after a learning trial whether the correct referent was selected or not, but that performance in the experimental condition is indistinguishable from that in the control condition. This would support the multiple hypotheses model of statistical learning but at the same time suggest that at some point after learning, memory of alternative hypotheses is discarded, supporting a pruning model of statistical learning. Alternatively, it could be found that performance improvement after a learning trial depends on whether the correct referent was selected or not, but that performance in the experimental condition is substantially better than that in the control condition. This would substantiate a propose-yet-verify model of statistical learning, but at the same time suggest that memories of previously entertained hypotheses are still maintained somehow throughout learning, even though they don't necessarily enhance performance, supporting what might be characterized as a wasteful tracking model of statistical learning. If this outcome turned out to be true, it seems likely that distinct learning mechanisms operate in parallel to guide statistical learning, with one perhaps enabling flexibility in performance when statistical structure changes, and the other guiding initial learning. Still, these additional two possibilities are unlikely outcomes of the experiment.

Method

Participants

65 participants (36 women; median age 25) were recruited on Amazon Mechanical Turk to participate in the experiment. 30 were assigned to the experimental condition while 35 were assigned to the control condition. Participants were compensated $3.50 for their participation, and spent a median of 18 minutes on the task. Additional participants were recruited for the experiment, but to ensure data quality, outliers with respect to time spent performing the experiment were removed from the data set.

Apparatus

The experiment was programmed in JavaScript using jsPsych, a JavaScript library for creating behavioral experiments that run in a web browser (de Leeuw, 2015). Individuals participated in the experiment remotely, so the device and display specifications of the study depended on their whims and may have varied considerably. Participants were asked to sit in a quiet, well-lit area where they could focus on the procedure.

Statistical Structure

Principles of organization. Here we specify the statistical structure of the word-referent pairings and their presentation before going into the details of how they were presented. Trials were generated such that 1) each word had a unique main referent and half of words also had a unique secondary referent; 2) main referents were paired with their respective main referent 100% of the time and if applicable paired with their secondary referents 60% of the time; 3) other referents were paired with words randomly, being sampled randomly from a pool without replacement. The experiment consisted of 16 pseudowords and 24 possible image referents.

Words were presented with 4 possible referents, and in each of 4 blocks, words were presented 5 different times.

Algorithm. For each word an arbitrary main referent was selected; half of words were also assigned a secondary referent. Also for each word, a shuffled pool of potential referents was generated to create trials, with the word's main and secondary referent removed, when applicable. For each trial, either 2 or 3 items (2 if a secondary referent is to be present, 3 otherwise) were randomly pulled from the referent pool to be included with the word's main and potentially secondary referent in the trial. For words with secondary referents, the secondary referent was included in 60% of trials but excluded in the other 40% of trials. Referents that were neither a word's secondary nor main were never paired with that word more than once within a block.

Procedure

Instructions. Participants received the following instructions: "You will be presented with a series of trials on which you will see 6 pictures of objects on the screen and a single \"mystery\" word that refers to one of the objects presented. There will be 80 trials in each of 4 blocks, and over the course of the experiment you will be shown 16 different mystery words, each with a different meaning. Your job is to figure out what each mystery word means. Click on the object that you think the word may refer to. Make your best guess, and try to keep improving your guesses as the experiment continues." These instructions were repeated to them at the end of each block along with an encouragement to take a break if necessary.

Presentation. Each trial was presented as a string of text "Oh look! A <pseudoword>!" paired with 4 referent images (**Figure 1)**. These pseudo-words were computer generated to broadly sample phonotactically probable English forms while referent images were selected from

the Cross-Situational Word Learning Stimulus Pool (Kachergis, Yu, & Shiffrin, 2014). Each trial was shown on the screen for at most 10000ms; if participants selected an image, a bold border would appear around the referent to indicate their selection. Within each of 4 blocks, 80 trials were presented this way; each were organized so that words occurred randomly except with the constraint that the second instance of a word never presented until all the first instances of all the other words were presented, and so on.

Analysis

Additional participants were recruited for the experiment, but to ensure data quality, outliers with respect to time spent performing the experiment (i.e., who on average used less than 15% of the time allotted to respond to stimuli) were removed from the data set. However, 65 participants were nonetheless included in the analyzed data set, including 30 in the experimental condition and 35 in the control condition.

All analyses were run using R. A one-way between-subject ANOVA was first conducted with condition as the predictor variable to determine if accuracy on the modified 4th block word-

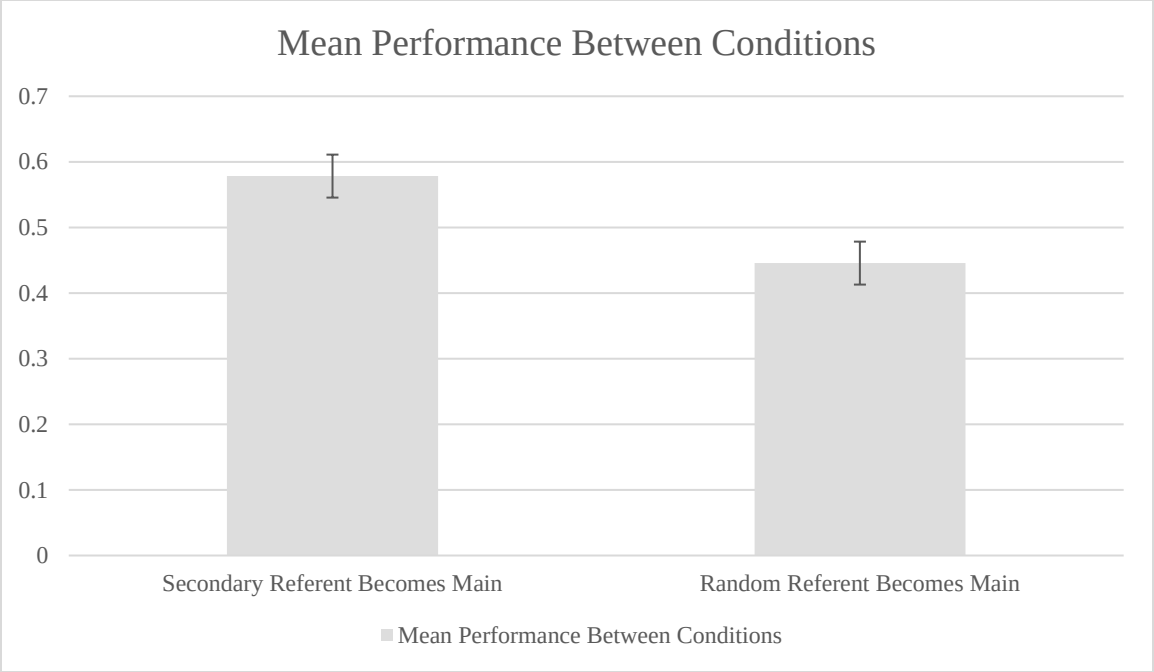
pairing trials was significantly different between conditions. A follow-up student's t-test was also performed to compare performance between the two conditions.

Next, a repeated measures ANOVA was performed to confirm that performance improved (i.e. accuracy selecting the main referent for every word) between in the first 3 blocks of the experiment. Additionally, the Pearson's product-moment correlation between block and accuracy was tested.

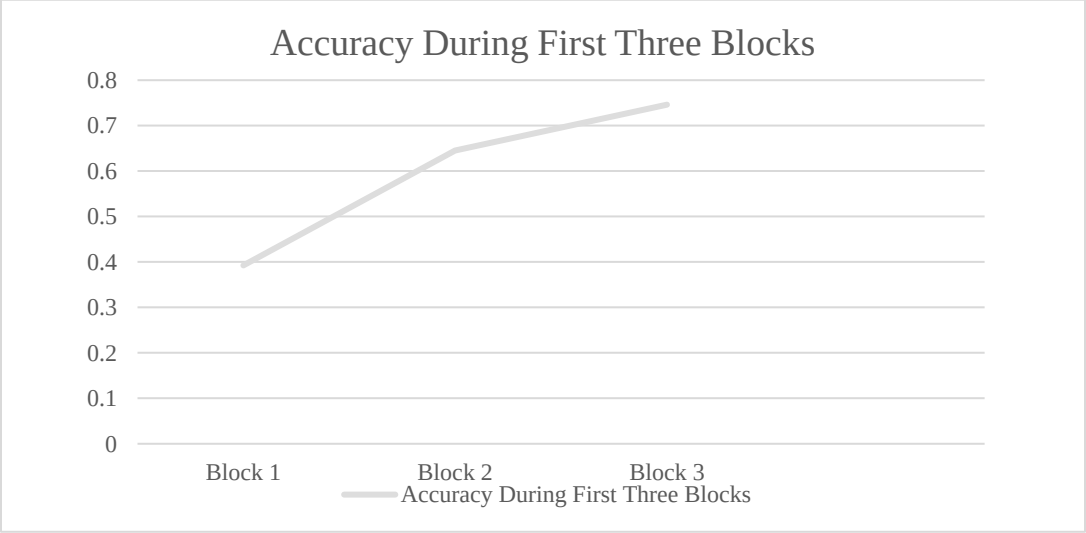
Finally, accuracy of word-referent pair matching within the first three blocks after a trial correctly matching the same word with its main referent as well as accuracy of word-referent matching after a trial incorrectly matching the same word with its main referent were compared. A one-way ANOVA examined the effect of the accuracy on a given trial for a given word on the accuracy on the succeeding including of that word in the first 3 blocks. Performance succeeding an incorrect response and performance succeeding an correct response were also subjected to one-sample t-tests to confirm whether performance differed significantly from chance-level accuracy.

Results

As shown in **Figure 2**, a one-way between-subject ANOVA conducted with condition as the predictor variable revealed a significant difference between group means for accuracy on the modified 4^{th}-block word-pairing trials at the $p<.05$ level for the two conditions [$F(2) = 46.24$, $p=1.293e-11$]. A follow-up student's t-test also performed to compare performance between the two conditions found a significant difference in the scores for the experimental ($M=.5783$, $SD = .494$) and control ($M=.4457$, $SD = .4972$) conditions; $t(2542.4) = 6.8034$, $p = 1.27e-11$.



Next, a repeated measures ANOVA was performed to confirm that performance improved between in the first 3 blocks of the experiment, revealing a significant effect of time on accuracy, $F(1) = 1433$, $p < 2.2e\text{-}16$. Additionally, a Pearson's product-moment correlation between block and accuracy was tested, revealing a significant relationship ($cor = .2901$; $t(15598) = 37.855$, $p < 2.2e\text{-}16$), shown in **Figure 3**.



Finally, accuracy of word-referent pair matching within the first three blocks after a trial correctly matching the same word with its main referent as well as accuracy of word-referent matching after a trial incorrectly matching the same word with its main referent were compared. A one-way ANOVA examined the effect of the accuracy on a given trial for a given word on the accuracy on the succeeding including of that word in the first 3 blocks and revealed a significant difference between performance after a correct and after an incorrect response [$F(1) = 2452$, $p=2.2e-16$]. Performance succeeding an incorrect response and performance succeeding an correct response were also subjected to one-sample t-tests to confirm whether performance differed significantly from chance-level accuracy, revealing significant differences between performance and chance level in both cases. For responses after an incorrect response, mean performance was .3770; $t(6306) = 29.009$, $p < 2.2e-16$. For responses after a correct response, mean performance was .8143; $t(4350) = 104.19$, $p < 2.2e-16$. Even during only the first block, mean performance after an incorrect response was 37.70%, significantly above chance ($p < 2.2e-16$). In all, the data confirms the reality of humans' capacity to quickly extract word-referent pair mappings and offers unequivocal evidence supporting the multiple hypotheses model of statistical learning over the propose-yet-verify model.

Accuracy After a Correct Response	Accuracy after a Wrong Response
.8143	**.3770**

Discussion

In consonance with previous results on statistical learning for word-referent mapping, participants managed to quickly extract word-referent mappings after a relatively small number of experiences; indeed, by the third block, average performance on the task was 74%, several times better than chance. As hypothesized, participants demonstrated both during the course of learning and after learning not just memory of alternative hypotheses but also readiness to apply these memories to improve performance. When participants selected a wrong pairing for a while on a trial, they demonstrated memory of alternative hypotheses for the word-pairing by selecting the main referent of the word upon its successive appearance at a rate substantially above chance. Later, when some words' main referents were swapped with their secondary referents, participants again demonstrated memory of the presence of these secondary referents in previous blocks by adapting to the change at a rate substantially better than control participants did when some words' main referents were swapped with a random referent instead. These results confirm that participants had attended to alternative hypotheses throughout the experiment.

These results strongly contradict the results of a similar experiment (Trueswell et al, 2013); the paper did not include a secondary referent manipulation, but it did analyze performance succeeding trials when a participant selected an incorrect word-referent; it found no significant difference in accuracy from chance in these successive trials, substantiating the propose-yet-verify model of statistical learning. It's not clear why results differ so much between the studies; it could be that including 4 rather than 5 possible referents for each word in each trial made the task of tracking alternative hypotheses easier. It may also be the case that the studies' differing subject pools (Mechanical Turk workers compensated $3.50 versus undergraduate students compensated $10) have something to do with the differences in outcomes. Still, this

study's marshalling of several, rather than only one, strategy for measuring memory of alternative hypotheses helps corroborate results despite these differences.

In many ways, the debate between the propose-yet-verify and multiple hypotheses models obscure the wide variety of possible accounts of statistical learning that might turn out to be accurate. The cognitive schemata proposed by the framework to explain performance in many cases are difficult to generalize when marshalled to understand other kinds of statistical learning. For example, the propose-yet-verify model seems to struggle to explain statistical learning where pairings are ambiguous (for example, a transitional conditional probability above .5 but below 1) since hypotheses are supposedly abandoned upon disconfirming experience. Similarly, the multiple hypotheses model's characterization of learning in terms of concurrent statistic tracking only ambiguously coheres with findings in cognitive neuroscience and psychology, and leaves unclear issues like how minds select which potential hypotheses are tracked and for how long the tracking occurs. This vagueness leaves open the possibility that pruning theories or even wasteful memory models of statistical learning as discussed in the possible outcomes section of this thesis might still readily compete with the multiple hypotheses model even if memory of alternative hypotheses is confirmed. Given all this, it might be argued that the word-referent pair matching paradigm has little to offer researchers interested in characterizing the nature of statistical learning more broadly. Still, shared cognitive principles underlying different cases of statistical learning such as the tracking of multiple possible pairings can be more clearly elucidated under this paradigm than others, rounding out gaps in theoretical work that might otherwise be ignored.

Future research evaluating the propose-yet-verify and/or multiple hypotheses models as broad models of statistical learning might attempt to generalize the paradigm and its findings to other domains of learning. For example, the construct of secondary pairings might be used in

transitional statistical learning research to explore how possible representations compete as participants learn artificial grammar rules. Alternatively, other work might attempt to test general ideas about statistical learning in the context of word-referent pair matching; for example, midstream statistical structure manipulations in this thesis are in some ways analogous to those discussed elsewhere (Gebhart, Aslin, Newport, 2009) with reports of strong interference of prior experience on new learning; these effects weren't observed as decisively in the experiments reported here, suggesting that differences in our designs might mediate the primacy effect.

Finally, the broad consilience of these findings with more general accounts of statistical learning substantiate the hypothesis that statistical learning is a general process that can account for learning in a variety of domains. A more concerted effort that the arguments discussed here (which gloss over differences to facilitate theoretical crossover) to account for word-referent learning phenomena in terms of more general theoretical learning principles might substantially advance psycholinguistic research.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. Trends in Cognitive Sciences, 15(3), 122-131. doi:10.1016/j.tics.2011.01.003

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. Cognition, 106(3), 1382-1407. doi:10.1016/j.cognition.2007.07.005

Conway, C. M., & Christiansen, M. H. (2005). Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(1), 24-39. doi:10.1037/0278-7393.31.1.24

Conway, C. M., & Christiansen, M. H. (2006). Statistical Learning Within and Between Modalities: Pitting Abstract Against Stimulus-Specific Representations. Psychological Science, 17(10), 905-912. doi:10.1111/j.1467-9280.2006.01801.x

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? Progress in Brain Research Essence of Memory, 323-338. doi:10.1016/s0079-6123(07)00020-9

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. Trends in Cognitive Sciences, 19(3), 117-125. doi:10.1016/j.tics.2014.12.010

Gilbert, S. F. (2012). Ecological developmental biology: Environmental signals for normal animal development. Evolution & Development, 14(1), 20-28. doi:10.1111/j.1525-142x.2011.00519.x

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing Structures in Midstream: Learning Along the Statistical Garden Path. Cognitive Science, 33(6), 1087-1116. doi:10.1111/j.1551-6709.2009.01041.x

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. Journal of Vision, 15(4), 16. doi:10.1167/15.4.16

Hansson, B., & Stensmyr, M. (2011). Evolution of Insect Olfaction. Neuron, 72(5), 698-711. doi:10.1016/j.neuron.2011.11.003

Johansson, T. (2009). Strengthening the Case for Stimulus-Specificity in Artificial Grammar Learning. Experimental Psychology, 56(3), 188-197. doi:10.1027/1618-3169.56.3.188

Jungé, J. A., Scholl, B. J., & Chun, M. M. (2007). How is spatial context learning integrated over signal versus noise? A primacy effect in contextual cueing. Visual Cognition, 15(1), 1-11. doi:10.1080/13506280600859706

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. Cognition, 83(2). doi:10.1016/s0010-0277(02)00004-5

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. Cognition, 122(2), 241-246. doi:10.1016/j.cognition.2011.10.007

Matthias, B. (2012). How Sensitive Is the Human Visual System to the Local Statistics of Natural Images? Front. Comput. Neurosci. Frontiers in Computational Neuroscience, 6. doi:10.3389/conf.fncom.2012.55.00053

Oskarsson, A. T., Boven, L. V., Mcclelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. Psychological Bulletin, 135(2), 262-285. doi:10.1037/a0014821

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. Journal of Experimental Psychology: General, 125(2), 123-138. doi:10.1037/0096-3445.125.2.123

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. Science, 274(5294), 1926-1928. doi:10.1126/science.274.5294.1926

Sandoz, J. C. (2011). Behavioral and Neurophysiological Study of Olfactory Perception and Learning in Honeybees. Front. Syst. Neurosci. Frontiers in Systems Neuroscience, 5. doi:10.3389/fnsys.2011.00098

Schnee, J. E. (1977). Predicting the unpredictable: The impact of meterological satellites on weather forecasting. Technological Forecasting and Social Change, 10(3), 299-307. doi:10.1016/0040-1625(77)90026-9

Stamps, J. A., Briffa, M., & Biro, P. A. (2012). Unpredictable animals: Individual differences in intraindividual variability (IIV). Animal Behaviour, 83(6), 1325-1334. doi:10.1016/j.anbehav.2012.02.017

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. Psychological Bulletin, 139(4), 792-814. doi:10.1037/a0030801

Voss, J. L., Federmeier, K. D., & Paller, K. A. (2011). The Potato Chip Really Does Look Like Elvis! Neural Hallmarks of Conceptual Processing Associated with Finding Novel

Shapes Subjectively Meaningful. Cerebral Cortex, 22(10), 2354-2364. doi:10.1093/cercor/bhr315

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. Psychological Review, 104(2), 301-318.

Fitelson, B., & Osherson, D. (2012). Remarks on random sequences. Retrieved from http://arxiv.org/abs/1205.5865

Ladouceur, R., Paquet, C., & Dube´, D. (1996). Erroneous perceptions in generating sequences of random events. Journal of Applied Social Psychology, 26, 2157–2166.

Montada, L. & Lerner, M.J. (1998). Preface, in Leo Montada & M.J. Lerner (Eds.). Responses to victimizations and belief in a just world (pp. vii–viii). Plenum Press: New York.

Oskarsson, A., Van Boven, L., McClelland, G., & Hastie, R. (2009). What's next? Judging sequences of binary events. Psychological Bulletin, 135(2), 262-285.

Tversky, A., & Kahneman, D. (1871). Belief in the law of small numbers. Psychological Bulletin, 76(2), 105-110.

Zhao, J., Hahn, U., & Osherson, D. (2014). Perception and identification of random events. Journal of Experimental Psychology: Human Perception and Performance, 40(4), 1358-1371.

Kachergis, G., Yu, C., & Shiffrin, R. (2012). Cross-situational word learning is better modeled by associations than hypotheses. 2012 IEEE International Conference On Development And Learning And Epigenetic Robotics (ICDL). http://dx.doi.org/10.1109/devlrn.2012.6400861

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. Proceedings Of The National Academy Of Sciences, 108(22), 9014-9019. http://dx.doi.org/10.1073/pnas.1105040108

Yu, C., & Smith, L. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. Psychological Science, 18(5), 414-420. http://dx.doi.org/10.1111/j.1467-9280.2007.01915.x

Turk-Browne, N., & Scholl, B. (2009). Flexible visual statistical learning: Transfer across space and time. Journal Of Experimental Psychology: Human Perception And Performance, 35(1), 195-202. http://dx.doi.org/10.1037/0096-1523.35.1.195

Turk-Browne, N., Scholl, B., Chun, M., & Johnson, M. (2009). Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. Journal Of Cognitive Neuroscience, 21(10), 1934-1945. http://dx.doi.org/10.1162/jocn.2009.21131

Trueswell, J., Medina, T., Hafri, A., & Gleitman, L. (2013). Propose but verify: Fast mapping meets cross-situational word learning. Cognitive Psychology, 66(1), 126-156. http://dx.doi.org/10.1016/j.cogpsych.2012.10.001