Inferences about Uniqueness in Statistical Learning

Anna Leshinskaya (alesh@sas.upenn.edu) and **Sharon L. Thompson-Schill (sschill@psych.upenn.edu)**
Department of Psychology, University of Pennsylvania
Stephen A. Levin Building, 425 S. University Ave
Philadelphia, PA 19104

Abstract

The mind adeptly registers statistical regularities in experience, often incidentally. We use a visual statistical learning paradigm to study incidental learning of predictive relations among animated events. We ask what kinds of statistics participants automatically compute, even when tracking such statistics is task-irrelevant and largely implicit. We find that participants are sensitive to a quantity governing associative learning, ΔP, independently of conditional probabilities and chunk frequencies, as previously considered. ΔP specifically reflects the uniqueness, as well as strength, of conditional probabilities; we find that uniqueness is equally affected by a single strong alternative predictor as by several weak predictors. Performance is well captured with an adapted version of the Rescorla-Wagner delta learning rule (Rescorla & Wagner, 1972). We conclude that incidental predictive learning is governed by considerations of uniqueness, and that this is computed by normalizing conditional probabilities by events' base-rates. This opens the possibility of common mechanisms between statistical learning, associative learning, and causal inference.

Keywords: statistical learning; associative learning

Introduction

Our minds adeptly register stable patterns in experience. For example, we spontaneously encode the predictive relations among sequentially presented stimuli (Saffran, Aslin, & Newport, 1996; Turk-Browne, Jungé, & Scholl, 2005). This phenomenon, *statistical learning*, takes place without conscious effort, feedback, or reward. It suggests the existence of a learning mechanism operating in the background of our minds to produce mental models of our world. Two major questions are just how inferentially complex this mechanism might be, and what are its relations to other learning phenomena.

The literature on causal reasoning, contingency learning, and classical conditioning (for a review: Mitchell, De Houwer, & Lovibond, 2009) universally describes a core phenomenon of some inferential complexity: learners do more than register that two stimuli co-occur, but also compute whether they predict each other uniquely and independently, as if attempting to determine a causal model. Suppose two events A and B coincide, such that after most occurrences of A, B occurs. However, B also occurs *without* A at a very high rate. One would not represent a strong link between A and B in this case. This consideration is captured by a foundational learning formula, ΔP (Allan, 1980; Rescorla & Wagner, 1972; Shanks, 1985):

$$\Delta P = P(A|B) - P(A|{\sim}B)$$

This equation states that learning is a product of both how often B follows A, as well as how often it appears without it. This consideration is at the core of the Rescorla-Wagner learning rule, which accounts for phenomena concerning uniqueness in classical conditioning and contingency learning, termed *cue selection* or *blocking* effects (Kamin, 1968; Rescorla & Wagner, 1972; Shanks, 1985). After learning that stimulus B coincides with an outcome (B+), learning that both A and B yield an outcome (AB+) leads to weak associative strength between A and the outcome; this less the case than if participants saw only AB+ trials. Thus, judgments about the predictive importance of A are affected by whether a different stimulus is also predictive. That is, learners care about the unique predictiveness of stimuli.

Surprisingly, whether participants' learning follows this uniqueness principle in statistical learning tasks has not been tested (c.f., Sobel & Kirkham, 2006, 2007). The extension is not trivial because statistical learning tasks differ substantially from those of causal reasoning and classical conditioning. There are two important differences. The first is that in statistical learning of temporal relations, stimuli typically appear one by one, and do not appear in compounds like in classical blocking paradigms. The uniqueness of a prediction from A to X would thus depend on whether X follows stimuli other than A (e.g., B) on separate trials. While different in form, this still captures the deeper principle behind cue selection or blocking: that causes should increase the probability of their effects above and beyond what one might expect otherwise. However, the influence of such BX trials on AX representations would require a more sophisticated computation than the Rescorla-Wagner model in standard form (Kruschke, 2008) for reasons we discuss in depth later, and furthermore, may not occur as robustly as classic blocking.

A second difference is that learning in statistical learning tasks is largely incidental. In associative and causal tasks, the goal to predict the relevant outcomes is strongly incentivized by task instructions or the inherent value of stimuli (shocks or rewards). By contrast, in statistical learning paradigms, participants passively observe streams of events, with no strategic advantage or instruction to identify predictive relations. The number of stimuli, and their rapid, continuous presentation, prevents explicit tracking of their rates of co-occurrences. Learning thus takes place incidentally, with contents typically unavailable for conscious report (Brady & Oliva, 2008; Kim, Seitz, Feenstra, & Shams, 2009). This is unlike all causal reasoning experiments and many conditioning experiments

(Mitchell et al., 2009). But while the learning situations differ, principles of learning may be preserved (Sobel & Kirkham, 2007).

In Experiment 1, we demonstrate that statistical learning is subject to considerations of uniqueness: that learning reflects not just the conditional probability relating two events, a *cause* and an *effect*, but whether that relation is unique. Experiment 2 suggests that non-uniqueness can arise from either a strong alternate predictor, or the overall base-rate of the effect. We then adapt the Rescorla-Wagner learning model (Rescorla & Wagner, 1972) to account for these results. Overall, we suggest that during incidental learning, conditional probabilities are spontaneously normalized by events' base-rates, and that learning takes place when these relative values are high. This results in sensitivity to uniqueness. In this way, we suggest that there is more in common between statistical and associative learning than previously considered. Experiment materials, data, and code are available at https://osf.io/up8qz/.

Experiment 1

We tested whether learners are sensitive to uniqueness in a visual statistical learning (VSL) task. Participants saw two distinct event sequences, each composed of a unique set of animated events (Figure 1A). Each sequence contained one strongly predictive event pair—a *cause* and an *effect*—whose uniqueness we varied. In both sequences, the first term in the ΔP formula above was matched: the probability that the effect appeared given that the cause appeared on the previous trial was equally high in both[1]. However, in the low ΔP sequence, we increased the value of the second term, P(effect|~cause), by having the effect follow two other events and itself more often than in the high ΔP sequence. Thus, the two conditions were matched in terms of the transition probability from cause to effect, as well as in the number of times a cause-effect pair appeared overall (*chunk frequency*), but differed in terms of how uniquely the cause, rather than other events, predicted the effect. We expected learning to be worse in the low ΔP condition.

Method

Participants 100 participants were recruited and tested via Amazon Mechanical Turk. Participants provided electronic consent and procedures were approved by the Institutional Review Board of the University of Pennsylvania. Compensation was $3, with a bonus of up to $2.50 based on task accuracy. Participants were excluded if they had previously participated in a related experiment or this one (10), for failing an attention measure (8), for missing data (1) or reporting a technical glitch (1). 80 participants were included (43 female, age $M = 33$, range 19 – 62).

[1] ΔP has been used to describe simultaneous co-occurrences as well as sequential ones. For example, a patient taking a medicine and experiencing a headache afterwards is a common example of the latter (Cheng, 1997).

Stimuli & Design Stimuli were sequences of animated events which took place surrounding or involving a continually present object. Each participant was shown two types of sequences, *low* Δ*P* and *high* Δ*P,* each cued by a distinct object. Each sequence was 500 events long, split into a passive preview (100) and two cover task segments (200 each), and contained eight unique events (1.2s duration), plus a static event showing object standing still (3.6s). The eight events formed 4 pairs, which were subtly visually altered versions of each other (e.g., blue vs. pink bubbles). One alternate appeared 10% of the time instead of the other; this was for the purposes of the cover task (see below). Eight different events composed the other sequence. The specific events and novel object assigned to each sequence type were selected randomly for 20 sets of materials; then, yoked materials were created by swapping the event assignments between the two sequence types. Each yoked set was then used twice, once for each possible order of presentation of the two sequences, creating materials for 80 participants fully counterbalanced for order and stimuli. The order of events in each sequence type was governed by a distinct pairwise transition matrix (Figure 1B); this was specified over the 4 event types plus static (the rare alternate replaced its pair on 10% of randomly chosen instances). Each matrix specified that the *effect* followed the

B

high △P condition (.97)

trial n	trial n+1: cause	effect	random1	random2	static
cause	0.00	0.98	0.00	0.00	0.02
effect	0.07	0.00	0.47	0.45	0.00
random1	0.15	0.01	0.38	0.41	0.05
random2	0.16	0.01	0.40	0.38	0.05
static	0.28	0.01	0.28	0.29	0.14
frequency	0.13	0.14	0.35	0.35	0.04

low △P condition (.61)

cause	0.00	0.97	0.00	0.00	0.03
effect	0.03	0.14	0.41	0.41	0.00
random1	0.27	0.66	0.00	0.00	0.08
random2	0.28	0.65	0.00	0.00	0.08
static	0.27	0.01	0.29	0.28	0.16
frequency	0.13	0.44	0.20	0.20	0.04

Figure 1. A) Static images depicting several of the stimuli, with common vs. rare alternates depicted in the top vs. bottom row, and the two objects which cued the distinct sequences. B) Mean transition matrices governing the appearance of events in each condition, and event frequencies below.

cause with a high conditional probability (~98%), whereas other transitions were lower. The critical manipulation was of the probability of the effect given other events, including itself, computed as the proportion of time any event other than the cause was followed by the effect vs. was not. In the *high* ΔP sequence, the effect was rarely preceded by any other event, yielding a high ΔP (97%). In the *low* ΔP sequence, the effect was often preceded by one of the two other, non-static events or itself, yielding a lower ΔP (61%). Individual sequences were generated stochastically according to an idealized transition matrix; the mean obtained matrix values are shown in Figure 1B. The sequence conditions were closely matched on the conditional probability of seeing the effect given the cause (*high*, 97.8%; *low* 97.4%), and their chunk frequency (number of times participants saw a cause-effect pair; *high* 63.28; *low* 62.18). The conditional probability of the cause given the effect was lower for the low ΔP sequence (*high* .074, *low* .027) so that the number of times they saw the effect-cause pair (*high* 5.12; *low* 6.02) was relatively matched. We address any concerns about minor differences in the sequence properties in the Results.

Procedure Participants' cover task was to learn to identify the "common" vs. "rare" version of each event type. To ensure task comprehension, static images depicting both versions were shown (but not which were which), and a 20-trial practice task required them to hit the spacebar in response to the start of each distinct event; 80% accuracy was required to continue. Subsequently, they watched a preview video to learn to identify the common vs. rare events, followed by 2 videos in which they were asked to hit 'r' for rare and 'o' for common. They then saw the second sequence type, starting again with static images of the new events, the practice task, preview, and 2 task videos. They were given feedback on cover task accuracy following each task video. Order of the sequence types was counterbalanced across subjects within yoked stimulus set.

Following *all* videos, a surprise forced-choice test probed participants' knowledge of the cause-effect relation in both sequences separately. The critical questions (repeated 3 times and responses averaged) showed the cause followed by the effect in one video, and the effect followed by the cause in the other; participants had to choose the video that seemed more typical or familiar. We chose this as the test question so as to match the two videos on individual stimulus frequency, and to match question difficulty across conditions (ensuring they have equal relative transition probabilities). Only the common stimulus versions were used in testing. The stimulus pairs appearing in these questions were also matched in chunk frequency between the conditions, and accordingly had a slightly larger difference in transition probability in the low ΔP condition (which would make it easier, counter to our hypothesis). Two filler questions were also presented to avoid giving the impression that the test was stuck on a single question.

To probe explicit (verbalizable) access to the sequence statistics, participants were given a freeform question asking "Did certain events follow each other more often than others? Describe any you noticed for the first set and for the second set of videos". They were also asked, "How confident are you that you detected any systematic order to the events in the [first/second] set of videos?", for each set, and responded on a 1-5 rating scale, with 1 = Definitely False, 3 = Unsure, and 5 = Definitely True.

Results

Participants performed well on the cover task for both sequences (*high* ΔP M = 86.99, SE = 1.03; *low* ΔP M = 87.37, SE = 0.94; $t(79) = -0.39$ p = .698). On the critical questions of the forced choice test, participants were above chance for the high ΔP sequence (M = 61.83%, SE = 3.90%, $t(79) = 2.79$, p = .007, d = 0.31) but below chance for the low ΔP sequence (M = 41.67%, SE = 3.89%, $t(79) = -2.16$, p = .034, d = -0.24), which were significantly different from each other (CI [8.43, 29.90], $t(79) = 3.55$, $p < .001$, d = 0.55), as shown in Figure 2. This supports the idea that participants had a weaker representation of the cause-effect relationship in the low ΔP condition—despite the fact that in both conditions, cause-effect transitions occurred twelve times as often as effect-cause transitions. To rule out that this was due to the minor difference in the number of times participants saw a cause-effect transition relative to an effect-cause transition in the two conditions, we computed the difference in the number of times each participant saw a cause-effect pair vs. an effect-cause pair between the two conditions (M = 2). This difference was uncorrelated with the difference in accuracy between the two conditions ($r(78) = -0.03$, p = .791). There was no effect of training order (which sequence participants saw first), p = .273.

Participants' confidence that they noticed any systematic order among the events was not reliably above 'unsure' for either condition (*high* M = 3.20, SE = 0.13, $t(79) = 1.57$, p = .121; *low* M = 3.06, SE = 0.13, $t(79) < 1$), with no effect of condition ($t(79) = 1.52$, p = .132). In their freeform responses, 20/80 participants described a relation between the cause and effect for one sequence, but only 2/80 did so for both. However, participants were more likely to describe it for the high ΔP events (19/80) than the low ΔP events (5/80; X^2 (1) = 9.61, p = .002). Thus, the high ΔP condition

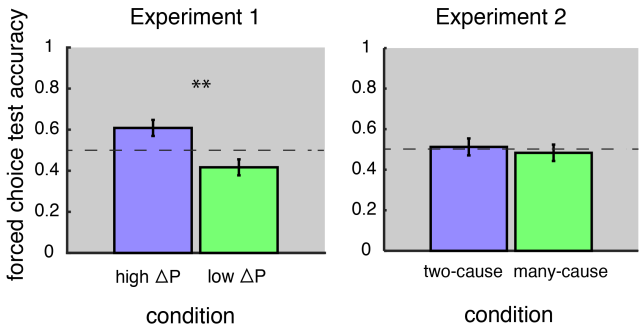

Figure 2. Forced-choice test accuracy by condition in Experiments 1 and 2. ** = $p < .001$.

enabled participants to better notice the predictive pattern. In contrast, participants were not likely to report effect-cause patterns in either condition: there were no cases of this for the low ΔP condition, and one in the high ΔP condition. This argues against the idea that they strongly believed in the effect-cause relation in the low ΔP condition.

Discussion

During a VSL task, participants' learning of a highly likely transition between two events (a *cause* and an *effect*) was weaker when the effect was also preceded by other events (low ΔP condition), relative to when the effect was uniquely predicted by the cause (high ΔP condition). The cause-effect transition had a higher chunk frequency and transitional probability than the effect-cause transition in both conditions: participants saw cause-effect pairs a similarly high number of times, and effect-cause pairs a similarly small number of times. However, when asked which was more typical, participants chose the cause-effect pair reliably only in the high ΔP condition.

This finding indicates that participants' incidental learning is automatically informed by computations of uniqueness, in that neither participants' cover task nor the test questions demanded it or benefitted from it. Answers based on chunk frequency or conditional probability were both valid, and computationally simpler. Thus, the computation of ΔP was a product of participants' incidental learning process. Furthermore, a high ΔP between cause and effect led to a higher rate of reporting of this pattern when asked to describe any systematic order to the events. It seems unlikely that participants had noticed that the low ΔP cause was nearly always followed by the effect, but were less willing to report it. We thus conclude that the rate of awareness was itself influenced by ΔP.

These results are in line with a related finding that participants have poor *memory*—not just low causal ratings—for non-unique ("blocked") cause-effect relations (Mitchell, Lovibond, Minard, & Lavis, 2006). They are also consistent with demonstrations of more traditional blocking effects in infants in a paradigm somewhere in between classical conditioning and statistical learning (Sobel & Kirkham, 2006, 2007). Here we show that effects of uniqueness hold in a traditional statistical learning paradigm, characterized by sequentially presented events, incidental learning, absence of incentive or reward, and conditions which exceed the capacity to track event statistics explicitly.

A puzzling fact of these data is the below-chance performance in the low ΔP condition: participants selected the effect-cause transition more often than cause-effect. It did not appear that participants strongly believed in this effect-cause transition, as they did not mention it even once in their freeform responses. This suggests that participants may have had an *inhibited* (implicit) belief in the cause-effect relation.

An additional open question is the source of this blocking-like effect. Here and elsewhere, it is ambiguous whether the weakened belief in a low ΔP cause-effect pair is due to a strong associative representation of an alternate event-effect pair, or from an increased base-rate of the effect's occurrence. For example, participants might have developed a strong belief that random event 1 or 2 actually caused the effect, and subsequently inhibited their belief in other possible predictors. Alternatively, participants could compute ΔP by normalizing the conditional probabilities of the effect-cause pair by the base-rate of both the effect and the cause, having tracked that the effect event occurs often overall. We test this possibility in Experiment 2.

Experiment 2

As in Experiment 1, participants were exposed to two VSL sequences, each containing a strong predictive relation between a *cause* and *effect*. In the *two-cause* stream, the effect also followed a second, equally strong predictor, but no other events. In the *many-cause* stream, the effect followed the three other events and itself with a medium probability (~18%). Thus, the cause-effect pair in both conditions had an equal ΔP value (.80), but from different sources: a strong alternative or a high overall base-rate. If learning is impaired specifically due to a salient alternative cause, participants should perform worse in the *two-cause* stream relative to the *many-cause* stream.

Method

Participants 107 participants were recruited and tested via Amazon Mechanical Turk. Participants provided electronic consent and procedures were approved by the Institutional Review Board of the University of Pennsylvania. Compensation was $3, with a bonus of up to $2.50 based on task accuracy. Participants were excluded if they had previously participated in a related experiment or this one (7), for failing an attention measure (19), or for missing data (1). 80 participants were included (47 female, age M = 34, range 19 – 71).

Stimuli & Procedure Stimuli and procedures were similar to Experiment 1, except that each sequence contained 5 events plus static, and the first task video was slightly shorter (150 events). The transition matrices were similar, except that the *two-cause* condition had a 96% transition probability between random1 and effect; and the *many-cause* condition had a ~18% transition probability between each non-static event and the effect. Conditions had similar counts for cause-effect and effect-cause pairs, and similar ΔP value between cause and effect (.80).

Results

Participants performed well on the cover task for both sequences (*two-cause*: M = 85.43, SE = 1.11; *many-case*: M = 86.63, SE = 1.09, $t(79) < 1$). On the critical trials in the forced-choice test (Figure 2), performance was not different from chance in either condition (*two-cause*: M = 51.25%, SE = 4.17, $t(79) < 1$; *many-cause*: M = 48.33%, SE = 4.04, $t(79) < 1$) with no difference between them ($t < 1$).

Discussion

When shown a cause-effect relation accompanied by either an alternative strong predictor of the same effect (*two-cause* condition) or several weak predictors (*many-cause* condition), participants showed equally poor learning of the effect-cause relation. It seemed not to matter for learning whether there was a salient alternative cause, or simply a higher base rate of the effect. Of course, this null finding must be bolstered by positive evidence that performance is worse on the *many-cause* condition relative to an appropriately matched high ΔP condition. Preliminary data from an experiment with this manipulation, and otherwise identical methods to Experiment 1, are indeed in line with this prediction ($t[37] = 2.30$, $p = .027$, $d = 0.53$). Overall, the available evidence makes it unlikely that blocking (i.e., weak representation of non-unique relations) strictly requires a strong alternative predictor, and that rather, it can arise equally well from a high overall base rate.

Modeling

As introduced earlier, the Rescorla-Wagner (R-W) learning rule was developed to account for uniqueness effects (i.e., blocking) in associative learning tasks (Rescorla & Wagner, 1972). Here we use an adaptation of this learning rule for sequential stimuli to test whether it can account for our effects. We thus assume that learners acquire all pairwise associative weights among the stimuli, which they can compare to solve the forced choice task.

We first found that a straightforward implementation of the standard R-W rule did not account for our effect in Experiment 1: it was only minimally sensitive to the difference between the two conditions. This is consistent with prior observations that the R-W rule has difficulty accounting for certain types of blocking, specifically those which involve updating weights for stimuli using evidence (trials) in which those stimuli are not actually present—that is, using BX trials to update AX representations (Kruschke, 2008; Shanks, 1985). Bayesian models like the Kalman filter circumvent this problem because they require the weights from all causes to an effect to sum to 1, since they must reflect probabilities (Kruschke, 2008). We adopted just this property here, by adding a normalization step at each trial: we required the set of weights *to* each event from all others to sum to 1. This meant that weights could trade off, such that stronger AX weights would naturally reduce BX weights. We found that this simple adaptation enabled us to capture the difference between conditions.

Model Details

The model learns the weight of links among all pairs of stimuli 1 to s, represented by matrix **W** with $s \times s$ entries. We allow the learner to know the number of stimuli, and to begin with the assumption that all transitions are equally likely; thus, all entries in **W** are initially set to $1/s$. Learning operates sequentially over observed input stream **o**, which consists of a sequence of the stimuli from set {1:s}. At each observation $\mathbf{o}_j$, stimulus i is shown. The model compares this observation with its prediction for $\mathbf{o}_j$ on the basis of the preceding n stimuli ($\mathbf{o}_{j-n}$ ... $\mathbf{o}_{j-1}$, which consists of stimuli $k_{1:n}$ also in set {1:s}) and the weights between $k_{1:n}$ and i, as represented in **W.** Its error in anticipating the stimulus is used to adjust the entries in **W** between stimuli $k_{1:n}$ and stimulus i, which are then used in subsequent predictions. Formally, the degree of anticipation of stimulus i on trial j is given by:

$$a_i = \sum_{k=1}^{n} \boldsymbol{w}_{\mathrm{ki}}$$

For simplification, we set $n = 1$, such that this is the entry between the current stimulus i and the preceding stimulus k. The error is computed by:

$$d_i = 1 - a_i$$

Because stimulus i occurred in binary fashion, its observed value is 1. If this was well anticipated by the previous stimuli, a_i should be close to 1, resulting in a low error. The corresponding entries in **W** are adjusted by this error multiplied by a learning rate parameter α:

$$\Delta \mathbf{W}_{ki} = \alpha\, d_i$$

Critically, following this adjustment, the columns of **W** (input weights to each event) are normalized to sum to 1.

Model Results & Discussion

The event sequences given to human participants in Experiment 1 served as inputs to the model. The outputs were the adjusted matrices **W**, for each set of materials ('runs'). We assumed that forced-choice behavior reflects the relative weight of links between cause & effect and effect & cause pairs, and thus, compare these entries in **W.** Using a learning rate of .5, we found significantly stronger weights for cause-effect than effect-cause links, in both conditions (*high* ΔP, cause-effect $M = 0.70$, $SE = 0.02$; effect-cause $M = 0.11$, $SE = 0.01$, $t(79) = 21.92$, $p < .001$; *low* ΔP, cause-effect $M = 0.19$, $SE = 0.01$, effect-cause $M = 0.07$, $SE = 0.01$, $t(79) = 9.07$, $p < .001$), with a significantly larger difference in the *high* ΔP condition ($t(79) = 17.35$, $p <$

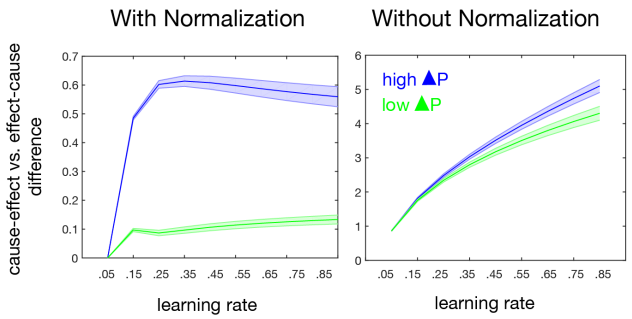

Figure 3. Difference in weights between cause-effect and across runs.

.001). The same pattern is seen consistently across settings of the learning rate parameter (.05 - .85; Figure 3). This is important because, given the relatively weak human learning, we expect that the actual learning rate is fairly low.

Without column normalization, the model does less well; while it shows effects of condition at high learning rates (>.35), it does not reliably do so at lower learning ones, and across rates, effects are much weaker (Figure 3). This is intuitive: the weight from B to X will only be affected by evidence of A to X trials if the weights to X trade off, such that as one gets stronger, the rest weaken (Kruschke, 2008). This can also be seen as the effect of representing each link as proportional to the base rate of X, if the base rate of X is, as here, captured in how often it appears following the other events in the state space.

Normalization—conversion to relative rather than absolute values—is a cognitively realistic and adaptive mechanism, although future research should investigate the circumstances under which associative strengths are normalized or not. Furthermore, our simplistic model will fail to exhibit the advantages of a Bayesian formulation, such as in explicit representation of uncertainty (Kruschke, 2008) and should be compared to other versions of the R-W update rule (Dickinson, 2001). Our ongoing work investigates whether these factors also describe learning in VSL tasks.

Conclusions

Our key finding was that participants in a statistical learning task were sensitive to not only the conditional probability between two events, but also the uniqueness of that relation. We suggest that low uniqueness can be established by either a competing predictor, on the one hand, or an overall high base rate of the predicted effect, on the other. Both can be treated as the result of normalization: the assumption that predictors of the same effect trade off, and to be considered effective, must raise the probability of the effect above its rate of occurrence otherwise.

This finding brings statistical learning in closer contact with the rich literature in associative learning and causal reasoning, despite differences in the nature of these learning tasks. Work on causal reasoning has extensively shown that participants use the uniqueness of a predictive relation to attribute causality (Cheng, 1997). We find that this computation additionally takes place incidentally and automatically, suggesting it is one consideration for how we register the naturally occurring statistics of our observed world.

Acknowledgments

This work was supported by NIH grant R01DC015359 to S.L.T-S.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. *Psychological Science*, *19*(7), 678–685.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Dickinson, A. (2001). Causal Learning: Association Versus Computation. *Current Directions in Psychological Science*, *10*(1975), 127–132.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–31). Coral Gables, FL: University of Miami Press.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters*, *461*(2), 145–149.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior*, *36*(3), 210–226.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183. h

Mitchell, C. J., Lovibond, P. F., Minard, E., & Lavis, Y. (2006). Forward blocking in human learning sometimes reflects the failure to encode a cue-outcome relationship. *Quarterly Journal of Experimental Psychology*, *59*(5), 830–844.

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*(6), 64–99.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by eight-month-old infants. *Science*, *274*(5294), 1926–1928.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology*, *37*(B), 1–21.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*(6), 1103–15.

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science*, *10*(3), 298–306.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552–64.