Research Article

Psychological Science
1–11

DOI: 10.1177/0956797619868998
www.psychologicalscience.org/PS

Automated Study Challenges the Existence of a Foundational Statistical-Learning Ability in Newborn Chicks

Samantha M. W. Wood[1], Scott P. Johnson[2], and Justin N. Wood[1]

[1]School of Informatics, Computing & Engineering, Indiana University, and [2]Department of Psychology, University of California, Los Angeles

Abstract

What mechanisms underlie learning in newborn brains? Recently, researchers reported that newborn chicks use unsupervised statistical learning to encode the transitional probabilities (TPs) of shapes in a sequence, suggesting that TP-based statistical learning can be present in newborn brains. Using a preregistered design, we attempted to reproduce this finding with an automated method that eliminated experimenter bias and allowed more than 250 times more data to be collected per chick. With precise measurements of each chick's behavior, we were able to perform individual-level analyses and substantially reduce measurement error for the group-level analyses. We found no evidence that newborn chicks encode the TPs between sequentially presented shapes. None of the chicks showed evidence for this ability. Conversely, we obtained strong evidence that newborn chicks encode the shapes of individual objects, showing that this automated method can produce robust results. These findings challenge the claim that TP-based statistical learning is present in newborn brains.

Keywords

statistical learning, newborn, chick, transitional probability, controlled rearing, open data, open materials, preregistered

Received 9/14/18; Revision accepted 7/15/19

A core goal in psychology is to understand the origins of cognition. This requires characterizing the learning mechanisms in newborn brains. One candidate mechanism is statistical learning. A growing consensus across developmental psychology (Krogh, Vlach, & Johnson, 2012; Saffran, Aslin, & Newport, 1996; Smith, Suanda, & Yu, 2014) and computational neuroscience (DiCarlo, Zoccolan, & Rust, 2012; Wiskott & Sejnowski, 2002) is that animals learn to interpret sensory input through statistical learning by associating features that co-occur in the input stream. Are statistical-learning abilities present in newborn brains? What roles do experience and maturation play in the development of these abilities?

In a study published in *Current Biology*, Santolin, Rosa-Salva, Vallortigara, and Regolin (2016) reported that newborn chicks can encode the transitional probabilities (TPs) between sequentially presented shapes. The chicks were reared for 2 hr with a structured visual sequence of shapes, and the order of the shapes was defined by TPs within and between shape pairs. The chicks were then tested with a two-alternative forced-choice task to examine whether they could discriminate the familiar (structured) sequence from a novel sequence, either unstructured (Experiment 1) or with a new set of TPs between shapes (Experiment 2). The chicks showed a preference for the novel sequence, suggesting that chicks can encode the TPs between sequentially presented shapes.

Corresponding Authors:
Samantha M. W. Wood, Indiana University, School of Informatics, Computing & Engineering, 700 N. Woodlawn Ave., Bloomington, IN 47408
E-mail: sw113@iu.edu

Justin N. Wood, Indiana University, School of Informatics, Computing & Engineering, 700 N. Woodlawn Ave., Bloomington, IN 47408
E-mail: woodjn@indiana.edu

This is a potentially important finding for two reasons. First, this study indicates that TP-based statistical learning can be present in newborn brains during the earliest stages of visual learning. TP-based statistical learning might therefore be a foundational learning mechanism in newborn brains. Second, the Santolin et al. (2016) study suggests that newborn chicks have more powerful statistical-learning abilities than newborn humans. Although human infants encode some statistical relations at birth (Bulf, Johnson, & Valenza, 2011), they fail to encode TPs until 5 months of age (Marcovitch & Lewkowicz, 2009; Slone & Johnson, 2015). This pattern of results implies that maturation or experience (or both) play a role in the development of TP-based statistical learning for humans, but not for chicks. To understand TP-based statistical learning, we must therefore understand how (and why) its developmental trajectory could differ so radically across species. Given the potential import of the Santolin et al. study for understanding the origins, development, and evolution of TP-based statistical learning, it is crucial to ensure that the results are accurate and robust.

Although Santolin et al. (2016) tackled an important theoretical question, the study had three limitations: noisy measurements (low signal-to-noise ratio), small effect sizes, and high analytic flexibility (e.g., flexibility in terms of the number of subjects that were tested). These three factors are now known to be key contributors to the replication crisis (Munafò et al., 2017). When data are noisy, studies often produce estimates of performance that are considerably higher (or lower) than the true population performance (Loken & Gelman, 2017). These inflated estimates of performance can significantly increase false-positives rates, especially when researchers have flexibility in terms of the number of subjects that are tested, analyses that are performed, and results that are reported (Simmons, Nelson, & Simonsohn, 2011). Here, we attempted to reproduce the results from the Santolin et al. study using a preregistered design (which limits analytic flexibility) and an automated method (which allows large amounts of precise behavioral data to be collected from each chick).

Fueled by innovation in image-based tracking software, it is now possible to fully automate controlled-rearing experiments with newborn chicks (J. N. Wood, 2013). Automation removes the possibility of experimenter bias and allows chicks' behavior to be measured continuously (24 hr per day, 7 days per week), producing massive amounts of data per subject. Conversely, with nonautomated methods, researchers must collect data manually, limiting the amount of data collected per subject. For instance, Santolin et al. (2016) collected 6 min of test data per chick, whereas we collected 5,600 min of test data per chick in Experiment 1 and 1,600 min of test data per chick in Experiment 2. Thus, across two experiments, we collected 900 and 250 times more test data per subject than in the original study. This allowed us to obtain measurements that were 3 to 4 times more precise than those in the original study, substantially improving the power of the experimental design. Our study was therefore not a *direct* replication of the Santolin et al. study but, rather, an attempt to reproduce the findings with a more powerful method.

Experiment 1

In the first week of life, newborn chicks were exposed to an imprinting sequence consisting of a stream of four shapes presented in an order defined by the TPs between shapes (Fig. 1a; see also Video S1 in the Supplemental Material available online). In the second week, we tested the chicks using a two-alternative forced-choice task (Fig. 1b; see also Videos S2 and S3 in the Supplemental Material). In the TP condition, one monitor showed a familiar TP sequence (in which the TPs between shapes matched the imprinting sequence) and the other monitor showed a novel TP sequence (in which the TPs between shapes did not match the imprinting sequence). In the shape-recognition condition, one monitor showed a sequence of familiar shapes, and the other monitor showed a sequence of novel shapes.

Method

Subjects. Twelve Rhode Island Red domestic chicks (*Gallus gallus*) of undetermined sex were tested. No subjects were excluded from the analyses. The sample size was determined by a power analysis based on the results of prior automated controlled-rearing studies with newborn chicks (J. N. Wood, 2013, 2014). Specifically, in both experiments in the Santolin et al. (2016) study, the researchers reported group performance rates of 37% (chance level = 50%). As described below, our automated method reduces measurement error by increasing the amount of data collected per chick. Thus, for our power calculations, we assumed a performance rate that was similar to the one used in the original study (37%) but with less measurement error. Automated studies from our lab typically yield standard deviations (across subjects) of approximately 10% (S. M. W. Wood & Wood, 2019), whereas the standard deviation in the Santolin et al. study was approximately 35%. Thus, our approximate expected effect size (Cohen's *d*) was 1.3.[1] A sample size of 7 subjects was required to achieve 80% power for an effect size of 1.3 (R package *pwr*; R Core Team, 2015, Version 3.1.3). To perform a high-powered replication, we tested 12 chicks,

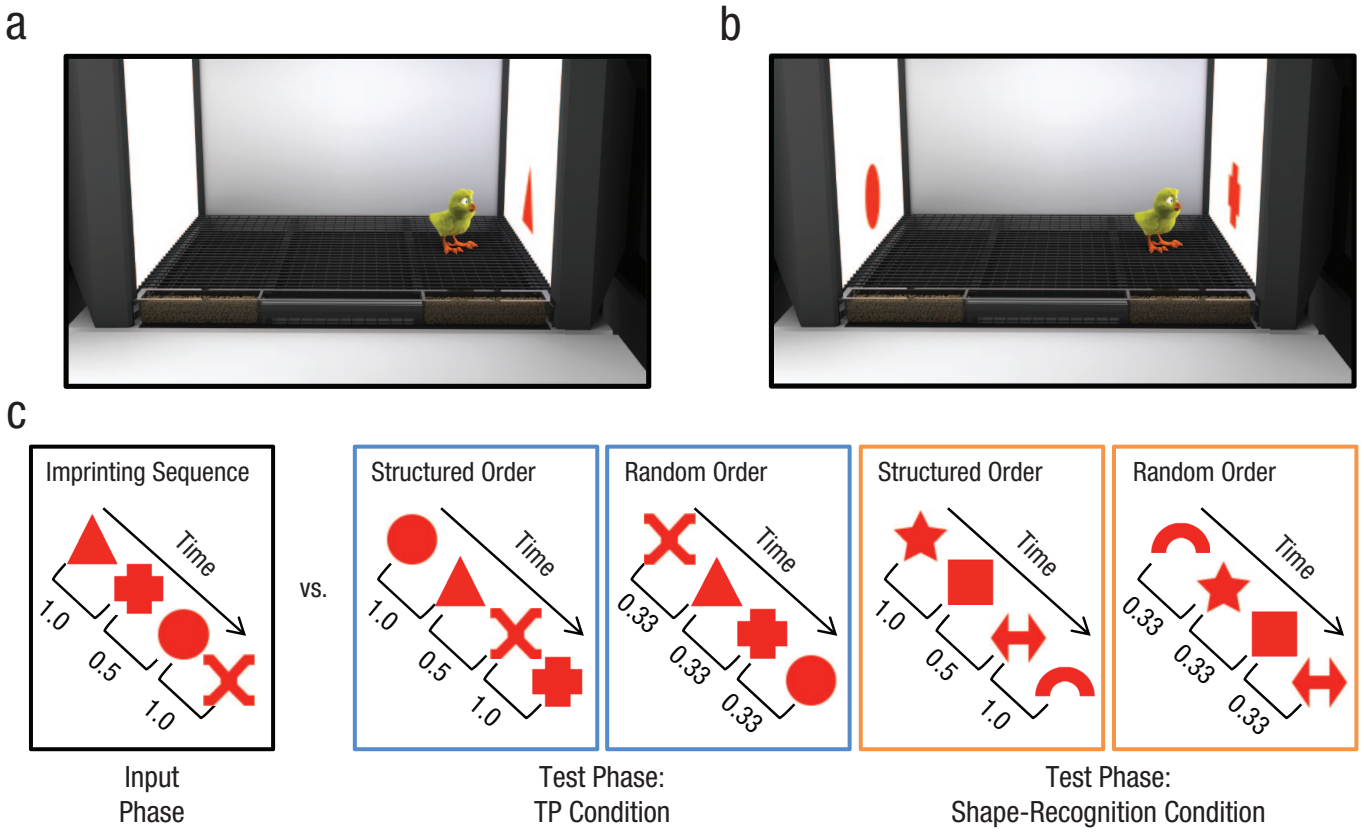

Fig. 1. Experiment 1 method. The illustrations in the top row show a controlled-rearing chamber (a) during the input phase and (b) during the test phase. The stimuli design (c) is shown in the bottom row. During the input phase, an imprinting sequence defined by the transitional probabilities (TPs) within and between shape pairs appeared on one display wall at a time. The imprinting sequence is shown in the box with the black border (c). During the test phase, we presented chicks with two-alternative forced-choice tasks. In each test trial, one display wall showed the imprinting sequence (c, box with black border), and the other display wall showed one of four novel sequences (c, boxes with blue and orange borders). In the TP condition (c, boxes with blue borders), chicks saw the same shapes as in the imprinting sequence, but in novel orders. In the shape-recognition condition (c, boxes with orange borders), the shapes were novel. In each condition, the novel sequence was presented in either a structured order or a random order. Numbers in (c) show the probability of transitioning from one shape to the next.

which powered our experiment to 98%, assuming the parameters above.

The eggs were obtained from a local distributor and incubated in darkness in an OVA-Easy incubator (Brinsea Products, Titusville, FL). The incubation room was kept in complete darkness. Within 24 hr after hatching, the chicks were moved from the incubator to the controlled-rearing chambers in darkness with the aid of night vision goggles. Each chick was raised singly within its own chamber. We maintained the room temperature at 80° F for the duration of the experiment. This research was approved by the University of Southern California Institutional Animal Care and Use Committee.

Procedure. Newborn chicks were reared for 2 weeks within specially designed controlled-rearing chambers. The chambers measured 66 cm (length) × 42 cm (width) × 69 cm (height) and contained no real-world (solid, movable) objects. To present object stimuli to the chicks, we projected virtual 2-D shapes on two display walls (Acer 19-in. LCD monitors with 1,440 pixel × 900 pixel resolution) situated on opposite sides of the chamber. We presented the stimuli 24 hr per day, 7 days per week, without a light/dark cycle. The average luminosity of the chambers was 45.3 lumens when fully lit. Food and water were available within transparent troughs in the ground that measured 66 cm (length) × 2.5 cm (width) × 2.7 cm (height). The floors of the chambers consisted of black wire mesh (0.5-in. grid spacing) supported over a black matte surface.

The chambers recorded all of the chicks' behavior (9 samples per second, 24 hr per day, 7 days per week) via microcameras in the ceilings and automated image-based tracking software (EthoVision XT; Noldus Information Technology, Leesburg, VA). Automation allowed large numbers of test trials (up to 140) to be collected from each chick; in total, approximately 4,032 hr of video footage (14 days × 24 hr per day × 12 subjects) were collected for this experiment. Accordingly, we were able

to measure each chick's performance with high precision. More generally, automated controlled-rearing methods produce measurements that are 3 to 4 times more precise and effect sizes that are 3 to 4 times larger than those produced by nonautomated studies (S. M. W. Wood & Wood, 2019). Automated studies also eliminate experimenter bias and allow for analyses on the individual subject level.

In the first week of life, the chicks were reared with animations of four shapes presented in an order defined by TPs between shapes (the imprinting sequence). The shapes were presented sequentially on one of the display walls (see Video S1). The display wall showing the imprinting sequence switched every 40 min. Each shape was presented for 2.0 s and loomed from 5 cm to 10 cm in height on the monitor. All of the shapes were red, and all of the sequences were shown on a white background to prevent reflection of the chicks on the monitor. The imprinting sequence consisted of a triangle, cross, circle, and X shape for half of the chicks and a star, square, arrow, and arc for the other half of the chicks. As in the Santolin et al. (2016) study, the imprinting sequence consisted of two shape pairs defined by statistical dependencies between and within pairs' elements. For example, the triangle was always followed by the cross (TP = 1.0), and the circle was always followed by the X shape (TP = 1.0). The item that appeared after the pair (i.e., following the cross or X shape) was the first element of one of the pairs (TP = 0.5). Repetitions of the same pair were allowed. The only cue to segment the pairs was the statistical structure of the sequence.

In the second week, the chicks were presented with two-alternative forced-choice test trials. The chicks received 140 test trials (20 trials per day). Each test trial lasted 40 min and was followed by a 30-min rest period. During the rest periods, the imprinting sequence was shown on one display wall, and the other display wall showed a white screen.

During the TP test trials (see Video S2), one display wall showed a sequence with familiar shapes defined by the TPs from the imprinting sequence (familiar TP sequence), whereas the other display wall showed a sequence with familiar shapes defined by a different statistical structure (novel-order sequences). On half of the TP test trials, the novel-order sequence showed the shapes in a random order (random novel-order sequence; cf. Santolin et al., 2016, Experiment 1). On the other half of the TP test trials, the novel-order sequence was structured by TPs, but the TP structure was different from the imprinting sequence (TP-based novel-order sequence; cf. Santolin et al., 2016, Experiment 2). Thus, if the chicks encoded the specific TP structure from the imprinting sequence, then they should have distinguished the familiar TP sequence from the novel-order sequence on both types of TP test trials. If the chicks were sensitive to only structured sequences versus unstructured sequences (as reported in newborn infants; Bulf et al., 2011), then the chicks should have distinguished the familiar TP sequence from the random novel-order sequences but not from the TP-based novel-order sequences. The shapes were the same color in the imprinting and test sequences.[2]

During the shape-recognition test trials (see Video S3), one display wall showed a sequence of familiar shapes defined by the TPs from the imprinting sequence (familiar-shape sequence), whereas the other display wall showed a sequence of novel shapes (novel-shape sequence). The novel shapes were the same color and size as the familiar shapes. On half of the shape-recognition test trials, the novel-shape sequence showed the shapes in a random order (random novel-shape sequence). On the other half of the novel-shape test trials, the novel-order sequence was structured by TPs (TP-based novel-shape sequence). If the chicks encoded the shapes from the imprinting sequence, then they should have distinguished the familiar-shape sequence from the novel-shape sequence on both types of shape-recognition test trials.

The order of the test trials was pseudorandomized to ensure that (a) each type of test trial was presented an equal number of times per day and (b) the familiar and unfamiliar sequences were presented an equal number of times on the same monitor as the imprinting sequence from the preceding rest period.

Results

Performance by trial type. To measure each chick's preference, we calculated the amount of time each subject spent within zones (22 cm × 42 cm) next to each monitor. Preference for a target stimulus was calculated as the time spent by the correct sequence divided by the time spent by both sequences. The results are shown in Figure 2. To assess whether performance differed across the test conditions, we performed a repeated measures analysis of variance (ANOVA) with within-subjects factors of trial type (TP vs. shape recognition) and novel-sequence order (random vs. TP based). The ANOVA revealed a significant main effect of trial type, $F(1, 11) = 20.655$, $p < .001$, $\eta_p^2 = .653$. The main effect of novel-sequence order, $F(1, 11) = 0.761$, $p = .402$, $\eta_p^2 = .065$, and the interaction, $F(1, 11) = 0.000$, $p = .990$, $\eta_p^2 = .000$, were not statistically significant. In the TP condition, the chicks failed to show a significant preference for either the familiar TP sequence or the novel sequence, $M = 49.406\%$, $SE = 1.087\%$, 95% confidence interval (CI) = [47.012%, 51.799%], $t(11) = 0.547$, $p = .596$, Cohen's $d = 0.158$, Bayes factor (BF)[3] in favor of the null hypothesis = 4.046 (see Fig. 2a). Conversely, in the shape-recognition condition, the chicks spent significantly more time with the

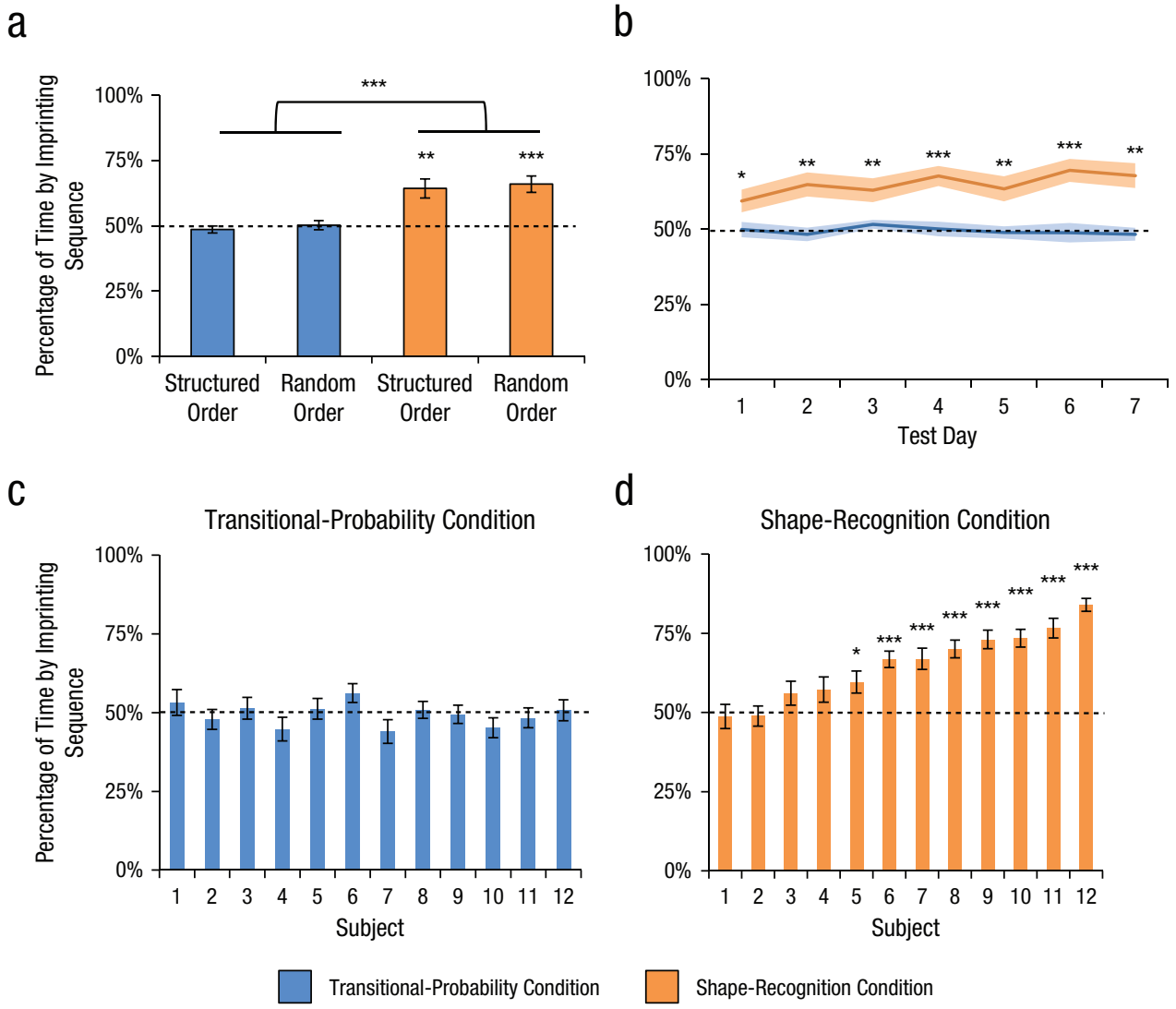

Fig. 2. Performance in Experiment 1. The overall percentage of time that chicks spent with the imprinting sequence versus the novel sequence is shown (a) for the two novel sequences in the transitional-probability condition and the two novel sequences in the shape-recognition condition. Performance is also shown (b) separately for each test day for the two conditions. Individual subject performance is shown for (c) the transitional-probability condition and (d) the shape-recognition condition. Chance performance was 50% (dashed line). The colored ribbons in (b) and the error bars in the other three panels denote ±1 *SE*. Asterisks above brackets indicate that differences between conditions were significant, and asterisks above data bars or data points indicate that results were significantly different from chance (*$p < .05$, **$p < .01$, ***$p < .001$). Individual subject's p values have been corrected for multiple comparisons using Holm-Bonferroni correction.

familiar shapes than the novel shapes, $M = 65.163\%$, $SE = 3.217\%$, 95% CI = [58.083%, 72.243%], $t(11) = 4.714$, $p < .001$, Cohen's $d = 1.361$, BF in favor of the alternative hypothesis = 64.702 (see Fig. 2a). Paired-samples t tests showed that performance was significantly higher in the shape-recognition test trials than the TP test trials, mean difference = 15.758%, $SE = 3.465\%$, 95% CI = [–23.385%, –8.131%], $t(11) = 4.547$, $p < .001$, Cohen's $d = 1.313$, BF in favor of the alternative hypothesis = 50.812. Thus, the chicks were not sensitive to the order of objects in the sequence, but they were sensitive to the shapes of the individual objects.

Individual-level performance. Because we collected approximately 90 hr of test data per chick, we were also able to perform individual-level statistical analyses and assess whether each chick performed above or below chance level. With large amounts of data from each chick, these individual-level analyses are highly powered: Many reach 5-sigma levels of statistical significance (the statistical threshold for new discoveries in theoretical physics) for individual subjects (e.g., J. N. Wood, 2013). Another benefit of individual-level analyses is that each subject can serve as a replication attempt of an effect. If we consider an experiment as being on an individual level (rather than

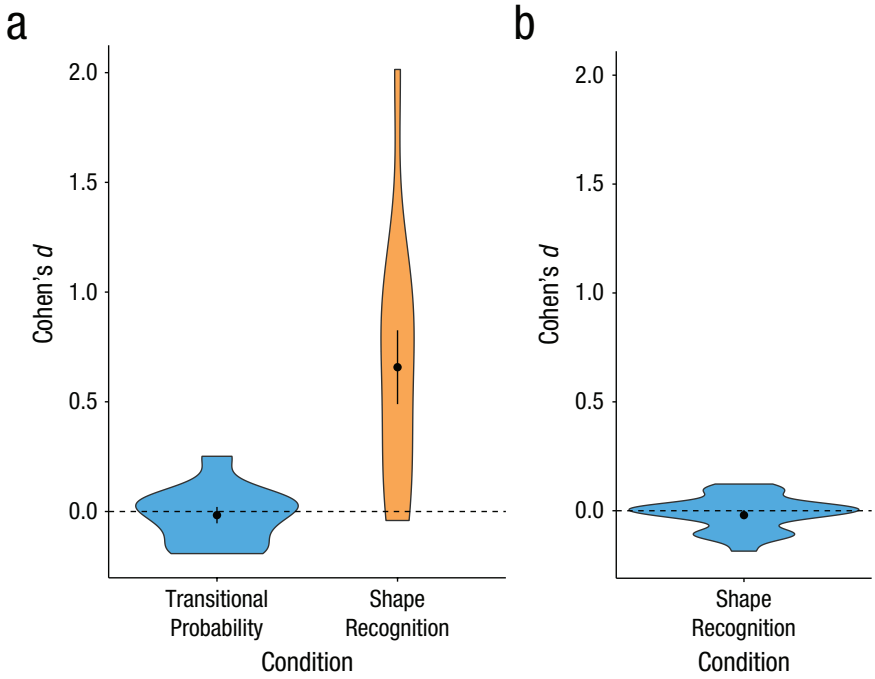

Fig. 3. Violin plots showing the distribution of the individual subjects' effect sizes (Cohen's *d*s) for the transitional-probability and shape-recognition conditions in (a) Experiment 1 and (b) Experiment 2. The dot and vertical line inside each violin plot denote the mean and standard error, respectively. The width of the plot indicates the distribution of the data. The dashed line marks an effect size of 0, or no effect.

group level), then each subject provides an opportunity to replicate an effect. Experiments at the individual level are possible only with enough measurements from each subject to reliably reject or fail to reject the null hypothesis for each individual. Thus, our automated method permits individual-level internal replications.

In the individual-level analysis, no chicks showed evidence of TP-based statistical learning (all *p*s > .05; see Fig. 2c). The average Cohen's *d* across individual subjects in the TP condition was –0.016, 95% CI = [–0.099, 0.066] (see Fig. 3). Conversely, in the shape-recognition condition, 8 of the 12 chicks showed a statistically significant preference for the familiar shapes (7 chicks: $p < .0001$; 1 chick: $p < .05$; see Fig. 2d). The average Cohen's *d* across individual subjects in the shape-recognition condition was 0.658, 95% CI = [0.287, 1.028] (see Fig. 3). Thus, on the shape-recognition task, our method produced robust results—both on the individual and group levels—with the same group of chicks that failed to encode TPs between shapes.

Change over time. Performance by test day is shown in Figure 2b. To test whether performance changed across the test phase, we performed a repeated measures ANOVA with the within-subjects factors of trial type (TP vs. shape recognition) and test day (1–7). The ANOVA revealed a significant main effect of trial type, $F(1, 11) = 20.364$, $p = .0009$, $\eta_p^2 = .649$. The main effect of test day, $F(6, 66) = 1.162$, $p = .337$, $\eta_p^2 = .096$, and the interaction of trial type and test day, $F(6, 66) = 1.405$, $p = .226$, $\eta_p^2 = .113$, were not significant. Performance was significantly above chance level on every test day in the shape-recognition condition (Holm-Bonferroni corrected, all *p*s < .05), whereas performance did not exceed chance level on any test day in the TP condition (all *p*s > .25). Thus, the chicks in Experiment 1 showed no sensitivity to the TPs between shapes across the test phase, despite showing robust recognition of the individual shapes.

Discussion

Despite showing a strong ability to distinguish novel shapes from familiar shapes, the chicks failed to encode TPs between shapes. Because our method was able to detect robust shape recognition in chicks, the chicks' failure to encode TPs between shapes cannot be explained by problems with the method. Indeed, we used an automated method that eliminates experimenter bias and reduces measurement error, two essential steps for improving the power and precision of empirical studies.

There were, however, a number of methodological differences between the Santolin et al. (2016) study and

the present study, and it is possible that these differences caused the divergent findings across studies. To reconcile these differences, we performed a second experiment in which we substantially reduced the differences between the studies.

Experiment 2

Method

The methods in Experiment 2 were identical to those used in Experiment 1, except in the following ways. First, as did Santolin et al. (2016), we started testing newborn chicks on Day 1 after hatching. By testing for statistical learning on Day 1, Experiment 2 allowed us to detect early emerging statistical-learning abilities. Second, as did Santolin et al., we exposed the chicks to the imprinting sequence for 120 min before initiating the first test trial. By exposing the chicks to the imprinting sequence for 120 min, Experiment 2 allowed us to detect statistical-learning abilities that may emerge from only small amounts of experience. Third, as did Santolin et al., we used shorter test trials, reducing the length of the test trials from 40 min to 20 min. This allowed us to collect trials that were closer in duration to those used in the Santolin et al. study, while also collecting more data per subject to reduce measurement error. Fourth, as did Santolin et al., we reared the chicks in darkness for 30 min after they were exposed to the imprinting sequence. Experiment 2 therefore allowed us to detect statistical-learning abilities that may emerge only when chicks are maintained in a dark environment after exposure to an imprinting sequence. Fifth, as did Santolin et al., we limited the design to a single type of test trial (imprinting sequence vs. random sequence; cf. Bulf et al., 2011). In Experiment 1, we had four types of test trials, but it is possible that this might have somehow contaminated the results. Experiment 2 addressed this concern directly.

We presented the experimental cycle (120-min imprinting sequence, 30-min darkness period, 20-min test trial) repeatedly, 8 times per day, for 10 days. By repeating the experimental cycle, we could (a) attempt to replicate the Santolin et al. (2016) study in the first experimental cycle and (b) continue to look for evidence of statistical learning across the first 10 days of life. We preregistered the design and analyses for this experiment. The preregistration can be found at osf.io/cmxya/.

Results

Preregistered analyses. The results are shown in Figure 4. The chicks' overall performance was not significantly different from chance level, $M = 49.692\%$, $SE = 0.467\%$, 95% CI = [48.726%, 50.657%], $t(23) = 0.661$, $p = .515$, Cohen's $d = 0.135$, BF in favor of the null hypothesis = 5.169 (see Fig. 4a). We also performed one-sample t tests to determine whether any individual chicks succeeded at the task (see Fig. 4b). To compute a t test for an individual chick, we calculated the percentage of time the chick spent in proximity to the TP-based sequence versus the random-order sequence for each trial throughout the experiment. No chicks performed significantly above or below chance level (all ps > .10). The average effect size (Cohen's d) across individual subjects was –0.013, 95% CI = [–0.045, 0.020].

We also tested whether performance changed across the test days, as it is theoretically possible that chicks could prefer novel stimuli in the first 1 to 2 days of life but eventually imprint to both stimuli, thus preferring both equally in later test days. Contrary to this hypothesis, our results showed that performance did not differ significantly across the test days (repeated measures ANOVA), $F(9, 207) = 1.215$, $p = .287$, $\eta_p^2 = .050$ (see Fig. 4c).

Post hoc analyses. The results of our test-day analysis (ANOVA) did not support the hypothesis that chicks encode and detect TPs in the first days of life but then lose interest in TP information during later days of life. To provide a more direct test of this hypothesis, we analyzed whether the chicks performed above chance level on any of the test days. Performance was not significantly different from chance level during any test day (all ps > .05; see Fig. 4c), including Day 1, $M = 51.759\%$, $SE = 1.704\%$, 95% CI = [48.235%, 55.283%], $t(23) = 1.032$, $p = .313$, Cohen's $d = 0.211$, BF in favor of the null hypothesis = 3.847, or Day 2, $M = 51.881\%$, $SE = 2.457\%$, 95% CI = [46.799%, 56.964%], $t(23) = 0.766$, $p = .452$, Cohen's $d = 0.156$, BF in favor of the null hypothesis = 4.816.

Because Santolin et al. (2016) collected a single test trial per chick, we also explored the possibility that sensitivity to TP information is present on the first test trial but no longer exists in subsequent test trials. Performance was not significantly different from chance level on the first test trial, $M = 52.607\%$, $SE = 11.110\%$, 95% CI = [29.432%, 75.783%], $t(20) = 0.235$, $p = .817$, $d = 0.051$, BF in favor of the null hypothesis = 5.838.

Discussion

In Experiment 2, we found no evidence for statistical learning, either on the group level or on the individual-subject level. We also found no evidence for statistical learning during Days 1 to 2 of life, contradicting the results from the Santolin et al. (2016) study. During the entirety of the 10-day experiment, the chicks' environment contained the imprinting sequence 100% of the time (other than the periods of darkness). Thus, the

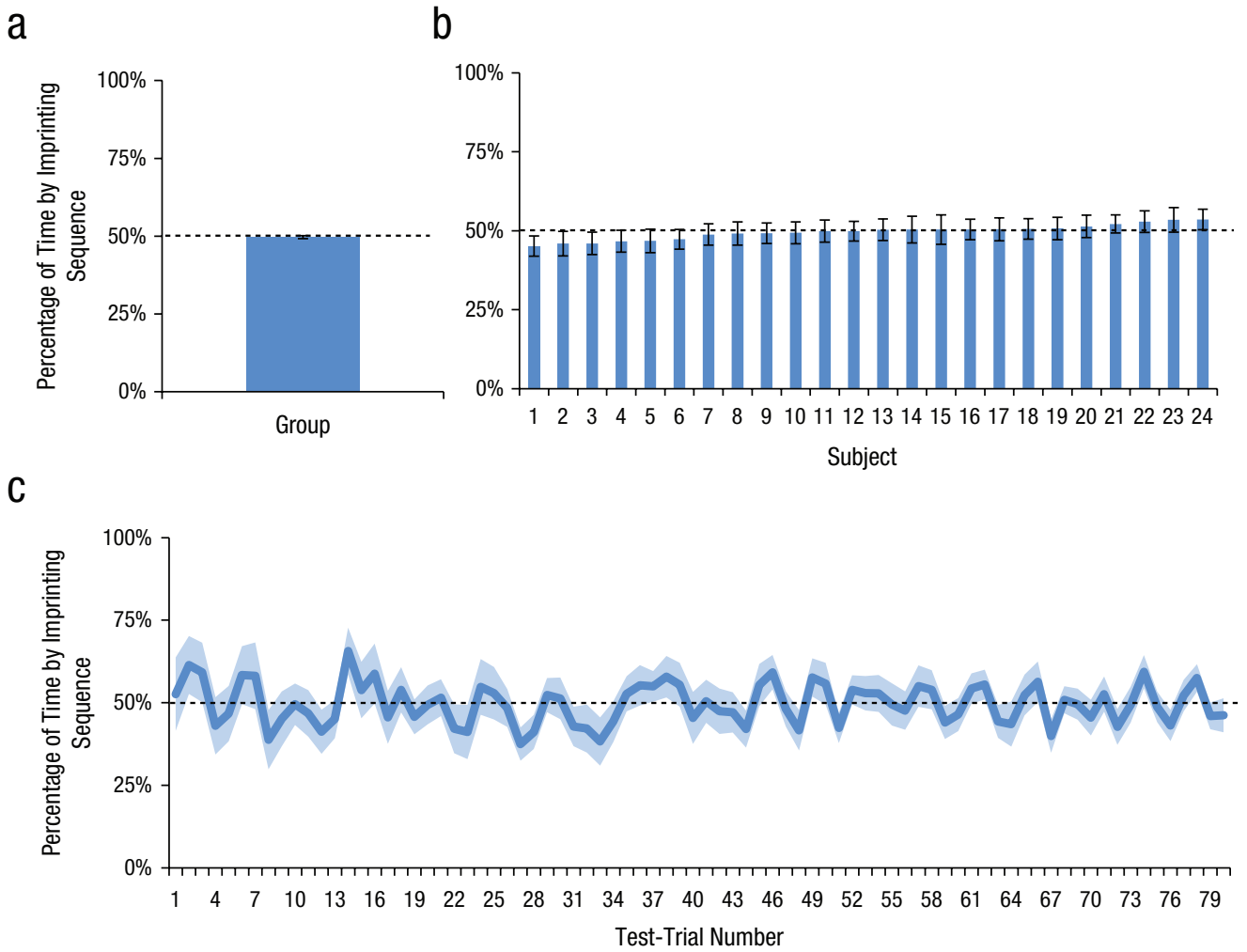

Fig. 4. Performance in Experiment 2. The graphs show (a) group performance (percentage of time chicks spent with the imprinting sequence vs. a novel sequence), (b) individual subject performance, and (c) group performance by test-trial number. Chance performance was 50% (dashed line). Error bars in (a) and (b) and the colored ribbon in (c) denote ±1 *SE*.

chicks had ample opportunity to learn the statistical structure of the sequence. In agreement with Experiment 1, Experiment 2 indicates that newborn chicks do not encode TPs between sequentially presented shapes.

General Discussion

Across two experiments, we attempted to reproduce the findings from the Santolin et al. (2016) study showing that newborn chicks can encode the TPs between sequentially presented shapes. Our experiments revealed no evidence for TP-based statistical learning in chicks.

In Experiment 1, the chicks failed to learn the statistical structure of the sequence despite acquiring 1 week of experience with that sequence before testing. A critic might argue that we failed to find evidence for TP-based statistical learning in this experiment because the chicks received too much time with the imprinting sequence or because the chicks were tested for prolonged periods of time. However, studies with humans have shown that TP-based learning gets stronger with greater exposure to a sequence (Pena, Bonatti, Nespor, & Mehler, 2002), and many studies have shown that testing chicks for prolonged periods yields robust effects in visual-discrimination tasks, including object recognition (J. N. Wood, 2013), face recognition (S. M. W. Wood & Wood, 2015), and action recognition (Goldman & Wood, 2015). Moreover, prolonged testing increased the effect size in the shape-recognition condition by threefold to fourfold, while also producing a threefold to fourfold decrease in measurement error (see Fig. 5). Thus, our method produced far more precise measurements than Santolin et al.'s (2016) did.

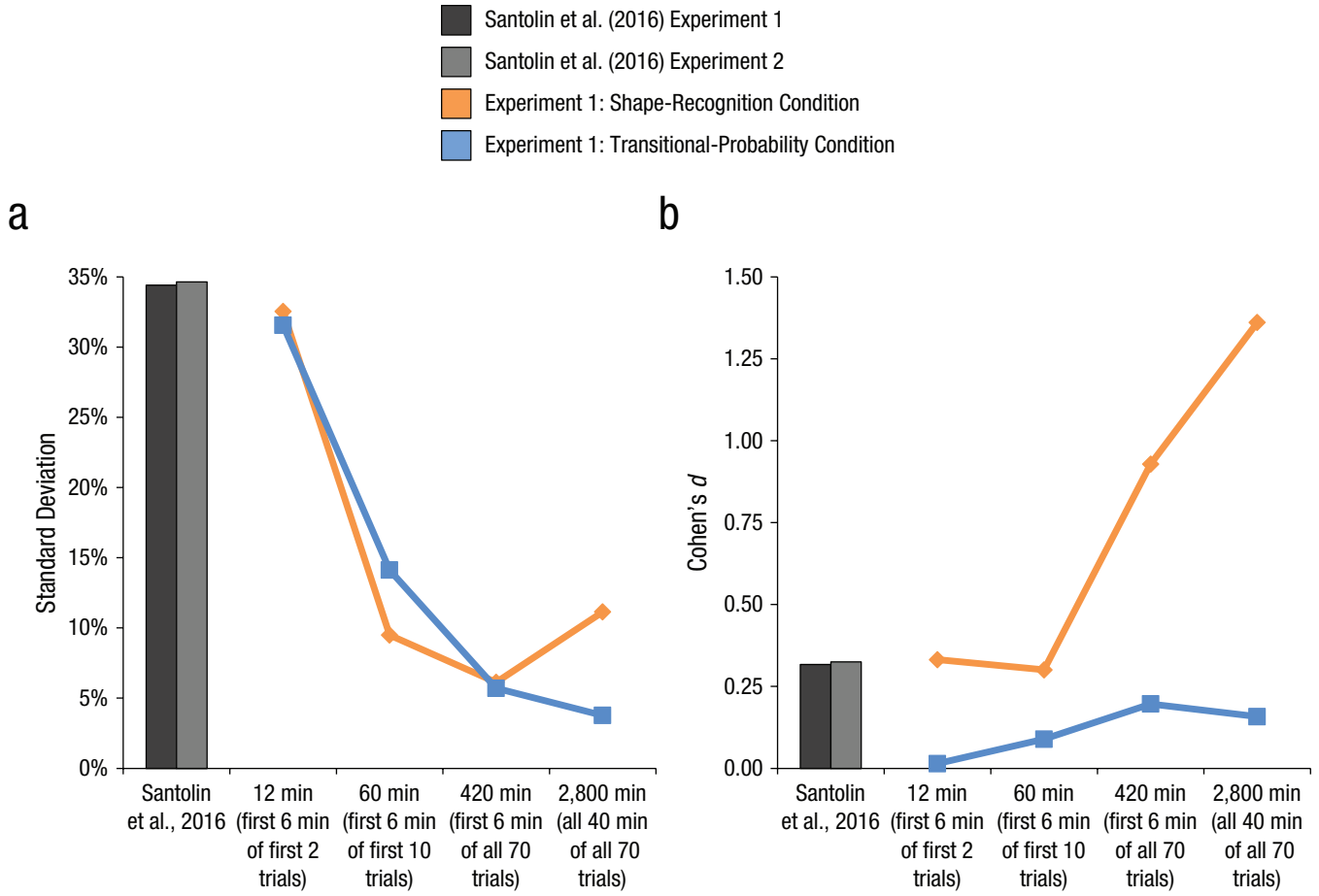

Fig. 5. A comparison of the measurement error and effect size obtained in the present study and the original Santolin, Rosa-Salva, Vallortigara, and Regolin (2016) study. The gray bars show the (a) measurement error (standard deviation) and (b) effect size (Cohen's *d*) from the two experiments by Santolin et al. The orange and blue lines show the measurement error and effect size from the shape-recognition and transitional-probability conditions in the present study.

To directly examine whether methodological differences led to the divergent findings across studies, we performed a second experiment in which we minimized these differences. As in the Santolin et al. (2016) study, the chicks were exposed to the imprinting sequence for 120 min, reared in darkness for 30 min, and then presented with a short test trial. We then repeated this experimental cycle 8 times per day, for 10 days, to explore whether TP-based statistical learning emerges during early or late stages of development. We found no evidence for TP-based statistical learning during any stage of the experiment.

The Santolin et al. (2016) study had noisy measurements, small effect sizes, and high analytic flexibility—three factors that increase false-positive rates. Our study overcame these problems by using a preregistered design (which limits analytic flexibility) and automation (producing large amounts of precise behavioral data from each chick). Given that our measurements were 3 to 4 times more precise than those in the Santolin et al. study, our findings cast doubt on claims that newborn chicks use TP-based statistical learning.

Importantly, our results do not rule out the possibility that newborn chicks use other types of statistical learning to build visual representations. For example, recent automated controlled-rearing studies indicated that newborn chicks require slow and smooth visual-object input to learn to recognize objects (J. N. Wood, 2016; J. N. Wood, Prasad, Goldman, & Wood, 2016; J. N. Wood & Wood, 2016, 2018). These results accord with unsupervised temporal-learning models in computational neuroscience that involve a type of statistical learning in which the brain encodes slow and smooth signals from the environment to build up accurate visual representations of the world (DiCarlo et al., 2012; Foldiak, 1991; Rolls, 2012; Stone, 1996; Wiskott & Sejnowski, 2002). We hypothesize that newborn chicks are capable of this type of statistical learning (encoding slow and smooth visual signals) but not other types of statistical learning (encoding TPs between objects).

Detecting slow and smooth signals requires encoding co-occurrences of retinal inputs (which can be implemented with models using spike-timing-dependent plasticity-learning rules; Sprekeler, Michaelis, & Wiskott, 2007). Conversely, detecting TPs is more complex because TPs entail predictive relations among items across longer time windows rather than integrating input signals over the span of milliseconds. Thus, TPs may rely on working memory and higher-level processing mechanisms that encode information at longer time scales (Kiebel, Daunizeau, & Friston, 2008).

More generally, newborn chicks' failure to encode TPs is consistent with previous automated controlled-rearing experiments showing that chicks fail to encode the order of images in a sequence, despite encoding the individual images (J. N. Wood et al., 2016). These results are also consistent with prior studies showing that although human infants encode some statistical relations at birth (e.g., discrimination of structured from random sequences; Bulf et al., 2011), they fail to encode TPs until 5 months of age (Marcovitch & Lewkowicz, 2009; Slone & Johnson, 2015). Thus, it may not be necessary to postulate differences in statistical-learning abilities across species to explain why newborn chicks succeed in TP-based statistical learning whereas newborn humans fail (Santolin & Saffran, 2018). Both newborn chicks and newborn humans appear to be incapable of TP-based statistical learning, suggesting that this ability requires extended maturation or experience (or both) in order to develop.

In sum, a core goal in psychology is to understand the learning mechanisms in newborn brains. Our study indicates that TP-based statistical learning is not one of those mechanisms (at least in the visual domain for chicks). TP-based learning may emerge later in development, as with humans, or may never emerge in some species. Of course, it is also possible that something about our method prevents TP-based statistical learning while enabling object learning, and other methods could provide robust, replicable evidence for this ability in chicks. If so, our results would still indicate that there are strong constraints on the types of situations that can successfully elicit TP-based statistical learning. Understanding the development of TP-based statistical learning—and its distribution across the animal kingdom—is an important avenue for future research.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

All the authors designed the research. S. M. W. Wood and J. N. Wood ran the experiments. S. M. W. Wood analyzed the data. All the authors wrote the manuscript and approved the final manuscript for submission.

ORCID iD

Samantha M. W. Wood https://orcid.org/0000-0002-2219-0285

Acknowledgments

This work was conducted at the University of Southern California.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by National Science Foundation CAREER Grant BCS-1351892 (to J. N. Wood), a James S. McDonnell Foundation Understanding Human Cognition Scholar Award (to J. N. Wood), and National Institutes of Health Grant R01-HD073535 (to S. P. Johnson).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797619868998

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at osf.io/uqmh4/. The direct links are osf.io/szvfp (Experiment 1 data), osf.io/5gd9m (Experiment 2 data), osf.io/6sxat (Experiment 1 materials), and osf.io/a396j (Experiment 2 materials). The design and analysis plans for Experiment 2 were preregistered at osf.io/cmxya. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797619868998. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. One-sample Cohen's *d* was calculated as follows: absolute value (average performance – chance performance)/standard deviation = absolute value [37% – 50%]/10% = 1.3.
2. Santolin et al. (2016) changed the objects' color across the imprinting and test phases, whereas we held object color constant. We introduced this change because in the vast majority of studies on visual statistical learning, object color is held constant across the learning and test phases. In terms of replication, holding object color constant should have made it easier for the chicks to recognize the familiar TP sequence because the chicks did not need to generalize across a color change.

3. For all BFs, we used the scaled Jeffreys-Zellner-Siow BF with scale (r) of 1.0 (Rouder, Speckman, Sun, Morey, & Iverson, 2009). For reference, BFs between 0 and 3 are often considered anecdotal evidence, BFs between 3 and 10 are considered moderate evidence, and BFs greater than 10 are considered strong evidence.

References

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*, 127–132.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*, 415–434.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194–200.

Goldman, J. G., & Wood, J. N. (2015). An automated controlled-rearing method for studying the origins of movement recognition in newly hatched chicks. *Animal Cognition*, *18*, 723–731.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLOS Computational Biology*, *4*(11), Article e1000209. doi:10.1371/journal.pcbi.1000209

Krogh, L., Vlach, H. A., & Johnson, S. P. (2012). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology*, *3*, Article 598. doi:10.3389/fpsyg.2012.00598

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585.

Marcovitch, S., & Lewkowicz, D. J. (2009). Sequence learning in infancy: The independent contributions of conditional probability and pair frequency information. *Developmental Science*, *12*, 1020–1025.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, Article 0021. doi:10.1038/s41562-016-0021

Pena, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*, 604–607.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rolls, E. T. (2012). Invariant visual object and face recognition: Neural and computational bases, and a model, VisNet. *Frontiers in Computational Neuroscience*, *6*, Article 35. doi:10.3389/fncom.2012.00035

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Santolin, C., Rosa-Salva, O., Vallortigara, G., & Regolin, L. (2016). Unsupervised statistical learning in newly hatched chicks. *Current Biology*, *26*, R1218–R1220.

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*, 52–63.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Slone, L. K., & Johnson, S. P. (2015). Infants' statistical learning: 2- and 5-month-olds' segmentation of continuous visual sequences. *Journal of Experimental Child Psychology*, *133*, 47–56.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences*, *18*, 251–258.

Sprekeler, H., Michaelis, C., & Wiskott, L. (2007). Slowness: An objective for spike-timing-dependent plasticity? *PLOS Computational Biology*, *3*, 1136–1148.

Stone, J. V. (1996). Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, *8*, 1463–1492.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*, 715–770.

Wood, J. N. (2013). Newborn chickens generate invariant object representations at the onset of visual object experience. *Proceedings of the National Academy of Sciences, USA*, *110*, 14000–14005.

Wood, J. N. (2014). Newly hatched chicks solve the visual binding problem. *Psychological Science*, *25*, 1475–1481.

Wood, J. N. (2016). A smoothness constraint on the development of object recognition. *Cognition*, *153*, 140–145.

Wood, J. N., Prasad, A., Goldman, J. G., & Wood, S. M. W. (2016). Enhanced learning of natural visual sequences in newborn chicks. *Animal Cognition*, *19*, 835–845.

Wood, J. N., & Wood, S. M. W. (2016). The development of newborn object recognition in fast and slow visual worlds. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1829), Article 20160166. doi:10.1098/rspb.2016.0166

Wood, J. N., & Wood, S. M. W. (2018). The development of invariant object recognition requires visual experience with temporally smooth objects. *Cognitive Science*, *42*, 1391–1406.

Wood, S. M. W., & Wood, J. N. (2015). Face recognition in newly hatched chicks at the onset of vision. *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*, 206–215.

Wood, S. M. W., & Wood, J. N. (2019). Using automation to combat the replication crisis: A case study from controlled-rearing studies of newborn chicks. *Infant Behavior & Development*, *57*, Article 101329. doi:10.1016/j.infbeh.2019.101329