Infant Behavior and Development 57 (2019) 101329

Contents lists available at ScienceDirect

Infant Behavior and Development

journal homepage: www.elsevier.com/locate/inbede

Using automation to combat the replication crisis: A case study from controlled-rearing studies of newborn chicks

Samantha M.W. Wood*, Justin N. Wood

Indiana University School of Informatics, Computing, and Engineering, 700 N Woodlawn Ave, Bloomington, IN 47408, United States

ARTICLE INFO

Keywords:
Replication crisis
Automation
Controlled rearing
Measurement error
Chick

ABSTRACT

The accuracy of science depends on the precision of its methods. When fields produce precise measurements, the scientific method can generate remarkable gains in knowledge. When fields produce noisy measurements, however, the scientific method is not guaranteed to work – in fact, noisy measurements are now regarded as a leading cause of the replication crisis in psychology. Scientists should therefore strive to improve the precision of their methods, especially in fields with noisy measurements. Here, we show that automation can reduce measurement error by ~60% in one domain of developmental psychology: controlled-rearing studies of newborn chicks. Automated studies produce measurements that are 3–4 times more precise than non-automated studies and produce effect sizes that are 3–4 times larger than non-automated studies. Automation also eliminates experimenter bias and allows replications to be performed quickly and easily. We suggest that automation can be a powerful tool for improving measurement precision, producing high powered experiments, and combating the replication crisis.

1. Introduction

The scientific method has produced astonishing gains in human knowledge. In the past decade, however, it has become clear that the scientific method does not always work, producing large bodies of unreproducible findings (Ioannidis, 2005). In social psychology, for example, rigorous replication studies have failed to replicate many published findings, including textbook phenomena such as ego depletion (Hagger et al., 2016) and social priming (Klein et al., 2014). Likewise, in a large-scale replication study of 100 experiments from three high-ranking psychology journals, only one-third to one-half of the original findings were observed in the replication studies (Open Science Collaboration, 2015). Thus, some findings in psychology are reproducible, whereas others are not. What distinguishes reproducible from non-reproducible findings, and how can we improve our methods to produce more reproducible results?

The field of metascience – the scientific study of science itself – has identified a number of factors contributing to the replication crisis. The most prominent factors are underpowered studies (low sample size relative to effect size), high measurement error (noisy measurements), small effect sizes, experimenter bias, analytic flexibility (also known "P-hacking" and "the garden of forking paths"), and conflicts of interest (Ioannidis, 2005; Loken & Gelman, 2017; Munafò et al., 2017; Simmons, Nelson, & Simonsohn, 2011). Producing accurate and reproducible research therefore requires combating these factors.

Here, we present a case study of how automation can be used to address many of these factors contributing to the replication crisis. Specifically, we focus on controlled-rearing studies of newborn chicks – an area with deep historical roots in developmental

* Corresponding author.
E-mail addresses: samantha.m.w.wood@gmail.com (S.M.W. Wood), justin.n.wood@gmail.com (J.N. Wood).

https://doi.org/10.1016/j.infbeh.2019.101329
Received 6 May 2018; Received in revised form 18 January 2019; Accepted 1 May 2019

psychology. Our paper is organized as follows. First, we briefly describe controlled-rearing studies and their importance for developmental psychology. Second, we sample a range of non-automated controlled-rearing studies with newborn chicks and find that these studies tended to produce noisy measurements and small effect sizes – two leading factors in the replication crisis. Third, we show how automation can solve this problem, by allowing newborn chicks to be observed and tested for long periods of time. We found that increasing the amount of data collected from each chick (via automated tracking software) reduced measurement error by 60% and increased the effect size of experiments by about 300%. Automation also eliminates experimenter bias and allows replications to be performed quickly and easily. We conclude that automation can significantly improve the power and accuracy of controlled-rearing studies with newborn animals.

2. Using controlled rearing to explore the origins of intelligence

Controlled-rearing studies are designed to explore the role of experience in shaping perception, cognition, and behavior. Despite widespread interest in this topic, there is little agreement on fundamental issues, including which mechanisms exist in newborn brains and how those mechanisms transform sensory input into knowledge. On one view, newborn animals are equipped with a collection of domain-specific mechanisms, each shaped by evolution to perform a particular function. Some researchers argue that newborns have numerous domain-specific systems (e.g., Cosmides & Tooby, 1994Steven, 2003), while others argue that newborns have a smaller number of specialized core knowledge systems (i.e., systems for representing objects, actions, number, and space; Carey, 2009; Spelke & Kinzler, 2007). According to this class of “nativist” theories, newborns have domain-specific learning mechanisms at the onset of post-natal experience.

On the second view, newborn brains contain flexible and adaptive domain-general learning mechanisms for building domain-specific knowledge through experience (Karmiloff-Smith, 1995; Quartz & Sejnowski, 1997; Thelen and Smith, 1994). This view originally traces back to Enlightenment thinkers such as Locke (Locke, 1689) and Hume (Hume, 1748). More recent accounts from computational neuroscience (e.g., Hassabis & Kumaran, 2017; Hinton, 1992; McClelland & Rumelhart, 1985) argue that the representational abilities of the brain are built from the dynamic interaction between neural growth mechanisms and environmentally produced neural activity. This class of “empiricist” theories predict that domain-specific systems emerge from domain-general learning mechanisms as newborns acquire experience with the world.

Modern psychologists generally accept that newborn brains are equipped with both domain-specific and domain-general learning mechanisms. Nonetheless, there is still little consensus on fundamental issues: researchers disagree about what these mechanisms are, what they consist of, what causes them to emerge, and how they change with experience. How can we advance this debate and characterize the learning mechanisms in newborn brains? As scientists have long noted, human infants are not an optimal population for addressing these questions. When infants participate in experiments, they have already been shaped by days, weeks, or months of experience with the natural visual world. Consequently, for almost any ability found in young human infants, it is not possible to determine whether that ability is hardwired by natural selection or learned from experience. Indeed, there is growing evidence from neuroscience that natural sensory experience plays an important role in building domain-specific knowledge (e.g., for faces and symbols: Arcaro, Schade, Vincent, Ponce, & Livingstone, 2017; Srihasam, Vincent, & Livingstone, 2014). Studies of monkeys also show that brains change rapidly in response to statistical redundancies in the environment, with significant neuronal rewiring occurring in as little as 1 h (Li & DiCarlo, 2008, 2010). These findings allow for the possibility that even early emerging domain-specific abilities are learned from post-natal experience.

In contrast to studies of human infants, controlled-rearing studies of nonhuman animals can reveal the precise role of post-natal experience on development (Held & Hein, 1963; Hubel & Wiesel, 1962; Walk, Gibson, & Tighe, 1957). By systematically manipulating the experiences provided to newborn subjects and observing the effects of those manipulations on behavior, we can distinguish the experiences that are causally related to developmental change from those that are not. Controlled-rearing experiments therefore provide an experimental avenue for probing the role of experience in the development of mental abilities.

For example, recent automated controlled-rearing studies have started to reveal the role of visual experience in building object concepts. These studies indicate that newborn animals (domestic chicks) can begin building abstract (view-invariant) object concepts at the onset of vision (Wood, 2013; Wood & Wood, 2015). From an engineering perspective, this is an impressive computational feat: computer vision systems typically require thousands to millions of diverse training images to build abstract object concepts. However, this ability only emerges when newborn chicks are raised with objects that move slowly and smoothly over time (Wood & Wood, 2016, 2018). When objects move too quickly or non-smoothly, newborn chicks build inaccurate object concepts. Thus, to learn how to perceive objects, chicks need experience with objects that adhere to the spatiotemporal properties of objects in the real world, akin to human children (Smith & Slone, 2017). From a mechanistic perspective, these results indicate that newborn brains contain unsupervised temporal learning mechanisms, which leverage the spatiotemporal characteristics of natural visual environments to build object concepts. This machinery has been sculpted by evolution to generate abstract object concepts, but only when the brain receives input from a natural visual world.

2.1. Newborn chicks as a model system for studying perceptual and cognitive development

Newborn chicks (*Gallus gallus*) are uniquely suited for studying the development of perception and cognition. Unlike commonly-used animal models in psychology (e.g., rats, pigeons, and monkeys), chickens are a precocial species (mobile in the first day of life) and can be raised in strictly controlled environments immediately after hatching. Newborn chicks can be raised for weeks in environments containing no real-world objects or agents (Wood, 2013). Thus, studies of newborn chicks allow us to test critical hypotheses in the nativist–empiricist debate. With chicks, we can discover mechanisms that are present in newborn brains, and explore how those mechanisms are shaped by experience. While mechanisms found in chicks may or may not have also evolved in humans (the last common ancestor of humans and chickens lived around 310 million years ago), these studies provide an existence proof of the types of mechanisms that are present in newborn brains.

Moreover, recent work from avian neuroscience indicates that cortical mechanisms are largely conserved across birds and mammals (reviewed by Jarvis et al., 2005; Karten, 2013; Shanahan, Bingman, Shimizu, Wild, & Gunturkun, 2013). At the circuit-level, avian and mammalian brains contain homologous cortical circuits for processing sensory information; these circuits share similarities in terms of cell morphology, the connectivity pattern of the input and output neurons, gene expression, and function. At the macro-level, avian and mammalian brains share the same large-scale organizational principles: both are modular, small-world networks with a connective core of hub nodes that includes prefrontal-like and hippocampal structures. Since cortical mechanisms are shared across birds and mammals, studies of chicks can reveal general characteristics of vertebrate brain development. While chicks have smaller brains than humans, this can provide an advantage to researchers. When attempting to understand a particular phenomenon, it is often valuable to use the simplest system that demonstrates the properties of interest. The fields of neuroscience and biology rely heavily on this strategy – for example, researchers use sea slugs to study the physiological basis of memory storage in neurons (Kandel, 2007), worms to study the mechanisms of molecular and developmental biology (Brenner, 1974), and fruit flies to study the mechanisms of genetics (Bellen, Tong, & Tsuda, 2010). In a similar vein, controlled-rearing studies of newborn chicks can provide an important (and unique) perspective on perceptual and cognitive development. Next, we briefly describe non-automated controlled-rearing studies with newborn chicks, with a focus on studies that had a significant influence on theories of cognitive development.

3. Non-automated studies of newborn chicks

Controlled-rearing studies of newborn chicks have deep historical roots in developmental psychology (e.g., Gibson & Walk, 1960; Vallortigara, 2012). The majority of these studies were designed to explore whether chicks have domain-specific knowledge in the absence of post-natal experience with the domain.[1] In support of this premise, researchers have reported that newborn chicks have early emerging preferences for faces (Rosa-Salva, Regolin, & Vallortigara, 2010), biological motion (Vallortigara, Regolin, & Marconato, 2005), and self-propelled motion (Mascalzoni, Regolin, & Vallortigara, 2010), all without prior experience with faces, biological motion, and self-propelled motion. These findings have been interpreted as evidence for an innate (non-learned) mechanism for detecting animate agents. There have also been reports that newborn chicks have early emerging capacities for object cognition (Regolin & Vallortigara, 1995) and numerical cognition (Rugani, Regolin, & Vallortigara, 2010), and that newborn ducklings can perform abstract relational reasoning (Martinho & Kacelnik, 2016). Together, these studies have been interpreted as strong evidence for nativist theories of cognitive development (Carey, 2009; Spelke & Kinzler, 2007; Vallortigara, 2012). Since there was little to no opportunity for the animals in these studies to learn about objects, agents, and number prior to testing, empiricist theories of cognitive development cannot readily explain these findings. Given the importance of these findings, it seems vital to assess the precision of the measurements obtained in the studies. In all of the studies cited above, the chicks were tested in a two-alternative forced-choice task, within a single session lasting a short period of time (generally ~6 min). This is a short measurement period, so these studies tended to produce noisy measurements. In a sample of 10 non-automated controlled-rearing studies (Table 1), we found that the average standard deviation (measurement error) was 33%, ranging from 17% to 65% across studies (Fig. 1A). To illustrate how noisy of a measurement this is, Fig. 2 shows visualizations of distributions from noisy measurements (SD = 33%) versus precise measurements (SD = 10%). Note that for the noisy measurements, a large proportion of the data falls both above and below chance levels, hindering assessment of true population performance.

Furthermore, researchers often conclude that chicks succeeded at a task irrespective of whether the chicks performed above or below chance levels (i.e., both familiarity and novelty effects are considered evidence for an ability), providing increased flexibility to find an effect. Unfortunately, recent work from metascience shows that this combination of factors (noisy measurements and flexibility in the direction of the effect) can easily lead to the perception of false effects (Loken & Gelman, 2017; Munafò et al., 2017; Simmons et al., 2011).

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 positive rates, especially when researchers have flexibility in terms of the number of subjects that are tested, the analyses that are performed, and the results that are reported (Simmons et al., 2011). Inflated estimates of effect size lead researchers to underestimate the number of subjects needed for appropriate statistical power, resulting in underpowered studies.[2]

A critic might argue that if studies produce statistically significant results despite having noisy measurements, then the results

[1] These studies generally do not assume a special ecological niche that would differentiate domain-specific knowledge in chicks from domain-specific knowledge in other species. However, it is worth noting that chickens are a precocial species and may thus differ in some ways from humans. For example, chicks can begin exploring their environment at the onset of post-natal experience. This active motor exploration might play an important role in the development of perceptual and cognitive abilities.

[2] Effect sizes measure the magnitude of an effect relative to the noise (e.g., for a one-sample t-test, Cohen's d = (mean − μ)/SD). Because the denominator to compute d is standard deviation (SD), methods that produce larger standard deviations will have smaller effect sizes. Critically, effect sizes and power are inextricably linked: smaller effect sizes require larger numbers of subjects to attain the same level of power (see Fig. 1C).

Table 1
Measurement error (standard deviation) and effect size (Cohen's *d*) for all of the statistically significant conditions from 10 automated and 10 non-automated studies. Each point in Fig. 1 corresponds to one entry in the table.

Authors	Year	Title	Journal	Experimental condition	SD	Cohen's *d*
Non-automated studies						
Regolin & Vallortigara	1995	Perception of partly occluded objects by young chicks	Perception & Psychophysics	Exp 1: Complete triangle vs. fragmented triangle	22%	0.38
				Exp 1: Complete triangle vs. scrambled triangle	34%	0.36
				Exp 1: Partly occluded triangle vs. fragmented triangle	29%	0.40
				Exp 1: Partly occluded triangle vs. inversion of occlusion	26%	0.43
				Exp 2: Reared with occluded object, preference for complete triangle vs. Fragmented triangle	26%	0.55
				Exp 2: Reared with fragmented triangle, preference for fragmented triangle	21%	0.74
Vallortigara, Regolin, & Marconato	2005	Visually inexperienced chicks exhibit spontaneous preference for biological motion patterns	PLOS Biology	walking hen vs. rigid motion	27%	0.38
				walking hen vs. random motion	26%	0.38
				scrambled hen vs. rigid motion	28%	0.18
				scrambled hen vs. random motion	25%	0.43
				walking cat vs. rigid motion	28%	0.28
				walking cat vs. random motion	27%	0.27
Mascalzoni, Regolin, & Vallortigara	2010	Innate sensitivity for self-propelled causal agency in newly hatched chicks	PNAS	Exp 1	33%	0.32
Rosa-Salva, Regolin, & Vallortigara	2010	Faces are special for newly hatched chicks: evidence for inborn domain-specific mechanisms underlying spontaneous preferences for face-like stimuli	Developmental Science	Exp 1: Imprinted chicks	38%	0.61
				Exp 1: Non-imprinted, 1st 3 min only	44%	0.25
				Exp 4	38%	0.28
Rugani, Regolin, & Vallortigara	2010	Imprinted numbers: newborn chicks’ sensitivity to number vs. continuous extent of objects they have been reared with	Developmental Science	Exp 1: Imprinted on 3	24%	0.28
				Exp 1: Imprinted on 1	32%	0.22
				Exp 2: Imprinted on 2 objects	32%	0.35
				Exp 2: Imprinted on 3 objects	31%	0.30
				Exp 3: Imprinted on one big object	32%	0.44
				Exp 3: Imprinted on many objects	30%	0.44

(*continued on next page*)

Table 1 (*continued*)

Authors	Year	Title	Journal	Experimental condition	SD	Cohen's *d*
Non-automated studies						
				Exp 4: Imprinted on heterogeneous sets	33%	0.26
Chiandetti & Vallortigara	2011	Chicks like consonant music	Psychological Science	Min 5 of testing	65%	0.24
				Min 6 of testing	60%	0.32
Regolin, Rugani, Stancher, & Vallortigara	2011	Spontaneous discrimination of possible and impossible objects by newly hatched chicks	Biology Letters	Occlusion condition	29%	0.45
				No occlusion	22%	0.37
				Naïve	17%	0.48
Martinho & Kacelnik	2016	Ducklings imprint on the relational concept of “same or different”	Science	Shape condition	32%	0.39
				Colour condition	36%	0.19
Rosa-Salva, Grassi, Lorenzi, Regolin, & Vallortigara	2016	Spontaneous preference for visual cues of animacy in naïve domestic chicks: The case of speed changes	Cognition	Exp 2	42%	0.62
				Exp 3	46%	0.44
				Exp 4	43%	0.36
				Exp 6	46%	0.29
Santolin, Rosa-Salva, Vallortigara, & Regolin	2016	Unsupervised statistical learning in newly hatched chicks	Current Biology	Exp 1	34%	0.35
				Exp 2	35%	0.35

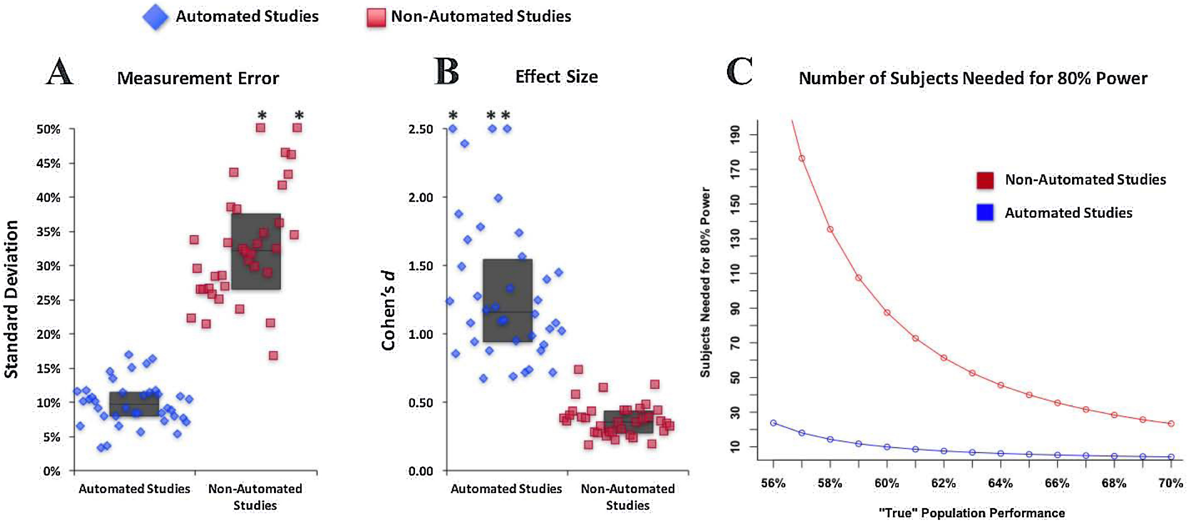

Fig. 1. Scatterplots and boxplots of the (A) measurement error and (B) effect sizes from samples of automated (blue points) and non-automated (red points) controlled-rearing studies. Each point represents the (A) standard deviation or (B) Cohen's *d* from a single condition (only statistically significant conditions are included). The boxplots show the range from 25th to 50th percentile and from 50th to 75th percentile. Points with an asterisk are beyond the range of the graph (greater than 50% standard deviation or greater than 2.5 Cohen's *d*). Across a wide range of studies, the effect sizes obtained with automated methods were much larger than the effect sizes obtained with non-automated methods. Automated methods also produced far more precise measurements than non-automated methods. (C) The number of subjects needed to achieve 80% power for a range of true population performance values for studies with low measurement error (standard deviation = 10%) and high measurement error (standard deviation = 33%). These standard deviations match those from our samples of automated and non-automated studies, respectively. Low measurement error massively reduces the number of subjects needed to achieve adequate experimental power. When measurement error is high, small decreases in true population performance require large increases in the number of subjects needed to achieve 80% power. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

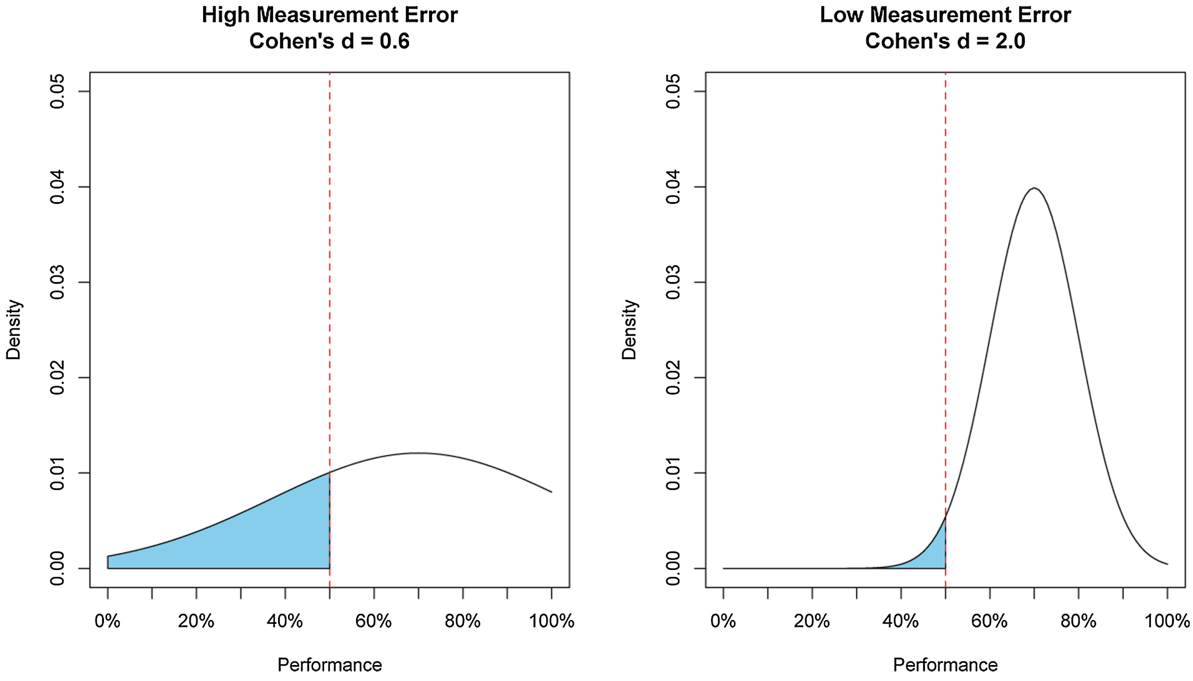

Fig. 2. Visualization of normal distributions of data with high measurement error (standard deviation = 33%) and low measurement error (standard deviation = 10%). These standard deviations match those from our samples of automated and non-automated studies, respectively. Despite having the same mean performance (70%), distributions with low measurement error produce much larger effect sizes than distributions with high measurement error (Cohen's $d = 2.0$ versus $d = 0.6$, respectively). Note that while the low measurement error data has a very narrow distribution, much of the high measurement error data falls both above and below chance level (despite the high average performance).

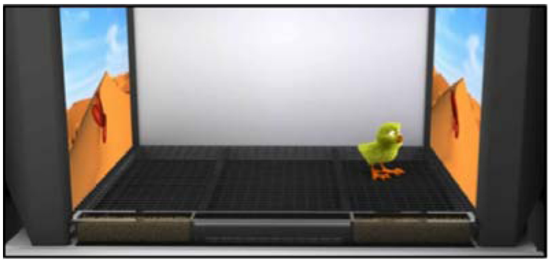

Fig. 3. Illustration of an automated controlled-rearing chamber. The chambers contain no objects other than the virtual objects projected on the display walls. Using microcameras embedded in the ceiling, we record chicks' behavior 24/7 via automated image-based tracking software.

must be particularly robust (because it was possible to detect the effects despite the noise in the signal). However, researchers typically have so many "researcher degrees of freedom" (flexibility in how they prepare, analyze, and report data) that statistical significance can easily be found even in the absence of underlying effects (Simmons et al., 2011), even without multiple hypothesis testing (Gelman & Loken, 2014). This problem is further exacerbated when data are noisy (Loken & Gelman, 2017). In other words, statistical significance conveys very little information when measurements are noisy and researchers have high analytic flexibility (for a detailed discussion see Gelman, 2018).

In sum, non-automated controlled-rearing experiments with newborn chicks tend to produce noisy measurements and small effect sizes. Thus, these studies show two of the factors contributing to the replication crisis (Munafò et al., 2017). When researchers collect a single short trial from subjects and have high analytic flexibility, then the resulting noisy measurements can masquerade as true effects. A method that could produce precise measurements and large effect sizes would therefore improve the accuracy of controlled-rearing studies. In the next section, we describe how automation can be used to produce precise estimates of performance from newborn chicks, resulting in large effect sizes and more powerful experimental designs.

4. Using automation to increase the precision and power of controlled-rearing experiments

Historically, newborn chicks' behavior has been measured through direct observation by trained researchers. This approach has limitations: a researcher can typically observe only a single subject at a time, and these observations are susceptible to experimenter bias. Experimenter bias can alter subjects' behavior and the measurement of that behavior. In other fields-including physics, astronomy, and chemistry-automation has been invaluable for minimizing experimenter bias in research procedures, while also removing sources of noise by standardizing data collection procedures.

To apply these benefits of automation to controlled-rearing research, our lab developed a fully automated controlled-rearing method (Wood, 2013). The method allows newborn chicks to be raised for weeks within automated controlled-rearing chambers. As illustrated in Fig. 3, these chambers provide complete control over all visual object experiences. The chambers contain extended surfaces only, and stimuli are presented to the chicks by projecting virtual objects on two display walls (LCD monitors) situated on opposite sides of the chamber. Thus, the chicks' visual object experiences are limited to the virtual objects projected on the display walls.

All stimuli presentation, data collection, and behavioral tracking are performed by computers, allowing the chicks' behavior to be measured continuously (24/7). The chambers track each chick's location at the rate of 9 samples/sec, using micro-cameras and automated image-based tracking software (EthoVision XT, Noldus Information Technology, Leesburg, VA, USA). Image-based tracking operates over a digital recording of the animal's behavior, which maintains an objective view of events, allows analyses to be repeated, and enables researchers to perform exploratory analyses of variables not originally considered.

This method also produces hundreds of hours of data per chick over their first few weeks of life. Accordingly, we can obtain precise estimates of each chick's performance and determine whether each chick succeeded or failed at the experimental task. Moreover, since the entire data collection process is automated, this method eliminates experimenter bias (e.g., bias that may occur when coding the subject's behavior, presenting stimuli to the subject, or deciding whether to include the subject in the final analysis).

To quantify the benefits of automation, we compared the measurement error and effect sizes from representative samples[3] of automated and non-automated controlled-rearing studies. For the automated studies, the average standard deviation was 10% (Table 2; blue points in Fig. 1A), whereas for the non-automated studies, the average standard deviation was 33% (Table 1; red points in Fig. 1A). Thus, automated studies produce measurements that are 3–4 times more precise than non-automated studies. Next, we analyzed whether automated studies produce larger effect sizes (Cohen's d) than non-automated studies (for reference, $d = 0.20$ is

[3] For our samples, we chose studies that used a two-alternative forced-choice test, sampled behavior during preselected measurement windows, and did not use a training procedure. For the non-automated studies, we selected studies that have had an influence on theories of developmental psychology (e.g., studies arguing for innate capacities for object, agent, and number cognition). For the automated studies, we selected the first 10 automated studies published from our lab with significant results. For both samples, we included all of the conditions reporting statistically significant effects. In the Supplementary Data, we provide an Excel file that includes all of the data and calculations used to produce the measurement error and effect size estimates across the samples.

Table 2
Automated studies. Measurement error (standard deviation) and effect size (Cohen's d) for all of the statistically significant conditions from 10 automated studies. Each blue point in Fig. 1 corresponds to one entry in the table.

Authors	Year	Title	Journal	Experimental condition	SD	Cohen's *d*
Automated studies						
Wood	2013	Newborn chickens generate invariant object representations at the onset of visual object experience	PNAS	Exp 1	12%	1.24
				Exp 2	7%	2.94
				Exp 3	10%	0.86
Wood	2014	Newly hatched chicks solve the visual binding problem	Psychological Science	Color change	12%	1.88
				Shape change	11%	1.49
				Color-shape change	11%	2.39
				Binding change	10%	1.69
Goldman & Wood	2015	An automated controlled-rearing method for studying the origins of movement recognition in newly hatched chicks	Animal Cognition	Exp 1	9%	1.08
				Exp 2: Replace one movement	3%	0.94
				Exp 2: Replace all movements	8%	1.28
				Exp 3	4%	1.78
Wood	2015	Characterizing the information content of a newly hatched chick's first visual object representation	Developmental Science	Identity trials	15%	0.68
				Viewpoint trials	13%	1.17
Wood & Wood	2015	A chicken model for studying the emergence of invariant object recognition	Frontiers in Neural Circuits		8%	0.88
Wood & Wood	2015	Face recognition in newly hatched chicks at the onset of vision	JEP: ALC	Edges only	7%	5.15
				No color	9%	3.71
				Features only	9%	1.99
				Different gender coloring	17%	1.10
				Inverted	15%	1.10
				Different age	12%	1.20
Wood, Prasad, Goldman, & Wood	2016	Enhanced learning of natural visual sequences in newborn chicks	Animal Cognition	Novel image trials, natural sequence condition	9%	1.33
				Novel image trials, unnatural sequence condition	6%	0.69
Wood & Wood	2016	The development of newborn object recognition in fast and slow visual worlds	Proceedings of the Royal Society B	Exp 1: Slow object – view-invariant vs. view-dependent (within-subjects difference)	11%	0.95
				Exp 1: Fast object – view-invariant vs. view-dependent (within-subjects difference)	16%	1.74
Wood	2017	Spontaneous preference for slowly moving objects in visually naïve animals	Open Mind	Exp 1	11%	1.57
				Exp 2: 1.25 s vs. 2.5 s	16%	0.72
				Exp 2: 1.25 s vs. 5 s	12%	0.74
				Exp 2: 1.25 s vs. 10 s	11%	0.98
				Exp 2: 1.25 s vs. 20 s	9%	1.15
Wood & Wood	2017	Measuring the speed of newborn object recognition in controlled visual worlds	Developmental Science	Exp 1: 250 ms	7%	1.24
				Exp 1: 750 ms	9%	0.88
				Exp 1: 1750 ms	9%	0.92
				Exp 1: 3750 ms	8%	1.40
				Exp 2: 125 ms	5%	1.04
				Exp 2: 250 ms	11%	0.72
				Exp 2: 750 ms	8%	1.08
				Exp 2: 1750 ms	7%	1.45
				Exp 2: 3750 ms	10%	1.02

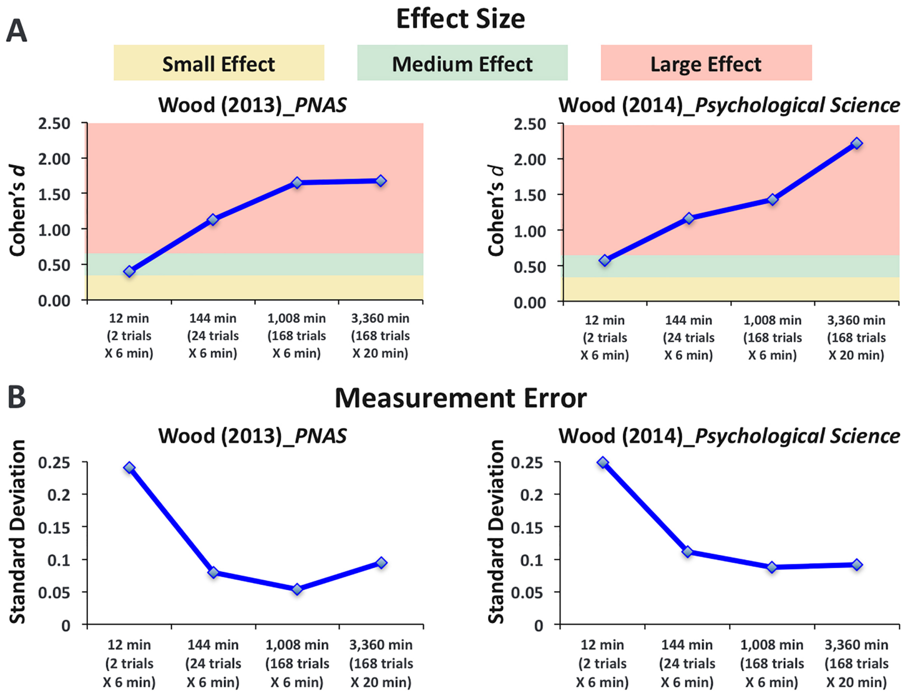

Fig. 4. Automation can significantly (A) increase the effect size and (B) decrease the measurement error of an experiment, by allowing large amounts of data to be collected from each subject. For example, for Wood (2013) and Wood (2014), increasing the amount of data collected from each chick increased the observed effect size by a factor of 4 and reduced the measurement error by a factor of 2.6. Thus, collecting more data per subject significantly increases the precision of data, leading to high powered experiments.

considered small, d = 0.50 medium, and d = 0.80 large). We found that automated studies tend to produce very large effect sizes (average Cohen's d = 1.43 in our sample, blue points in Fig. 1B). In contrast, non-automated studies produce small to medium effect sizes (average Cohen's d = 0.37 in our sample; red points in Fig. 1B). Thus, automated studies produce effect sizes that are 3–4 times larger than non-automated studies.

To better understand why automation improves measurement precision, we re-analyzed previously published data from our lab. In the analysis, we tested how the inclusion of more data per chick influenced the observed effect size for the first two studies published from our lab (Wood, 2013, 2014). As shown in Fig. 4, when only 12 min of data were analyzed per chick (the first 6 min of 2 trials = 12 min of data per subject), the average Cohen's d hovered in the medium range, $d \sim 0.5$. When we increased the data to include a full day of these shortened trials (the first 6 min of 24 trials = 144 min of data per subject), the average Cohen's d rose to $d \sim 1.1$. When we included all of our available data (20 min trials × 24 trials per day × 7 days = 3360 min of data per subject), the average Cohen's d rose to $d \sim 1.9$. Thus, increasing the amount of data collected per chick reduced the measurement error by a factor of 2.6 and increased the effect size by a factor of 4.

In addition to improving experimental accuracy, another benefit of reducing measurement error is that fewer subjects are needed to detect an effect. While the standard recommendation in response to the replication crisis is to *increase* sample size, statisticians are now suggesting that an even better focus is to improve the quality of measurements (reviewed by Gelman, 2018). Mathematically, reducing measurement error by a factor of 3 is as good as multiplying the sample size by 9.

To illustrate this effect, we computed the sample size needed to obtain 80% power for a range of "true" population performance magnitudes (Fig. 1C; Champely, 2015). Automated methods that produce precise measurements (SD = 10%) require far smaller numbers of subjects for 80% power than non-automated methods that produce noisy measurements (SD = 33%). For example, achieving adequate power to detect a true population performance of 58% (chance = 50%) with automated methods requires 15 chicks, while detecting the same effect with non-automated methods requires 136 chicks!

While automated controlled-rearing methods cannot be applied to human subjects, the principle of increasing the precision of measurements need not be unique to studies of newborn chicks. Methods that reduce statistical noise can improve the effect sizes and power of infant studies as well. For example, automated methods using tablets and head-mounted eye tracking may allow researchers to collect more data from each subject, while reducing experimenter bias in the data collection process.

5. Automation allows individual-level statistical analyses

After collecting a single, short trial from each chick, experimenters can compute a group-level measure of average performance. However, this group-level statistic does not tell us whether all of the chicks had the ability of interest or whether just a subset of the chicks had the ability. In fact, studies with noisy measurements typically find that a large proportion of the subjects performed in the *opposite direction* of the reported effect (Fig. 2). This is problematic from a theoretical perspective because if an ability is foundational to perception and cognition, then it should be present in most – if not all – newborn animals (provided we have sufficiently precise methods to detect the ability).

Automated testing provides a direct solution to this problem. By collecting hundreds of hours of data per chick, automation allows researchers to obtain precise measurements of each subject's performance and assess whether each chick performed above (or below) chance level (e.g., Wood, 2013, 2017). With massive amounts of data from each chick, these individual-level analyses are often highly powered: many reach 5-sigma levels of statistical significance (the statistical threshold for new discoveries in theoretical physics) for individual subjects.

Another benefit of individual-level analyses is that each subject can serve as a replication of an effect. If we consider an experiment as being on an individual level (rather than group level), then each subject provides an opportunity to replicate an effect. Experiments at the individual level are only possible with enough measurements from each subject to reliably reject or fail to reject the null hypothesis for each individual. Thus, automation permits individual-level internal replications.

Individual-level analyses may also be possible in studies of infants and toddlers. Online methods of testing allow children to participate in studies without having to continually travel back and forth to a physical laboratory. By recruiting infants and toddlers for multiple testing sessions, researchers of human development can also examine variation between individuals and use individual-level analyses as internal replications of an effect.

6. Eliminating analytic flexibility through preregistration and automation

Our paper has focused largely on improving measurement error and effect sizes. However, combating the replication crisis also requires reducing analytic flexibility ("researcher degrees of freedom") in experimental designs. For example, researchers can (1) engage in flexible data collection (e.g., continuing to collect data until the results reach statistical significance), (2) perform a range of analyses and report only those analyses that reached statistical significance, and (3) continue to test a hypothesis with slight methodological changes until a statistically significant result emerges, then frame the failed attempts as limits to the phenomenon (Peterson, 2016). These research practices dramatically increase false-positive rates (Simmons et al., 2011).

The most direct way to combat these problems is to preregister the design and analyses of experiments. Preregistration requires researchers to describe, in as much detail as possible, their plans for the study (e.g., number of subjects, stimulus materials, procedures, measures, rules for excluding data, plans for data analysis, predictions/hypotheses) and to post those plans in a time-stamped, locked file in an online repository that can be accessed by editors and reviewers (and, ultimately, by readers). By forcing researchers to make these decisions before collecting data, preregistration minimizes researcher degrees of freedom and reduces the probability of obtaining false-positives.

Automation can also be useful for combating experimental and analytic flexibility. To provide an example, our lab recently started automating data analysis procedures. We first create an R script of the analyses we plan to run, then use that script to analyze the raw data from the animal tracking computer. Thus, we formalize the statistical tests we plan to run *before* looking at the data. For each experiment, we then create a "reproducibility package" containing the R script and raw data. Running the package automatically reproduces the statistical analyses from the raw data. We are also starting to automate replication procedures, by creating "replication packages" containing all of the stimuli, materials, and code needed to replicate our experiments. Running the package automatically recreates all of the stimuli presented to the chicks, in exactly the same order as the original experiment. In the future, we plan to begin making these reproducibility and replication packages openly available to researchers (e.g., by posting them on our lab website), helping contribute to a culture of transparent and reproducible research. Automated data analysis procedures could also be used by laboratories that study infants and children to improve the reproducibility of statistical analyses.

7. Conclusion

A central goal in science is to produce robust and reproducible research. One strategy for improving reproducibility is to reduce measurement error (Gelman, 2018). In this paper, we show how automation can reduce measurement error by ~60% in controlled-rearing studies of newborn chicks. Specifically, automated controlled-rearing studies produce measurements that are 3–4 times more precise than non-automated studies and produce effect sizes that are 3–4 times larger than non-automated studies.

Our paper focuses on controlled-rearing experiments, but this strategy – which emphasizes improving measurement precision – could also be useful for other areas of developmental psychology. In response to the replication crisis, developmental psychologists have tended to focus on other strategies (e.g., increasing sample size, preregistration, open data policies, and performing exact replications). We agree that all of these strategies are valuable, but we suggest that researchers should be especially focused on reducing the amount of noise in their data.

How can researchers improve measurement precision in their experiments? One straightforward strategy is to collect more data from each subject. We found a strong positive correlation between the amount of data collected from each subject and the precision of the measurements (Fig. 4). This correlation is not limited to studies of non-human animals. Analyses of infant looking time designs have revealed that including more trials per infant can also result in large increases in statistical power (DeBolt, Rhemtulla, & Oakes, 2019). Moreover, by collecting more data per subject, researchers can determine the homogeneity or heterogeneity of their effects across infant and toddler populations.

Many researchers in developmental psychology are already moving in this direction. For example, while it is not possible to experimentally test human infants 24/7, new automated internet-based platforms, such as "Lookit", allow families to participate in behavioral studies via webcam (Scott, Chu, & Schulz, 2017; Scott & Schulz, 2017). Thus, researchers can begin recruiting larger numbers of infants and collecting repeated measurements. There is also growing use of automation to track infants' eye movements (Smith, Jayaraman, Clerkin, & Yu, 2018) and language acquisition (Roy, Frank, DeCamp, Miller, & Roy, 2015). Roy and colleagues, for example, used automation to collect more than 200,000 h of audio and video recordings from all of the rooms in a child's house, providing a detailed case study of the emergence of language.

Robust, replicable findings are vital for formulating and evaluating scientific theories. We conclude that automation can be a powerful tool for combating the replication crisis and improving the precision of scientific measurements in developmental psychology. Precise measurements make it easier to detect true effects, reduce the probability of detecting false effects, and decrease the number of subjects needed to achieve adequate power. Automated methods also protect against experimenter bias, enable individual-level statistical analyses, and allow replications to be performed quickly and easily.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.infbeh.2019.101329.

References

Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., & Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nature Neuroscience, 20*(10), 1404.

Bellen, H. J., Tong, C., & Tsuda, H. (2010). 100 years of drosophila research and its impact on vertebrate neuroscience: A history lesson for the future. *Nature Reviews Neuroscience, 11*(7), 514–522. https://doi.org/10.1038/nrn2839.

Brenner, S. (1974). The genetics of caenorhabditis elegans. *Genetics, 77*(1), 71–94.

Carey, S. (2009). *The origin of concepts. Oxford series in cognitive development*. Oxford; New York: Oxford University Press.

Champely, S. (2015). pwr: Basic functions for power analysis. *R package version, 1*(1).

Cosmides, L., & Tooby, J. (1994). Origins of domain-specificity: The evolution of functional organization. In L. Hirschfeld, & S. Gelman (Eds.). *Mapping the Mind: Domain-specificity in cognition and culture*. New York: Cambridge University Press.

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2019). *Robust data and power in infant looking time research: Number of infants and number of trials. Talk presented at the Society for Research in Child Development* Baltimore, MD.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin, 44*(1), 16–23.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465.

Gibson, E. J., & Walk, R. D. (1960). The "visual cliff". *Scientific American, 202*(4), 64–71.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., & Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546–573.

Hassabis, D., & Kumaran, D. (2017). Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology, 56*(5), 872.

Hinton, G. E. (1992). How neural networks learn from experience. *Scientific American, 267*(3), 144–151.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154.

Hume, D. (1748). *An enquiry concerning human understanding.* Oxford: Clarendon Press 1975 edition.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124.

Jarvis, E. D., Gunturkun, O., Bruce, L., Csillag, A., Karten, H., Kuenzel, W., Medina, L., Paxinos, G., Perkel, D. J., Shimizu, T., Striedter, G., Wild, J. M., Ball, G. F., Dugas-Ford, J., Durand, S. E., Hough, G. E., Husband, S., Kubikova, L., Lee, D. W., Mello, C. V., Powers, A., Siang, C., Smulders, T. V., Wada, K., White, S. A., Yamamoto, K., Yu, J., Reiner, A., & Butler, A. B. (2005). Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews: Neuroscience, 6*(2), 151–159. https://doi.org/10.1038/nrn1606.

Kandel, E. R. (2007). *In search of memory: The emergence of a new science of mind.* WW Norton & Company.

Karmiloff-Smith, A. (1995). *Beyond modularity: A developmental perspective on cognitive science.* MIT Press.

Karten, H. J. (2013). Neocortical evolution: Neuronal circuits arise independently of lamination. *Current Biology, 23*(1), R12–R15. https://doi.org/10.1016/j.cub.2012.11.013.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Štepán, Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., & Brumbaugh, C. C. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142.

Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science, 321*(5895), 1502–1507.

Li, N., & DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron, 67*(6), 1062–1075. https://doi.org/10.1016/j.neuron.2010.08.029.

Locke, J. (1689). In R. Woolhouse (Ed.). *An Essay Concerning Human Understanding.* London: Penguin.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585.

Martinho, A., & Kacelnik, A. (2016). Ducklings imprint on the relational concept of "same or different". *Science, 353*(6296), 286–288. https://doi.org/10.1126/science.aaf4247.

Mascalzoni, E., Regolin, L., & Vallortigara, G. (2010). Innate sensitivity for self-propelled causal agency in newly hatched chicks. *Proceedings of the National Academy of Sciences of the United States of America, 107*(9), 4483–4485. https://doi.org/10.1073/pnas.0908792107.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*(2), 159.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251).

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius, 2*, 1–10.

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences, 20*(4), 537–556.

Regolin, L., & Vallortigara, G. (1995). Perception of partly occluded objects by young chicks. *Perception & Psychophysics, 57*(7), 971–976.

Rosa-Salva, O., Regolin, L., & Vallortigara, G. (2010). Faces are special for newly hatched chicks: Evidence for inborn domain-specific mechanisms underlying spontaneous preferences for face-like stimuli. *Developmental Science, 13*(4), 565–577. https://doi.org/10.1111/J.1467-7687.2009.00914.X.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences, 112*(41), 12663–12668.

Rugani, R., Regolin, L., & Vallortigara, G. (2010). Imprinted numbers: Newborn chicks' sensitivity to number vs. continuous extent of objects they have been reared with. *Developmental Science, 13*(5), 790–797.

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind, 1*(1), 4–14.

Scott, K., Chu, J., & Schulz, L. (2017). Lookit (part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind, 1*(1), 15–29.

Shanahan, M., Bingman, V. P., Shimizu, T., Wild, M., & Gunturkun, O. (2013). Large-scale network organization in the avian forebrain: a connectivity matrix and theoretical analysis. *Frontiers in Computational Neuroscience, 7*, 89. https://doi.org/10.3389/fncom.2013.00089.

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*(11), 1359–1366.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology, 8*, 2124. https://doi.org/10.3389/fpsyg.2017.02124.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10*(1), 89–96. https://doi.org/10.1111/J.1467-7687.2007.00569.X.

Srihasam, K., Vincent, J. L., & Livingstone, M. S. (2014). Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nature Neuroscience, 17*(12), 1776.

Steven, P. (2003). *The blank slate: The modern denial of human nature.* Penguin.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to cognition and action.* Cambridge, MA: MIT Press.

Vallortigara, G. (2012). Core knowledge of object, number, and geometry: A comparative and neural approach. *Cognitive Neuropsychology, 29*(1-2), 213–236.

Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually inexperienced chicks exhibit spontaneous preference for biological motion patterns. *PLoS Biology, 3*(7), 1312–1316. https://doi.org/10.1371/journal.pbio.0030208.

Walk, R. D., Gibson, E. J., & Tighe, T. J. (1957). Behavior of light-and dark-reared rats on a visual cliff. *Science*.

Wood, S. M. W., & Wood, J. N. (2015). A chicken model for studying the emergence of invariant object recognition. *Frontiers in Neural Circuits, 9*, 7. https://doi.org/10.3389/fncir.2015.00007.

Wood, J. N. (2013). Newborn chickens generate invariant object representations at the onset of visual object experience. *Proceedings of the National Academy of Sciences of the United States of America, 110*(34), 14000–14005.

Wood, J. N. (2014). Newly hatched chicks solve the visual binding problem. *Psychological Science, 25*(7), 1475–1481. https://doi.org/10.1177/0956797614528955.

Wood, J. N. (2017). Spontaneous preference for slowly moving objects in visually naïve animals. *Open Mind, 1*(2), 111–122.

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), 20160166.

Wood, J. N., & Wood, S. M. W. (2018). A smoothness constraint on the development of abstract object representations. *Cognitive Science*. https://doi.org/10.1111/cogs.12595.