Received: 30 June 2021 | Revised: 24 August 2022 | Accepted: 3 September 2022

DOI: 10.1111/mila.12450

SUBMITTED ARTICLE

WILEY

In defense of epicycles: Embracing complexity in psychological explanations

Ansgar D. Endress

Department of Psychology, City, University of London, London, UK

Correspondence
Ansgar D. Endress, Department of Psychology, City, University of London, Northampton Square, London EC1V 0HB, UK.
Email: ansgar.endress.1@city.ac.uk

Is formal simplicity a guide to learning in humans, as simplicity is said to be a guide to the acceptability of theories in science? Does simplicity determine the difficulty of various learning tasks? I argue that, similarly to how scientists sometimes preferred complex theories when this facilitated calculations, results from perception, learning and reasoning suggest that formal complexity is generally unrelated to what is easy to learn and process by humans, and depends on assumptions about available representational and processing primitives. "Simpler" hypotheses are preferred only when they are also easier to process. Historically, "simpler", easier-to-process, scientific theories might also be preferred if they are transmitted preferentially. Empirically viable complexity measures should build on the representational and processing primitives of actual learners, even if explanations of their behaviour become formally more complex.

KEYWORDS

Bayesian learning, induction, language acquisition, learning constraints, Occam's razor, perceptual or memory primitives, simplicity

1 | INTRODUCTION

We like to explain things. Luckily we are not particularly good at it and can thus enjoy stories of gods creating other gods by throwing around godly genitals (Hesiod., 1914) as opposed to less

dramatic and "simpler" descriptions of celestial bodies forming other celestial bodies when their dust moves peacefully through space and accretes (Armitage, 2008).

While it might not make for good stories, a particularly prominent metric of the quality of an explanation is its simplicity. This principle is known as Occam's razor (e.g., Baker, 2016; Fitzpatrick, 2022): When two theories explain the same phenomenon, the simpler one, requiring fewer assumptions, should be given preference.

As simplicity is a good metric to gauge the quality of *scientific* theories (but see, e.g., Baker, 2016; Fitzpatrick, 2022), and as our mental representations are arguably some form of *mental* theory of the world, many authors propose that simplicity might also be a good guiding principle for the kinds of mental representations we entertain and, more generally, for the kinds of inductive biases we have when we learn about the world to establish these mental representations to begin with.

However, the history of science suggests that a formal notion of simplicity is not the only extra-empirical yardstick to judge the quality of a scientific theory. For example, Russo (2000) argues that the often ridiculed Ptolemaic model of planetary motion was an exceptionally efficient computational device:

> Because "epicycles" is still a byword for clumsy and backward attempts at science, we spell out the two reasons why the method was supremely well-adapted to the purposes to which it was put.
>
> First, accounting for the observed motion of planets as the composite of several uniform motions on circular orbits (the first centered on the earth ... and each of the others, called epicycles, centered on the point obtained on the preceding circumference) is equivalent to a modern expansion in Fourier series, and allows and efficient description of observed data with increasing precision as the number of epicycles grows. ... Second, since the main computational tool of Hellenistic mathematics was geometric algebra performed with ruler and compass, decomposition into circular motions was the most efficient possible system for computing the observable position of planets. (Russo, 2000, pp. 90–91).

In other words, although heliocentric models have been known for centuries at the time of Claudius Ptolemaeus, the computational tools available to Hellenistic astronomers made a geocentric model efficient for their purposes even though the heliocentric model is simpler (though this depends on what exactly we mean with "simplicity"; see, e.g., Baker, 2016; Fitzpatrick, 2022).

Here, I argue that a similar situation arises when it comes to mental representations and the learning thereof. We are endowed with a certain set of computational tools, and any learning problem (as well as the application of what has been learned) needs to make do with the tools we have at our disposal. As a result, any psychologically viable notion of simplicity needs to take into account the computational machinery we have evolved, both for learning and, later on, for applying what has been learned (i.e., for processing).

Specifically, I will make three claims. First, what is simple from a formal point of view is generally unrelated to what is simple from a learning and processing point of view, maybe because it is easier to learn and process what is more frequent in our environment, which, in turn, makes it more likely that we have evolved machinery to deal with these frequent occurrences.[1]

[1]While the proposals discussed below make no explicit reference to ease of learning, they do specify those regularities that should be acquired preferentially. In empirical and behavioral terms, however, the regularities that are acquired preferentially are indistinguishable from those that are acquired more easily.

Further, percepts, from sensory input to utterances, often require active interpretations of the stimuli, and these interpretations are unlikely to be the formally simplest interpretations. Conversely, when people prefer simpler explanations of empirical phenomena over more complex explanations, they do so because the simpler explanations happen to be easier to process and not due to a genuine preference for simplicity. Historically, easier-to-process hypotheses might also be learned and transmitted preferentially, leading to a normative preference for subjectively "simpler" theories.

Second, formal claims that learning is guided by simplicity considerations either misstate mathematical facts or are empirically unsupported. Such claims depend on auxiliary assumptions, have alternative (and simpler) explanations, are not necessarily consistent with the data they seek to explain, make incorrect predictions, or do not scale up to learning problems in actual humans. Third, I suggest that any empirically adequate concept of simplicity requires a detailed understanding of the mental computations and processing constraints.

While the arguments below focus on the question of whether simplicity guides learning in the common human, very similar arguments have been advanced in the philosophy of science. Simplicity is hard to define, a preference for "simpler" explanations is hard to justify, historically, it is unclear whether simplicity really guided scientific practice (or rather was a rhetorical device use to justify decisions already taken), the relative simplicity of two theories depends on the formalism in which the theories are described, and, even in cases where simplicity reputedly favored one theory of another (e.g., the Copernican vs. the Ptolemaic model of planetary motion), it is not always clear which of two theories is the simpler one (Baker, 2016; Fitzpatrick, 2022). Even when simplicity is valued as an attribute of a good scientific theory, its role is to "[bring] order to phenomena that in its absence would be individually isolated and, as a set, confused" (Kuhn, 1977, p. 322), and thus essentially a (collective) memory strategy that helps scientists to represent a set of phenomena, similarly to how, at the individual level, Hellenistic and later memory techniques embed memoranda in coherent contexts (Yates, 1966).

2 | IS FORMAL COMPLEXITY RELATED TO EASE OF PROCESSING?

2.1 | Extraneous information in everyday inference making

Occam's razor has a venerable history in science (e.g., Baker, 2016; Fitzpatrick, 2022). As simpler models, based on only few overarching principles, are easier to falsify (but see Fitzpatrick, 2022), Popperian, falsification-based scientific paradigms will thrive with simpler hypotheses. But people are not scientists. They indulge in conspiracy theories that require numerous auxiliary hypotheses and like to think that chemicals in the drinking water turn us gay, that 9/11 was an inside job, that Elvis is alive, and so forth. Scientists are not immune from such biases either. Giants of mathematics such as Leibniz and Newton were heavily influenced by occult religious (Hermetic) traditions (Keynes, 2010; Yates, 1966), and the inferences made by actually existing scientists depend on other aspects of their worldview (e.g., Kuhn, 1977, Chapter 13, pp. 320–339; Latour & Woolgar, 1986). Actual humans thus do not necessarily favor the simplest available hypotheses. (In contrast to refutation-based scientific theories, they are

also remarkably resistant to correction, potentially due to their confirmation bias; e.g., Kahneman, 2011.)

This is not to say that the kinds of inferences people make are illogical or unhelpful for surviving in their environment (as has long been recognized in anthropology; e.g., Lévi-Strauss, 1962). Rather, making complex and sometimes extravagant inferences might not be so much a bug of the mind as a critical and necessary feature. For example, what people are told is generally underspecified. To borrow an example from Sperber and Wilson (1987), a perfectly reasonable answer to the question "Do you want some coffee" is "Coffee will keep me awake". However, this sentence is not an actual answer to the question. Rather, the listener has to infer from extraneous information that is little more than a physiological truism ("coffee will keep me awake") that the actual answer is "No". These kinds of situations are so widespread that entire theories have been devoted to them (e.g., Sperber & Wilson, 1987, 1995), and should abolish any hope that successful communication would be possible if a listener's inferences were "simple" in any relevant sense.

Of course, the tendency of perceivers to have rich interpretations of the world is not limited to high-level phenomena such as decision-making and communication (though they are amply documented in such domains; see, e.g., Gigerenzer & Goldstein, 1996; Todd & Gigerenzer, 2000). Rather, they routinely occur in perception as well. In the words of Ramachandran (1991) "perception is essentially a 'bag of tricks' ... through millions of years of trial and error the visual system has evolved numerous short-cuts, rules-of-thumb and heuristics which were adopted not for their aesthetic appeal or mathematical elegance but simply because they worked" (347). This insight is by no means specific to sensory or mental processes; rather, organismal traits are rarely optimized on a stand-alone basis but evolve within the evolutionary history and constraints of an organism (e.g., Gould et al., 1979).

For example, if we see two lines of equal length, the simplest inference is probably that they *are* of equal length. But this is not the inference the perceptual system draws: Even though the two lines are of the same length, a line surrounded by arrow tails (e.g., ≻—≺) is perceived as longer than one surrounded by arrow heads (e.g., ⟷). In this case, however, the reason for the arguably more complex inference is clear: In natural scenes, line segments surrounded by arrow tails tend to be longer than those segments surrounded by arrow heads (Howe & Purves, 2005). Given these environmental statistics, our perceptual system thus evolved to make inferences that are maybe not "simple" in any obvious formal sense (that does not take into consideration historical environmental statistics), but that are adaptive given those environmental statistics, and there is no shortage of other cognitive mechanisms that might have evolved due to environmental constraints (e.g., New et al., 2007; Pinker, 1998; Sperber, 1994; Sugiyama et al., 2002).

In the next section, I will argue that similar phenomena occur in learning and processing: Simplicity in a formal sense is a poor guide towards what is easy to learn or process. Rather, what is simpler or easier to learn is determined by the learning mechanisms that happen to be available. Following this, I will argue that, in cases where observers do choose simpler hypotheses, the simplicity bias might be an emergent property of other cognitive biases, and that putatively simpler inferences might be those that can be processed with more fluency.

2.2 | Simpler is not simpler

There is no shortage of examples where formally more complex operations are easier for humans. For example, (arithmetic) divisions are hard for humans but easy for computers, while we can do elaborate 3D rotations which require considerable processing power in computers.

Likewise, it is easier to notice that a figure is symmetric (e.g., 𝄜) than that it is composed of the copy of two shapes (e.g., 𝄝) even though symmetry entails more operations than copying (e.g., Baylis & Driver, 1994, 2001; Bruce & Morgan, 1975; Corballis & Roldan, 1974): In the second figure, the 𝄞 shape is copied and then translated to the right; in the first figure, the 𝄞 is again copied and then translated, but, crucially, also mirror-reversed. Still, it is easier to detect symmetry than translation.

Such cases suggest that formally "simpler" operations are not necessarily easier to process for humans. But ease of processing is not even constant for a given operation. I will now provide examples illustrating that the difficulty to learn or execute a given operation strongly depends on the domain in which it is applied (see Endress, 2019, for a review). If the difficulty of a learning problem depends on its domain, the difficulty cannot be driven by the problem's formal complexity.

This is true even for the simplest operations such as associations. For example, animals readily associate tastes with visceral sickness, and external events (e.g., sounds and light) with pain. In contrast, it is much more difficult or even impossible to associate taste with pain, or external events with sickness (e.g., Garcia & Koelling, 1966; Garcia et al., 1974, 1976; see, e.g., Domjan, 1983, 2015, for reviews). Evolutionarily speaking, this makes sense of course: Sickness typically results from what we ingest, while physical pain typically has external causes that we can perceive sensorily, which makes the pattern of preferential associations adaptive (see Alberts & Gubernick, 1984; Gemberling & Domjan, 1982; Gemberling et al., 1980; Gubernick & Alberts, 1984, for evidence that this pattern of preferential associations is not learned). From a formal perspective, however, both types of associations are equally complex, simply because they reflect the very same computation.

Likewise, if we hear, see or feel two objects frequently occurring together, we form associations between them (e.g., Aslin et al., 1998; Conway & Christiansen, 2005; Endress, 2010; Fiser & Aslin, 2002; Saffran et al., 1999; Saffran, Newport, & Aslin, 1996; Turk-Browne et al., 2005; Turk-Browne & Scholl, 2009). However, this form of associative learning works better for consonants than for vowels (Bonatti et al., 2005), although the reasons are debated (Bonatti et al., 2007; Keidel et al., 2007). Again, the learnability of equally complex operations seems to depend on their domain, suggesting that formal complexity is a poor guide to the ease of acquisition and processing of an operation.

A similar conclusion follows from an operation that has helped defining the complexity of phonological processes (e.g., Culy, 1985; Manaster-Ramer, 1986): repetition-patterns. Such patterns can be learned in many domains, but not in many others; as the formal complexity of learning the very same pattern across domain is presumably identical, this suggests again that learnability is not equivalent to the complexity of a rule.

Specifically, repetition-patterns are important in many languages. For example, Marshallese uses reduplications for derivative morphology (e.g., "takin" means sock, while "takinkin" means to wear socks; Moravcsik, 1978). In some form or another, reduplication occur in some 85% of the world's languages (Rubino, 2013). Other examples of repetition-patterns used by language include vowel harmony (e.g., Rose & Walker, 2011; Vroomen et al., 1998), the feature

repetitions resulting from assimilation rules (e.g., Darcy et al., 2009; Mitterer & Blomert, 2003) and constraints on consonant co-occurrence in Semitic languages such as the Obligatory Contour Principle (e.g., Berent & Shimron, 1997; Frisch et al., 2004; McCarthy, 1986; McCarthy & Prince, 1999).

Despite their importance for language, repetition-patterns can be perceived in many non-linguistic domains and by many non-linguistic animals. For example, even seven-month-old babies notice the repetition in syllable sequences such as *dubaba* and generalize it to new items (Marcus et al., 1999). When familiarized with syllable sequences such as *ledidi*, *wijeje* and so forth, infants behave as if they were more familiar with novel sequences with novel syllables that share the repetition-pattern (e.g., *bapopo*) than with other novel syllable sequences that do not share the pattern (e.g., *babapo*). Humans and non-human animals can compute repetition patterns not only for speech syllables, but also for tones and visual objects (e.g., Dawson & Gerken, 2009; Endress et al., 2007; Giurfa et al., 2001; Hauser & Glynn, 2009; Marcus et al., 2007; Marcus et al., 1999; de la Mora & Toro, 2013; Murphy et al., 2008; Neiworth, 2013; Pepperberg, 1987; Saffran et al., 2007; Smirnova et al., 2015; Versace et al., 2017; Yamazaki et al., 2012, but see Spierings & ten Cate, 2016; van Heijningen et al., 2013, for evidence that these relations are not equally salient to all species).

However, although the ability to detect such patterns is widely shared across animals and domains, there are some domains where they are easier to learn than in others. For example, adult speakers learn such patterns better when they are carried by vowels than when they carried by consonants, to the extent that they fail to detect the patterns on consonants (e.g., Toro, Bonatti, et al., 2008; see Hochmann et al., 2011; Pons & Toro, 2010, for similar results with infants), even when the salience of the vowels is much reduced (Toro, Shukla, et al., 2008). Further, given that rats learn repetition-patterns just as well on consonants as on vowels (de la Mora & Toro, 2013), the reason for the vowel advantage in humans does not seem to be a gross perceptual difference.

Likewise, adult speakers seem unable to learn repetition-patterns over syntactic categories (Endress & Hauser, 2009). When familiarized with word triplets conforming to either an AAB pattern (noun-noun-verb triplets such as town-leg-choose and verb-verb-noun triplets such as choose-speak-leg) or an ABB pattern (noun-verb-verb or verb-verb-noun), they are unable to decide whether test triplets made from new words were like the familiarization items. Again, there does not seem to be any particular formal reason for this failure: Participants readily access the categories, learn other sequential regularities about the categories that do not involve repetitions, and learn repetition-patterns over semantic, non-syntactic categories (e.g., animals and clothes). Endress and Hauser's (2009) conclusion was that repetition-patterns over syntactic categories are simply not in the learner's repertoire because they are not part of any human language. Critically, however, participants failed to learn an extremely simple regularity that they should learn based on simplicity considerations, as they can learn it in many other contexts. Taken together, this evidence thus suggests that the formal simplicity of an operation does not seem to predict how easily it is learned.

2.3 | Simpler is also easier

The examples presented so far suggest that learning and processing biases cannot be predicted based on simplicity considerations. Rather, learners process those regularities more easily that happen to be supported by the learning mechanism they come equipped

with, and the available learning mechanisms might even differ across domains (see also Endress, 2019; Endress et al., 2009). I will now suggest that the converse holds as well: Some inferences might be subjectively simpler (and thus preferred) because they happen to fit processing constraints.

The evidence comes from experiments where participants have to choose between (or rate) competing explanations of phenomena. While observers sometimes prefer more complex explanations containing irrelevant information (e.g., Hopkins et al., 2016; Weisberg et al., 2008), they sometimes also choose "simpler" explanations (see Lombrozo, 2016, for a review). However, I will now argue that the simpler explanations are also those more attuned to human processing constraints, and that a preference for simpler explanations might be an emerging property of other processing biases related to general pragmatic factors and memory strategies such as chunking.

Such processing biases might also contribute to why (subjectively) simpler theories might be preferred historically: Theories that fit our processing biases are easier to transmit, and, in line with prominent accounts of cultural evolution (e.g., Kirby et al., 2007), more likely to survive.

2.3.1 | Pragmatics versus breadth of explanation

Read and Marcus-Newhall (1993) asked whether people prefer single hypotheses explaining multiple phenomena, or rather multiple hypotheses explaining one phenomenon each. In one of their scenarios, Cheryl presents with three symptoms: nausea, weight gain and fatigue. Participants had to choose a diagnosis among three narrow explanations that each explained one of the symptoms, and a single broad explanation that explained all three symptoms. When participants were told about just one symptom, they favored a narrow explanation that did not account for the other symptoms; for example, they favored a stomach virus as the cause for nausea when no other symptoms were mentioned. In contrast, when informed about all three symptoms, participants favored a "simpler" and broader explanation that explained all three symptoms (i.e., pregnancy), even though, in principle, the symptoms could be explained by a conjunction of three separate and narrower explanations.

While such results are consistent with a preference for simpler hypotheses, they can also emerge from pragmatic factors. If participants expect to be provided with *relevant* information (e.g., Grice, 1975; Sperber & Wilson, 1987, 1995), they might reasonably opt for the explanation that is most strongly associated with the facts to be explained (e.g., using the availability heuristics, Tversky & Kahneman, 1973, or Grice's [1975] maxims of quantity). Similarly, they might expect to be provided with *all* relevant symptoms, especially when they are strongly associated with a condition (e.g., weight gain and pregnancy), and might not think about symptoms that are not mentioned. After all, the description of a big, four-legged mammal gracefully grazing on a pasture is hard to recognize as an elephant when the most strongly associated feature (e.g., the trunk) is not mentioned, and a listener will likely conclude that the animal in question does *not* have a trunk when none is mentioned. As a result, the participants preference for narrower explanations might emerge from assumptions about the experimenters' communicative intentions rather than from epistemic biases per se.

2.3.2 | Memory fluency versus model flexibility

Simpler hypotheses might also be preferred when they allow for more fluent memory processing. For example, Blanchard et al. (2018) asked if participants preferred hypotheses that are less flexible and can account for a narrower set of data. For example, a "simple" model of a series of coin tosses is a binomial model with the success probability fixed to .5; a more flexible model would finetune the success probability. Simpler, less flexible models should be favored for a variety of reasons (e.g., Baker, 2016), and in fact, are favored by standard model selection metrics such as the Akaike Information Criterion or the Bayesian Information Criterion.

To test whether humans are sensitive to model complexity, Blanchard et al.'s (2018) participants had to classify animals as "velmos" or "zorgits". Both species had red spots on their backs, but the typical number of spots differed across species. All participants were told that velmos can have a random number of spots, with three possible values. For zorgits, the possible numbers of spots varied across participants. Some participants were told that the number of spots was deterministic, others that the number of spots could take four possible values, and yet others that the number of spots could take 100 possible values. Results showed that participants were increasingly less likely to classify the animals as zorgits as the number of possible values increased. They thus rejected increasingly complex hypotheses.

Participants thus preferred simpler hypotheses. However, this preference might emerge from processing constraints. After all, it is presumably hard to picture a distribution with 100 possible values, or to keep the relevant situations in working memory to evaluate their likelihoods. Participants might thus reject hypotheses that are too hard to understand or to process. In line with this possibility, text that is hard to read is judged as less true (Reber & Schwarz, 1999), and authors of excessively complex prose are judged as less competent (Oppenheimer, 2006). If participants use the processing fluency as a cue to the likelihood of an explanation, a simplicity biases might thus emerge from processing considerations.

2.3.3 | Chunking versus simplicity

People might also prefer hypotheses that allow for more efficient memory processing through chunking. For example, Lombrozo's (2007) participants had to explain a set of facts by competing hypotheses that differed in the number of causes. For example, a space alien might have two symptoms. Each symptoms could each be caused by a separate disease (two causes), or both symptoms could reflect a single common disease (one cause). Participants overwhelmingly chose the single cause explanation.

Such results are consistent with a preference for simpler explanations with fewer causes. However, a single cause explanation might also reduce the working memory load of the explanations because it would allow participants to chunk the symptoms together. In fact, even young infants can retain more information in working memory when arbitrary verbal labels are given (e.g., Feigenson, 2008; Kibbe & Feigenson, 2014). For example, while they cannot keep track of four identical objects, they can do so when two of the objects are labelled as "dax", and

two as "blickets" (Feigenson, 2008), presumably because such labels allow infants to first chunk the information, and later act as retrieval cues. Mutatis mutandis, a single cause explanation might act as a chunking cue binding together symptoms and proximate causes, and thus increase processing fluency, and participants might prefer explanations that are easier to process. Alternatively, they might also prefer explanations with stronger associations between the symptoms and the diseases.[2]

Likewise, Pacer and Lombrozo (2017) contrasted two possible metrics for the simplicity of an explanation: The total number of causes ("node simplicity") and the number of *unexplained* causes ("root simplicity"). For example, a patient might present with weight loss and fatigue. One explanation provides separate causes for each of the symptoms (e.g., reduced appetite and insomnia); this explanation thus entails two causes in total, but both causes are unexplained. Another explanation provides a higher level explanation; for example, depression might cause both reduced appetite and insomnia, which cause weight loss and fatigue in turn. This explanation involves three causes in total, but only one of them is unexplained. Results showed that participants favored the second explanation, thus minimizing the number of unexplained causes rather than the total number of causes.

However, as in the case of Lombrozo's (2007) results, participants might minimize the number of unexplained causes to reduce the working memory load. Adding a deeper, more fundamental cause might allow participants to chunk both causes and symptoms, and thus to process them more fluently.[3] In fact, both Pacer and Lombrozo (2017) and Lombrozo (2007) showed that participants tend to misremember frequency information in a way that made it more consistent with their chosen hypothesis, suggesting that participants seek to construct an effective memory representation of their preferred explanation.

2.3.4 | Are simpler explanations transmitted better?

The examples in this section suggest that, in some cases, humans might end up preferring simpler explanations if they choose explanations that are easier to process. Applying Occam's razor

[2]A preference for the single cause explanation might also emerge from basic associative processes. Participants might associate symptoms with diseases. When they are informed that a disease can cause a symptom, they might set the conditional probability of the symptom given the cause P(symptom | cause) to one. In contrast, when a symptom is not mentioned, the pragmatic situation might lead participants to conclude that the disease does *not* predict the symptom and that the conditional probability is zero. For the two cause explanation, the average conditional probability of the symptoms given the disease P(symptom | cause) is thus .5, because each disease explains only one of the symptoms. In contrast, the average conditional probability is 1.0 for the single cause explanation, because the disease explains both symptoms.A similar picture emerges when considering the reversed conditional probabilities, P(disease | symptom). For the separate cause explanation, the average conditional probability across symptoms is .25, because only one of the symptoms is associated with each disease, and because each symptom is associated with two diseases (i.e., the single cause and the separate cause). In contrast, for the single cause explanation, the average conditional probabilities P(disease | symptom) across symptoms is .5. Normatively, computing average conditional probabilities is not how probabilities of disjunctions are evaluated. However, in statistical learning tasks (e.g., Aslin et al., 1998; Saffran, Aslin, & Newport, 1996), participants seem to average conditional probabilities and tend to treat them as non-directional (e.g., Endress & Wood, 2011; Pelucchi et al., 2009; Perruchet & Desaulty, 2008; Turk-Browne & Scholl, 2009), perhaps because they rely on simple Hebbian associations (Endress & Johnson, 2021). As a result, a preference for explanations involving fewer causes might emerge if participants evaluate the strength of association between causes and effects.

[3]A second, mutually non-exclusive, explanation is that people tend to prefer more reductive explanations even when the reductive information is irrelevant (Hopkins et al., 2016; Weisberg et al., 2008). For example, non-experts are more satisfied with (bad) explanations of psychological phenomena when irrelevant neuroscience information is added.

to the evidence for Occam's razor thus suggests that, in some cases, people might prefer subjectively simpler theories without explicitly considering their formal complexity. This view might also explain why simplicity might play a major role in the acceptance of scientific theories.

Such an account would combine the role of individual preference in theory choice (e.g., Kuhn, 1977, Chapter 13, pp. 320–339), and the role of learning biases in the cultural transmission of knowledge (e.g., Kirby et al., 2007). Specifically, scientific theories need to be reasonably adequate empirically, but they also need to be intelligible (at least to scientific experts). If experts have similar biases as the lay persons in the studies above, they might also prefer (and preferentially transmit to their students) theories that are subjectively simpler and easier to process. Given that experts in fields from chess (Chase & Simon, 1973) to ornithology (Gauthier et al., 2000) to mathematics (Amalric & Dehaene, 2016) represent information in their field of expertise differently from lay persons, theories that are subjectively simple for experts might still be complex or even unintelligible to lay persons. However, if experts preferentially transmit theories that they can process more fluently, they might promote *subjectively* simpler theories even when there is no obvious formal metric that would quantify the simplicity of the theories.

3 | ARE FORMAL NOTIONS OF SIMPLICITY EMPIRICALLY ADEQUATE?

So far, I suggested that, in every day reasoning, people are not particularly concerned about the simplicity of their inferences, that the simplicity of operations or learning problems does not predict how easily they are processed, and that, when people prefer simpler explanations, this preference might arise because the simpler explanations also happen to be easier to process.

I will now review some prominent formal measures of simplicity that have been proposed to guide learning. I will first consider two related measures of "simplicity" (or rather complexity), one that might be more intuitive for defining the complexity of *operations* to be learned (e.g., grammatical rules) and one more intuitive for the learning of *knowledge structures* (e.g., words). In both cases, I will argue that, in the absence of an exhaustive list of the processing and representational primitives used by human learners, neither form of complexity provides useful guidance towards the relative ease of a learning problem. This conclusion is by no means novel. It has long been known that both the underlying representations and the processing capacities of a machine influence how easily regularities can be learned and represented. For example, a finite state automaton can accept more complex grammars when it operates on structured trees rather than simple strings (e.g., Morgan, 1986; Thatcher, 1973) and both the processing elements and expressive power of a machine have a massive influence on how easily a given grammar can be described (e.g., Kutrib & Pighizzini, 2013; Meyer & Fischer, 1971). Likewise, in the philosophy of science, it is widely recognized that the relative simplicity of two theories depends on the terminology used to describe them (Baker, 2016; Fitzpatrick, 2022).

Based on these observations, I will consider a notion of simplicity that focuses on the simplicity of the inferences, and argue that there is no empirical support for it.

3.1 | Kolmogorov complexity

Regarding the learning of procedures, one way of defining the "simplicity" of a learning problem is to use its Kolmogorov complexity (KC; e.g., Chater, 1996, 1999; Chater & Vitányi, 2003;

Pothos & Chater, 2002). KC is basically the length of the shortest program used to describe an object. But of course, the length of a program depends on the language it is written in. According to proponents of the use of KC as a measure of cognitive complexity, this problem is immaterial because, in the words of Chater and Vitányi (2003), "the choice of programming language does not matter [for computing KC], up to a constant additive factor".

However, this statement is misleading. In fact, the *difference* between the program length in different languages is certainly bounded by a constant; that is, if $K_1(x)$ and $K_2(x)$ are the complexities of some procedure in two languages L_1 and L_2, then $|K_1(x) - K_2(x)| < C$, where C is some constant. The reason is simply that, with general-purpose programming languages, one can always write an emulator of L_2 in L_1 (or vice-versa). Once the emulator is written, any piece of code from L_2 will execute in L_1 and the total length of the code is the length of the original L_2 code plus the length of the code representing the emulator when the latter is needed. The constant is thus essentially the code length of the emulator.

Critically, this result does not guarantee that the two objects have the same *relative* KCs in two languages. For example, if L_1 has operations for multiplication and addition but L_2 only for addition, a program using multiplication will be longer than a program using addition in L_2 (because on has to write the multiplication operation first), but both programs will be equally long in L_1. As a result, in the absence of an exhaustive list of the computational primitives available to human learners, it is impossible to formally determine if one learning problem is easier than another one.

Of course, it is possible to turn this approach around and to determine the available primitives from their relative ease of processing (e.g., Goldsmith & Riggle, 2012; Halle, 1962; Katzir, 2015), but this is not how KC is usually employed.

3.2 | Minimum description length

A closely related measure of complexity is *minimum description length* (MDL; see Rissanen, 2008, for a basic introduction). For example, when fitting a polynomial through a number of points, a higher degree polynomial will fit the points better, so that we need fewer bits to represent the noise (i.e., the errors from the fit). The description length of the data is thus shorter with higher order polynomials. However, we also need to represent polynomials themselves, and the additional terms of the higher order polynomials make their description length longer. MDL thus seeks a compromise between a short description of the data and a short description of the function describing the data.

This approach has been the basis of one of the first computational model of how infants might learn words from fluent speech (Brent & Cartwright, 1996, though one can also use it to simultaneously learn a lexicon and rules that are applied to the lexicon, given that output forms are generated from a lexicon using rules; e.g., Katzir, 2015; Rasin & Katzir, 2016). Fluent speech is a continuous signal, according to many authors, with few consistent cues to word boundaries (e.g., Aslin et al., 1998; Aslin & Newport, 2012; Conway & Christiansen, 2005; Saffran, Newport, & Aslin, 1996; Saffran et al., 1999, but see, e.g., Brentari et al., 2011; Christophe et al., 2003; Endress & Hauser, 2010; Fenlon et al., 2008). As a result, infants need to figure out where words start and where they end before they can learn the meaning of any word.

Brent and Cartwright (1996) used a MDL-based algorithm similar to standard compression techniques. The basic idea is illustrated in Figure 1, where I show three candidate segmentations of the continuous sequence "Thedogbitesthedog". Each candidate segmentation will take up memory

Segmentation type	Words in lexicon	Sentence representation based on this lexicon	# Words in lexicon	# Letters in lexicon	# Words in sentence	Total
Maximal	b, d, e, g, h, i, o, s, t	9-5-3-2-7-4-1-6-9-3-8-9-5-3-2-7-4	9	9	17	**35**
Minimal	Thedogbitesthedog	1	1	17	1	**19**
Intermediate	bites, thedog	2-1-2	2	11	3	**16**

FIGURE 1 Illustration of the *minimum description length* principle for a segmentation of the chunk "The dog bites the dog". The maximal segmentation considers each letter as a word; the minimal segmentation considers the whole input as a single word. Finally, some intermediate segmentations will find recurring units that are larger than single letters and smaller than the entire input. As memory is taken up by representing the items in memory and by representing the sequence, the optimal segmentation that minimizes the total memory storage will be some intermediate segmentation.

space in three ways. First, we need to reserve memory space for each word in the lexicon. Second, we need to populate this memory space with content, and the longer each word, the more space it takes up. Third, we need to represent the sequence using the symbols from the lexicon. For our purposes, the optimal segmentation is the one that minimizes the sum of these memory components, though Brent and Cartwright's (1996) model is mathematically more sophisticated.

In the maximal segmentation, each letter is a word in the lexicon. This gives us nine lexical entries (as there are nine unique letters in the sequence), each of which takes up a single memory unit. Finally, we need 17 units to represent the 17 letters of the sequence, leading to a total memory score of $9+9+17=35$.

In the minimal segmentation, the entire sequence is stored as a unit in the lexicon. If so, we have a single unit in the lexicon that takes up 17 memory units, while the sequence can be represented with a single symbol, yielding a total memory score of $1+17+1=19$.

Finally, in the intermediate segmentation, we postulate the "words" *bites* and *thedog*. We thus have two memory items that take up 11 memory units in total, and that allow us to represent the sequence using only three units. The total memory score is thus $2+11+3=16$, and thus the lowest one of the three possible segmentations.

Brent and Cartwright (1996) showed that such an MDL approach successfully recovers many word boundaries. Brent and Cartwright's (1996) critical conclusion was that, given that the algorithm recovered word boundaries, there must have been distributional information that allowed the algorithm to do so and might also allow infant learners to find word boundaries. Critically, they pointed out that their model was not meant to be a psychologically realistic model of infant learning (it is fundamentally a compression algorithm after all) and that their model demonstrated the availability of distributional information that infants might exploit if they have required processing mechanisms.

While Brent and Cartwright's (1996) model successfully demonstrated that distributional information is available in speech streams, I provide three general reasons for why MDL-based approaches are unlikely to be good guides to learning in terms of psychological processes. First, as in the complexity case, we do not know what the underlying primitives are. To keep with the example of word segmentations, learners of different languages might well rely on different perceptual units (as do their adult counterparts who use stress-based units, syllables or moras, depending on their native language; e.g., Cutler et al., 1986; Cutler et al., 1992; Mehler et al., 1981; Otake et al., 1993). If the underlying units are different, so we will be the results from a MDL-based learner. For example, when learners encounter their sibling's exclamation "Mama, Papa!" (maybe to express their outrage at the suggestion that they might be simplicity-based learners), it is easy to verify from Brent and Cartwright's (1996) cost function

(Equation (5) in Supporting Information SI1) that learners using phonemes as basic units would posit the words *Ma* and *Pa*, while learners using syllables as basic units would be undecided between the words *Mama* and *Papa* and the words *Ma* and *Pa*, respectively. Further, even if learners use the same basic units such as syllables or moras, they might not perceive all units equally well, depending on their native language (e.g., Dupoux et al., 1997; Polka & Werker, 1994; Werker & Tees, 1984), or perceive extra units that are not present in the speech signal (such as epenthetic vowels; see, e.g., Dupoux et al., 1999).

These problems are more general than the word segmentation problem. For example, if learners in the curve-fitting example above have the ability to represent polynomials only up to a degree of three (maybe due to some memory limitations), or if they cannot represent polynomials at all, an optimal polynomial of degree four would simply not be in their learning repertoire. Relatedly, when MDL-based approaches are used to decide among possible grammars, Heinz and Idsardi (2013) argued that "there are multiple ways [certain classes of] languages can be represented [and that] relative length of generalization is not preserved across these formalisms". As a result, without knowing what the underlying representational and processing primitives are, MDL-based approaches do not give us any guidance as to which learning problems might be simpler and which might be harder.

A second problem with MDL-based approaches to learning is that there is no evidence that learners actually optimize the description length. To use the word-segmentation example again, a learner in a word learning experiment might be exposed to a sequence of word repetitions like *dog, dog, dog, ..., pig, pig, pig, ...* Would such a learner extract the words *dog* and *pig*, or some other units such as *dogdog...* and *pigpig...*? In Supporting Information SI1, I show that this depends on the specific familiarization: If each word is repeated N times, the optimal units have length $\sqrt{\frac{N}{3\log_2 5}}$. If each pronunciation of a syllable takes roughly 500 ms, learners should thus extract the units *dogdogdog* and *pigpigpig* after a 1 min familiarization, *dogdogdogdog* and *pigpigpigpig* after a 2 min familiarization, *dogdogdogdogdog* and *pigpigpigpigpig* after a 3 min familiarization, and so on, a prediction that seems implausible at best. As a result, MDL-based approaches per se cannot give us any guidance as to the hypotheses learners consider in the absence of independent evidence for the underlying units and for the ability of learners to perform the relevant optimizations (though again, we might use the output of actual learners to decide between different theories of the underlying representations and learning mechanism; Goldsmith & Riggle, 2012; Halle, 1962; Katzir, 2015).

A final problem with MDL-based approaches is that it is not clear that learners really need to optimize the total memory space they use up, given that they have a massive capacity for declarative memory both in the short-term (e.g., Endress & Potter, 2014; Endress & Siddique, 2016) and in the long-term (e.g., Brady et al., 2008; Standing, 1973; Standing et al., 1970), with an estimate that a typical college-aged adult knows about 50,000 words (Pinker, 1999). Further, to the extent that grammatical knowledge relies on procedural rather declarative memory (e.g., Pinker & Ullman, 2002; Ullman, 2001; Ullman et al., 1997), such memory does not appear to be particularly limited either in a species that manages to learn the motor commands associated with speaking, cycling, swimming and playing the *Fantasia Contrappuntistica*.

3.3 | The size principle

An alternative to defining the simplicity of a learning problem is to consider the restrictiveness of the solutions: If we have to choose between two hypotheses that are equally consistent with

the examples we have seen, we choose the hypothesis that is more restrictive and compatible with fewer potential data points, a strategy called the *size principle* (Tenenbaum & Griffiths, 2001). Learners are thus expected to prefer simpler and less flexible hypotheses (see also Blanchard et al. (2018) above).

A related strategy has been proposed for language acquisition (e.g., Hyams, 1986; Manzini & Wexler, 1987). In some versions of this proposal, humans evolved to acquire language following a sequence of acquisition steps that is consistent with the most restrictive grammar given the input, using specific "triggers" to move from a more restrictive grammar to a more permissive one (e.g., Gibson & Wexler, 1994). The underlying idea is that the triggers allow learners to "conclude" that their current grammars are not general enough, and to adjust them appropriately, while it is unclear how they could even notice that they started out with a grammar that is too general (see below for an interpretation of "triggers" in terms of developmental phenotypic plasticity). More generally, an inference strategy that starts with simpler hypotheses and accepts more complex hypotheses only when necessary is efficient in the sense that it minimizes the number of times an incorrect hypothesis is adopted, and the working hypothesis needs to be changed (Kelly, 2007a, 2007b).

In the literature following Tenenbaum and Griffiths (2001), this idea has been applied much more widely in domains ranging from basic probabilistic inference to language acquisition to social cognition. If these demonstrations are convincing, the size principle might provide a simplicity-based metric with which learning problems can be evaluated *in general*. However, I will now discuss some of the strongest evidence for the size principle in the domains of word learning, rule learning and probabilistic inference, and argue that these demonstrations provide good examples of what Glymour (2007) called "Ptolemaic Psychology", and that their success relies on numerous auxiliary assumptions (see, e.g., Jones & Love, 2011; Marcus & Davis, 2013; Sakamoto et al., 2008, for related criticisms). I will further argue that common-sense psychology provides much more straightforward accounts even though it does not provide any general metric for the evaluation of learning problems.

3.3.1 | The size principle and word learning (1)

Some of the strongest evidence for the size principle comes from Xu and Tenenbaum's (2007b) experiments on word learning (see also Navarro et al., 2012). They asked how learners assign meaning to novel nouns, and under what condition they would choose a meaning at the subordinate category level (e.g., "Dalmatian"), at the basic-level category level (e.g., "dog"), or at the superordinate category level (e.g., "animal").

Participants were presented with a novel word (e.g., "fep"), and shown one or three examples of the word's meaning (e.g., a Dalmatian). Following this, they were shown a test screen with potential examples of "feps", and had to select other feps.

The test screen comprised eight items of each of three superordinate categories (i.e., animals, vegetables, and vehicles). Within each category, there were two examples of the same subordinate category (e.g., two other Dalmatians), two examples of the same basic-level category (e.g., two non-Dalmatian dogs), and four examples of the same superordinate category (e.g., four non-dog animals). If participants infer that fep means "animal", they should choose all eight pictures of that category; if they infer that it means "dog", they should choose the four corresponding pictures; and if they conclude that it means "Dalmatian", they should choose the two Dalmatians only.

Results showed that, when familiarized with three Dalmatians, participants concluded that fep meant "Dalmatian". When familiarized with one Dalmatian and two other dogs, they concluded that fep meant "dog"; and when familiarized with one Dalmatian and two non-dog animals, they concluded that fep meant "animal".

These results are certainly consistent with the size principle. After all, there are more animals in the world than there are dogs, and there are more dogs than there are Dalmatians. Hence, if learners opt for the most restrictive inference, they should opt for a subordinate meaning when this it is consistent with the data, as this is the most restrictive one, and chose basic level or superordinate levels only when required by the data.

However, it is almost certain that learners do not apply the size principle, simply because it does not scale up to natural language acquisition. In fact, natural language learners are rarely shown test screens explicitly providing them with the nine possible meanings of a novel word as well as with the number of elements of each category, raising the question of how learners could possibly estimate the number of elements of a category—and even if they had access to this information, whether they could process it. In fact, there are an estimated 70 million dogs in the United States (American Veterinary Medical Society 2012 U.S. Pet Ownership & Demographics Sourcebook, retrieved on August 16, 2019 from https://www.avma.org/KB/Resources/Statistics/Pages/Market-research-statistics-US-pet-ownership.aspx). It is an entirely open question whether infants can process numbers of this magnitude, or which other information they might possibly exploit according to Xu and Tenenbaum (2007b).

In contrast, these results follow directly from standard approaches to word learning (e.g., Medina et al., 2011; Stevens et al., 2017): Learners acquire a word whenever they encounter a situation that is conducive for inferring its meaning. If subsequent occurrences are consistent with this guess, they stick with it, and revise it if they are not. By default, learners might assume that word meanings correspond to a basic-level category (e.g., Markman & Hutchinson, 1984; Waxman & Markow, 1995). If they see an example that is not part of the basic level category (e.g., a cat, which is an animal but not a dog), they might revise their guess and opt for a superordinate category interpretation, similar to how the "triggers" discussed above (e.g., Gibson & Wexler, 1994) might lead learners to move to more general hypotheses. Conversely, if they consistently see "bad" examples of a basic-level category (e.g., because Dalmatians are presumably not particularly prototypical dogs; Emberson et al., 2019), or if the variability of the exemplars is less than expected from a basic-level category (e.g., because learners are not shown the best examples), learners might opt for a subordinate interpretation. This model thus accounts for Xu and Tenenbaum's (2007b) data, without making any use of the size principle. It also makes testable predictions: Learners should be less willing to entertain a subordinate interpretation when the examples are more prototypical for the basic level category even if they all come from the same subordinate category. For example, if they are shown German shepherds rather than Dalmatians, a subordinate interpretation might be less available.

That being said, Xu and Tenenbaum (2007b) report another result that, at first sight, seems to provide strong evidence for the size principle. Specifically, they show that, when participants are familiarized with a single Dalmatian, they conclude that fep means "dog"; in contrast, when familiarized with three Dalmatians, they conclude that fep means "Dalmatian". However, as mentioned above, participants might have a tendency to use basic level categories, and conclude that the lone Dalmatian is an exemplar of the "dog". When shown three Dalmatians, however, they might be surprised that all of the dogs are Dalmatian

(e.g., because the examples are particularly prototypical dogs), and change their inference accordingly.[4]

3.3.2 | The size principle and word learning (2)

Like Xu and Tenenbaum (2007b), Xu and Tenenbaum (2007a) asked how learners choose between subordinate category and basic-level meanings. Participants (adults and three-to-four-year-olds) were presented with a display showing items from two categories. The categories were defined by their shapes and were spatially grouped together. Each category comprised three subordinate categories of four items each. Items in different subordinate categories shared their shape, but differed in texture.

Learning occurred in one of two conditions. In the teacher-driven condition, the experimenter pointed to three items from the same subordinate category and labeled them with a novel word (e.g., a blicket). In the learner-driven condition, the experimenter labeled only one object, and then encouraged participants to point to two *other* blickets. Results showed that all but one participant pointed to two other items from the same subordinate category.

Following this, participants were pointed to five more objects and asked whether these were blickets as well. Results showed that participants in the teacher-driven condition were more likely to infer that "blicket" referred to the subordinate category than participants in the learner-driven condition. Xu and Tenenbaum's (2007a) explain the preference for the subordinate category in the teacher-driven condition as follows. Since the teacher knows the meaning of the word, she will choose example objects to which this meaning applies. Hence, due to the size principle, if all examples are consistent with a subordinate meaning, a subordinate meaning should be preferred, because the number of items in that category is lower than that in a basic-level category. In the learner-driven condition, in contrast, the learner does not know the meaning of the word; hence, the size principle does not apply to favor smaller categories, such that the learner should be less likely to infer a subordinate-level meaning than in the teacher-driven condition.

At first sight, Xu and Tenenbaum's (2007a) assumption that the size principle applies only in the teacher-driven condition seems to contradict Xu and Tenenbaum's (2007b) assumption

[4]In contrast, according to Xu and Tenenbaum's (2007b) formal explanation, the likelihood of each category given all exemplars is the product of the likelihoods of each category given the individual exemplars. Importantly, one of the factors in the individual likelihoods comes from the size principle, and is inversely proportional to the number of items in the category. As a result, with more exemplars, the influence of the set size is more pronounced, and should favor smaller categories (e.g., subordinate categories if these are consistent with the data). However, this is true for any probability smaller than one that is raised to a power corresponding to the number of exemplars. To make this point, I show in Supporting Information SI2 that Xu and Tenenbaum's (2007b) results can also be explained based on the similarity between the training items and the test items. If the similarity score is converted to a probability score, the same qualitative predictions follow. That is, the a priori likelihood of a meaning is largest when the similarity between the training examples and the other items to which the meaning applies is largest; as such, this probability score favors small (i.e., subordinate) categories. Hence, when it is raised to the third power (due to the three examples), the preference for the subordinate category will be more pronounced. As a result, what guarantees the narrowing of the generalizations between one and three exemplars is not the size principle but rather the rest of Xu and Tenenbaum's (2007b) formalism.A similar conclusion applies to the results reported by Navarro et al. (2012). "Narrowing" of the inferences occurs simply due to raising probabilities to a power corresponding to the number of examples, but not due to the size principle per se. As such, neither Xu and Tenenbaum's (2007b) nor Navarro et al.'s (2012) results provide evidence for computations involving the size principle.

that the size-principle does apply in the absence of a teacher. Maybe more critically, if it is true that, for the size principle to apply, one needs an experimenter (or caretaker) who patiently points to all instances in the extension of a label, it would seem questionable whether the size principle has any relevance for language acquisition at all. Further, the size principle would not apply to the acquisition of verbs—for which deictic reference is much less felicitous and much more ambiguous than for nouns (e.g., "Look at this, that's digesting!")

Leaving aside this caveat, Xu and Tenenbaum's (2007a) results are actually inconsistent with their conclusions. In fact, all but one participant in the learner-driven condition selected items from the same subordinate category when asked to find other blickets; given that Xu and Tenenbaum's (2007a) model predicts that participants in the learner-driven condition should favor a basic-level interpretation, and that there are more than twice as many candidate blickets from different subordinate categories, one would expect them to preferentially choose items from different subordinate categories.

However, there is a simple alternative interpretation. Xu and Tenenbaum's (2007a) claims notwithstanding, participants clearly have a tendency to choose a subordinate-level interpretation, maybe because they are presented with novel non-sense objects that might not be readily assigned to conceptual categories (Callanan et al., 1994). In the teacher-driven condition, participants might just stick with this interpretation. In the learner-driven condition, in contrast, they might opt for the basic-level interpretation for purely pragmatic reasons after they initially chose a subordinate interpretation. As mentioned above, participants were asked to decide which *other* objects were blickets only after they had (correctly) identified two further blickets. Plausibly, the teacher pointing to further objects, asking whether they were blickets as well, might have given participants the impression that their initial interpretation (at the subordinate level) was not general enough, and that the experimenter expected a (more general) basic-level interpretation. If so, the difference between the teacher-driven condition and the learner-driven condition might be due to pragmatic factors that are unrelated to the size principle and rather reflect the participants' beliefs about the teacher's communicative intentions.

3.3.3 | The size principle and rule-learning

Frank and Tenenbaum (2011) use the size principle to explain how infants might learn the repetition-patterns discussed above. They propose that, when infants have to choose between multiple regularities that are consistent with examples they have heard, they choose the one that has fewer potential items conforming to it.

According to their model, infants might encounter a total of three syllables. Before encountering any syllable triplet, infants know that the three syllables allow for a total 27 triplets, that six of these triplets follow an ABB pattern (e.g., *pu-li-li*), that three of these triplets follow an AAA pattern (where all three syllables are identical), as well as the number of triplets that would conform to any conceivable rule. They then use the number of triplets that are consistent with each rule to choose among possible generalizations.

Frank and Tenenbaum (2011) applied this size-principle-based model to a variety of infant rule-learning experiments. However, in addition to the prima facie implausibility of the model, Endress (2013) showed that the models made incorrect predictions (e.g., that a change from human syllables to monkey vocalizations should be *less* salient than a relatively subtle change from AAB patterns to ABB patterns), assumed that infants can process about 900 triplets per second, made predictions that were subsequently refuted (Gervain & Endress, 2017), assumed that infants have severe

perceptual problems in some phases of an experiment and perfect perception in other phases, used model parameters that led their model to contradict the experimental data when the parameters were used in psychologically meaningful ways or wired in the phenomenon they sought to explain (Endress, 2013; see Frank (2013) and Endress (2014) for discussion). It thus seems that an account based on the size principle is unlikely to explain infant rule learning.

3.3.4 | The size principle and probabilistic inference

Gweon et al. (2010) presented 15-months-olds with a transparent box containing blue and yellow balls (see also Denison et al., 2013). The experimenter then removed a variable number of *blue* balls from the box and demonstrated that they squeaked upon squeezing them. Following this, infants were handed a *yellow* ball. Gweon et al. (2010) asked how likely infants were to conclude that this ball squeaked as well. The dependent measure was whether, and how often, infants would squeeze the yellow ball.

The critical manipulations were (1) how many balls the experimenter picked from the box (1 or 3) and (2) whether the majority of the balls in the box was blue or yellow: In blue majority populations, 75% of the balls were blue, while, in yellow majority populations, only 25% of the balls were blue.

When the experimenter extracted *three* blue balls from a blue majority population, infants squeezed the yellow ball more often than when the blue balls came from a yellow majority population. In contrast, when only a *single* blue ball was extracted from a yellow majority population, infants squeezed the yellow ball as much as when three blue balls were extracted from a blue majority population. In a crucial control condition, three blue balls were ostensibly drawn by chance from yellow majority population. In that condition, infants did not suppress squeezing the yellow ball.

To explain their data, Gweon et al. (2010) proposed that infants consider the four possibilities spanned by two factors: (i) Is the teacher cooperative and picks the balls only from the squeaky ones, or is she evil, and picks from all balls irrespectively of squeakiness? (ii) Are all balls squeaky, or only the blue ones? Infants then compute likelihoods of the results of the experimenter's actions according to all four possible scenarios, and compare these likelihoods to decide whether or not to squeeze the yellow ball. This likelihood ratio is given by, with β being the proportion of blue balls (and α being a parameter that is irrelevant for the current purposes):

$$L_\alpha = \frac{\beta^n}{\alpha + (1-\alpha)\beta^n}, \quad \alpha \in [0, 1], \ \beta \in]0,1]. \tag{1}$$

Should infants ever squeeze yellow balls?

Irrespective of whether it is plausible that infants have the processing abilities to deal with such a complex model, it is inconsistent with the data, for two reasons. First, the model always concludes that it is more likely that only blue balls are squeaky. In fact, it is easy to see that $L_\alpha = 1$ for $\alpha = 0$ or $\beta = 1$, and that $L_\alpha < 1$ for all $\alpha > 0$ and $\beta < 1$. Hence, Gweon et al.'s (2010) model predicts that infants should *never* squeeze the yellow ball at all.[5]

[5]It is easy to see that, irrespective of the proportion of blue balls, this effect should be more pronounced when more balls are drawn from the container, and that, eventually L_α goes to 0. Indeed, the partial derivative $\partial_n L_\alpha = \frac{\alpha\beta^n \ln\beta}{(\alpha+(1-\alpha)\beta^n)^2}$ is strictly smaller than zero since $\ln\beta$ is smaller than zero for $\beta < 1$. Further, $\lim_{n\to\infty} L_\alpha = 0$.

Are blue balls more fun than yellow balls?

Second, the model assumes that it is more fun to squeeze squeaky blue balls than to squeeze squeaky yellow balls, and that infants thus desire to find *blue* squeaky balls. A plausible alternative hypothesis is that infants are interested in squeaky balls irrespective of their color. As shown in in Figure 2, a version of Gweon et al.'s (2010) model that assumes that infants care about squeaky balls *irrespective* of their color predicts that infants should be more likely to squeeze the yellow ball in *yellow* majority populations, which is just the situation where infants are *less* likely to squeeze it. (However, in contrast to Gweon et al.'s (2010) model, the new model accounts for the fact infants squeeze the yellow ball in the first place.)

Specifically, given a proportion β of blue balls, one can derive different likelihoods for the teacher picking three squeaky blue balls. These likelihoods are presented in Figure 3. The two middle columns show the likelihoods from Gweon et al.'s (2010) model, assuming that infants desire *blue* squeaky balls. The two right-most columns show the likelihoods assuming that infants care about squeakiness but not about color. The likelihood of picking n squeaky blues balls can be obtained by averaging across the choice strategies of the teacher (see Gweon et al., 2010, for a justification of this average).

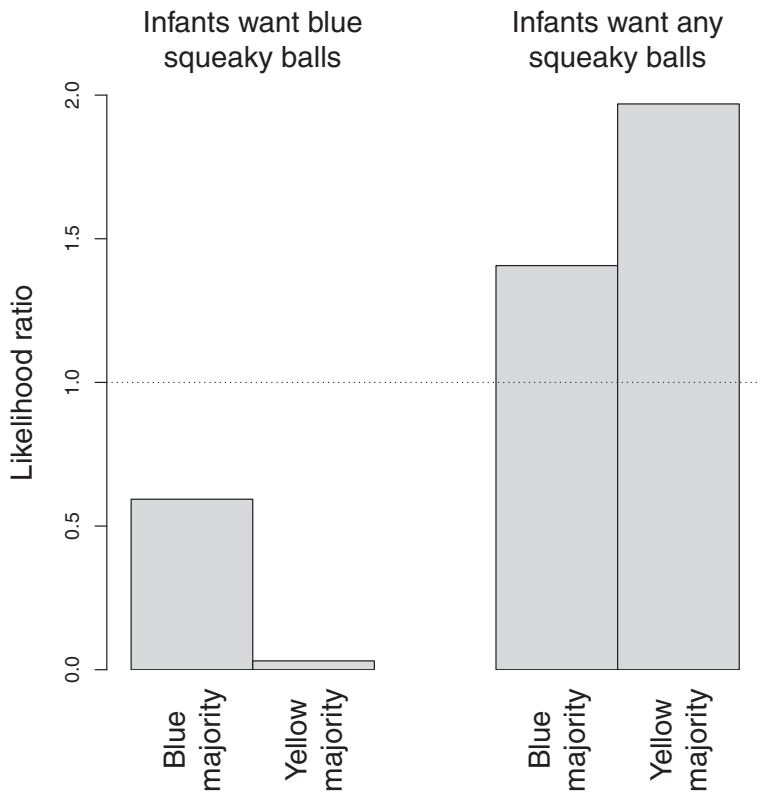

FIGURE 2 Predictions of Gweon et al.'s (2010) original model assuming that infants like to find squeaky *blue* balls (left) and a modified model where infants are interested in squeaky balls irrespective of color (right) when three balls are extracted from the container. In Gweon et al.'s (2010) model, infants are more likely to squeeze a yellow ball in when the container holds a majority of blue balls; in the modified model, infants are more likely to squeeze a yellow ball when the container holds a majority of yellow balls.

Squeaky balls	Teacher chooses among	Infants want *blue* squeaky balls: Explanation	*P*	Infants want squeaky balls: Explanation	*P*
only blue	squeaky	The hypothesis holds that only blue balls are squeaky, and that the teacher will only sample from these balls. Hence, according to this hypothesis, the teacher will choose squeaky blue balls with probablity 1.	1	The hypothesis holds that the teacher will only sample from squeaky balls. Hence, according to this hypothesis, the teacher will choose squeaky balls with probablity 1.	1
only blue	all	Since the teacher randomly picks balls, she has, for each ball, a chance β for picking a blue/squeaky ball.	β^n	Since the teacher randomly picks balls, she has, for each ball, a chance β for picking a blue/squeaky ball.	β^n
only blue	average		$(1+\beta^n)/2$		$(1+\beta^n)/2$
all	squeaky	While all balls are squeaky, the teacher has a chance of β to pick a blue ball.	β^n	Since all balls are squeaky, the probability of picking n squeaky balls is 1.	1
all	all	While all balls are squeaky, the teacher has a chance of β to pick a blue ball.	β^n	Since all balls are squeaky, the probability of picking n squeaky balls is 1.	1
all	average		β^n		1

FIGURE 3 Likelihoods of the teacher picking n balls of interest out of a box with a proportion of β blue balls. The middle two columns present Gweon et al.'s (2010) model, in which infants seek blue squeaky balls. The likelihood ratio in favor of the hypothesis that all balls are squeaky is given by $2\beta^n/(1+\beta^n)$. The rightmost columns present an alternative model, where infants just care about the squeakiness of the balls, irrespective of their color. In that case, the likelihood ratio is $2/(1+\beta^n)$.

Gweon et al. (2010) compared the ratio of (i) the likelihood of the teacher picking three squeaky blue balls if all balls are squeaky and (ii) the likelihood of the teacher picking three squeaky blue balls when only the blues one are squeaky. As mentioned above, this ratio is given by:

$$L=\frac{2\beta^n}{1+\beta^n}. \tag{2}$$

(Compared to Equation (1), and following Gweon et al. (2010), I set α to .5) When 75% of the balls are blue, and the teacher picks three balls, the ratio is .59; when 25% of the balls are blue, the ratio is .03. Hence, leaving aside the fact that the model predicts that infants should never squeeze yellow balls, they should be more likely to squeeze yellow balls when 75% of the balls are blue.

The alternative model, where infants care only about the squeakiness of the balls but not their color, reverses the predictions. As can be seen from Figures 2 and 3, the corresponding likelihood ratio is (with α set to 0.5):

$$L^{'}=\frac{2}{1+\beta^n}. \tag{3}$$

Keeping Gweon et al.'s (2010) α parameter as a variable, the likelihood ratio is given by:

$$L^{'}_{\alpha}=\frac{1}{\alpha+(1-\alpha)\beta^n}. \tag{4}$$

It is easy to see that $L^{'}_{\alpha}=1$ for $\alpha=1$ or $\beta=1$, and $L^{'}_{\alpha}>1$ for $\alpha<1$ and $\beta<1$. Hence, this model accounts for the fact that infants have a tendency to squeeze balls irrespective of their color.[6]

[6]In contrast to Gweon et al.'s (2010) model, the partial derivative $\partial_n L^{'}_{\alpha}=\frac{-(1-\alpha)\beta^n\ln\beta}{(\alpha+(1-\alpha)\beta^n)^2}$ is strictly positive for $\beta\in\ ]0,1[$ and $\alpha\neq 1$. Further, $\lim_{n\to\infty}L^{'}_{\alpha}=1/\alpha>1$.

However, for blue majority populations (with 75% of blue balls), the ratio is 1.41, while it is 1.97 for yellow majority populations (with 25% of blue balls). Hence, the alternative model (incorrectly) predicts that infants should be more likely to squeeze yellow balls when 25% of the balls are blue, though it does account for the fact that infant squeeze the yellow balls to begin with.

A simple psychological explanation of ball squeezing behavior

The considerations above suggest that the "success" of Gweon et al.'s (2010) model is not due to the size principle, but rather to extraneous assumptions about what infants are most interested in: They need to assume that squeaky blue balls are more fun than squeaky yellow balls.

In contrast to this model, there is a much simpler explanation. By default, infants might tend to squeeze the balls, because shape is a better predictor of function (i.e., squeaking) than color (e.g., Bloom, 1996; Brown, 1990; Hauser, 1997), and because squeezing them does not entail a huge cost. However, infants can also detect non-random behavior of an agent; they know that drawing three blue balls out of a box of mostly yellow balls is unlikely (Téglás et al., 2007) and can use this ability to detect non-random behavior in agents (e.g., Kushnir et al., 2010; though they sometimes expect random behavior from agents as well; Tauzin & Gergely, 2019). Further, infants know that humans are often communicative and might even courageously attempt to "teach" them (e.g., Csibra & Gergely, 2009). Hence, they might detect the non-random behavior of the agent, assume that the agent has a reason to behave in non-random ways, and, in line with the idea that infants and children are more likely to imitate when they perceive intentional behavior (e.g., Gergely et al., 2002; Lyons et al., 2007), imitate her more closely only in the condition where the agent shows clear non-random behavior. This idea accounts for all of Gweon et al.'s (2010) data. A similar account applies to Denison et al.'s (2013) data.

Taken together, these data thus do not provide support for another simplicity-based guide to learning: The size principle. Rather, they have alternative explanations based on simple psychological considerations, raising the question of what might constitute an empirically adequate guide to learning.

4 | CONCLUSIONS

In this paper, I first reviewed data suggesting that people do not necessarily draw the simplest inferences and that, generally speaking, the formal complexity of an operation is not necessarily a good measure of what is easy to process or learn for actual humans. When humans prefer "simpler" explanations of phenomena, they might do so because the simpler explanations are also easier to process. I then showed that prominent formal metrics of simplicity are unlikely to help in this respect. Claims to the contrary notwithstanding, approaches that rely on minimizing the code or description length of a problem crucially depend on assumptions about the underlying representational and processing primitives. Further, as shown by the empirical examples above, even when we have a sense about what the underlying computational primitives might be, certain operations are easier in some domains than in others, again for no obvious formal reason. Finally, approaches that quantify the simplicity of the inferences (e.g., the size principle) have, to the best of my knowledge no empirical support, strongly depend on sometimes arbitrary assumptions, and sometimes make incorrect predictions. As a result, there

does not seem to be a context-free, generally applicable notions of simplicity that explains the behavior of actual people.

As mentioned above, this conclusion also follows from purely theoretical considerations, given that the simplicity of a description depends on the representational primitives, the processing units and the representational strategies (e.g., Baker, 2016; Fitzpatrick, 2022; Kutrib & Pighizzini, 2013; Meyer & Fischer, 1971; Morgan, 1986; Thatcher, 1973). It also reflects debates in the philosophy of science about the definition, justification and historical relevance of simplicity in the practice of science (Baker, 2016; Fitzpatrick, 2022).

These results might also have implications for why simpler theories might be preferred in the history of science even when their simplicity is hard to quantify or even to justify (e.g., Baker, 2016; Fitzpatrick, 2022). In line with other accounts of cultural evolution (e.g., Kirby et al., 2007), scientists might prefer theories they can process more easily (though their mental representations of their domain of expertise likely differ from those of lay people; e.g., Amalric & Dehaene, 2016; Chase & Simon, 1973; Gauthier et al., 2000), *claim* that the easier theories are also simpler, and preferentially teach and transmit them. As a result, the scientists' learning and processing biases might shape the kinds of theories that will eventually find acceptance; further, these theories will appear simpler if they are easier to process.

Together, these arguments thus suggest that, for any empirically viable notion of "simplicity", we need to determine empirically what the underlying representational, learning and processing primitives are before any notion of simplicity can be fruitfully deployed. Once (and if) such a list of primitives becomes available, it will become possible to calculate the complexity of a learning or processing problem for an actual biological learner.

From an evolutionary point of view, such conclusions are utterly unsurprising. Just as it is impossible to know that a geocentric model of the solar system is computationally efficient (even when "simpler" heliocentric models are available) without knowing the processing constraints of Hellenistic astronomers (Russo, 2000), it is impossible to know what is simple for humans to infer and process without knowing the (evolutionary) history of their representational and processing abilities. Inductive biases might not follow simplicity prescriptions when considered by themselves. Rather, they might be those inductive biases that an organism with our evolutionary history happens to have, just as any other trait needs to evolve within the constraints of an organism's evolutionary history (e.g., Gould et al., 1979). In other words, while simplicity is often considered a guiding principle to science itself (e.g., Baker, 2016; Fitzpatrick, 2022; Goodman, 1943), the traits of biological systems cannot easily be predicted based on simple first principles: They result from an interplay of natural selection, evolutionary constraints, trade-offs, and mere accidents (e.g., Gould et al., 1979). Unless one is willing to entertain an evolutionary dualism, positing different kinds of evolutionary mechanisms for morphological traits on the one side and mental and behavioral traits on the other side, there is no a priori reason to consider simplicity a core feature of mental representations or inductive biases.

This is not to say that inductive biases have not evolved to acquire language. If so, they might be particularly conducive for learning the regularities that need to be learned given the (linguistic) environment in which they are learned (i.e., the input), just as we preferentially learn ecologically relevant associations.

An extreme version of the adaptive view is that learners have highly constrained feature detectors for certain abstract patterns, learn by matching their input to these patterns and reconfigure their representations according to these matches. Biologically speaking, this idea is similar to developmental phenotypic plasticity, where an organism "chooses" among

phenotypes depending on environmental stimuli (e.g., Alcock, 2001). For example, major phenotypic options such as sex are determined by the incubation temperature of the eggs in many reptiles (e.g., Shine, 1999). Mutatis mutandis, the "incubation language" might provide learners with trigger stimuli that lead them to adopt certain grammatical options, an idea seems rather similar to principles and parameter approaches (e.g., Baker, 2001; Huang & Roberts, 2016).

Irrespective of whether this speculation is true, the examples above show that, in some cases, models with epicycles are the better models: Just as the Ptolemaic model led to easier calculations for Hellenistic astronomers, the human mind needs to make do with the computational machinery it happens to come equipped with, and computations are simple when they match the available machinery.

ACKNOWLEDGEMENT

I am grateful to R. Katzir for helpful comments on an earlier version of this manuscript.

DATA AVAILABILITY STATEMENT

There are no data available.

ORCID

Ansgar D. Endress https://orcid.org/0000-0003-4086-5167

REFERENCES

Alberts, J. R., & Gubernick, D. J. (1984). Early learning as ontogenetic adaptation for ingestion by rats. *Learning and Motivation*, *15*(4), 334–359. https://doi.org/10.1016/0023-9690(84)90002-X

Alcock, J. (2001). *Animal behavior: An evolutionary approach* (7th ed.). Sinauer Associates.

Amalric, M., & Dehaene, S. (2016). Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 4909–4917. https://doi.org/10.1073/pnas.1603205113

Armitage, P. (2008). Planetary formation and migration. *Scholarpedia*, *3*(3), 4479. Retrieved from http://www.scholarpedia.org/article/Planetary_formation_and_migration (revision #186389). https://doi.org/10.4249/scholarpedia.4479

Aslin, R. N., & Newport, E. L. (2012). Statistical learning. *Current Directions in Psychological Science*, *21*(3), 170–176. https://doi.org/10.1177/0963721412436806

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Baker, A. (2016). Simplicity. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University Retrieved from https://plato.stanford.edu/archives/win2016/entries/simplicity/

Baker, M. (2001). *The atoms of language*. Basic Books.

Baylis, G. C., & Driver, J. (1994). Parallel computation of symmetry but not repetition in single visual objects. *Visual Cognition*, *1*, 337–400.

Baylis, G. C., & Driver, J. (2001). Perception of symmetry and repetition within and across visual shapes: Part-descriptions and object-based attention. *Visual Cognition*, *8*(2), 163–196.

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition*, *64*(1), 39–72.

Blanchard, T., Lombrozo, T., & Nichols, S. (2018). Bayesian occam's razor is a razor of the people. *Cognitive Science*, *42*, 1345–1359. https://doi.org/10.1111/cogs.12573

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, *60*(1), 1–29.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(8), 451–459.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2007). On consonants, vowels, chickens, and eggs. *Psychological Science*, *18*(10), 924–925. https://doi.org/10.1111/j.1467-9280.2007.02002.x

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1–2), 93–125.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, *54*(1), 49–72.

Brown, A. L. (1990). Domain-specific principles affect learning and transfer in children. *Cognitive Science*, *14*(1), 107–133.

Bruce, V. G., & Morgan, M. J. (1975). Violations of symmetry and repetition in visual patterns. *Perception*, *4*(3), 239–249.

Callanan, M. A., Repp, A. M., McCarthy, M. G., & Latzke, M. A. (1994). Children's hypotheses about word meanings: Is there a basic level constraint? *Journal of Experimental Child Psychology*, *57*(1), 108–138. https://doi.org/10.1006/jecp.1994.1006

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566–581.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *52*(2), 273–302.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics*, *31*(3–4), 585–598. https://doi.org/10.1016/S0095-4470(03)00040-8

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Corballis, M. C., & Roldan, C. E. (1974). On the perception of symmetrical and repeated patterns. *Perception & Psychophysics*, *16*(1), 136–142.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Culy, C. (1985). The complexity of the vocabulary of bambara. *Linguistics and Philosophy*, *8*(3), 345–351. https://doi.org/10.1007/BF00630918

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, *24*, 381–410.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of french and english. *Journal of Memory and Language*, *25*(4), 385–400.

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Fery, & R. van de Vijver (Eds.), *Variation and gradience in phonetics and phonology* (pp. 265–309). Mouton De Gruyter.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, *111*(3), 378–382. https://doi.org/10.1016/j.cognition.2009.02.010

de la Mora, D. M., & Toro, J. M. (2013). Rule learning over consonants and vowels in a non-human animal. *Cognition*, *126*(2), 307–312. https://doi.org/10.1016/j.cognition.2012.09.015

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: Evidence from 4.5- and 6-month-olds. *Developmental Psychology*, *49*(2), 243–249. https://doi.org/10.1037/a0028278

Domjan, M. (1983). Biological constraints on instrumental and classical conditioning: Implications for general process theory. *Psychology of Learning and Motivation*, *17*, 215–277. https://doi.org/10.1016/S0079-7421(08)60100-0

Domjan, M. (2015). The Garcia-Koelling selective association effect: A historical and personal perspective. *International Journal of Comparative Psychology*, *28*. https://escholarship.org/uc/item/5sx993rm

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in japanese: A perceptual illusion? *Journal of Experimental Psychology. Human Perception and Performance*, *25*(6), 1568–1578. https://doi.org/10.1037/0096-1523.25.6.1568

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A destressing "deafness" in french? *Journal of Memory and Language*, *36*(3), 406–421. https://doi.org/10.1006/jmla.1996.2500

Emberson, L. L., Misyak, J. B., Schwade, J. A., Christiansen, M. H., & Goldstein, M. H. (2019). Comparing statistical learning across perceptual modalities in infancy: An investigation of underlying learning mechanism(s). *Developmental Science*, *22*(6), e12847. https://doi.org/10.1111/desc.12847

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Acta Psychologica*, *135*(2), 182–190. https://doi.org/10.1016/j.actpsy.2010.06.005

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176. https://doi.org/10.1016/j.cognition.2012.11.014

Endress, A. D. (2014). How are Bayesian models really used? *Cognition*, *130*(1), 81–84. https://doi.org/10.1016/j.cognition.2013.09.003

Endress, A. D. (2019). Duplications and domain-generality. *Psychological Bulletin*, *145*(12), 1154–1175. https://doi.org/10.1037/bul0000213

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614. https://doi.org/10.1016/j.cognition.2006.12.014

Endress, A. D., & Hauser, M. D. (2009). Syntax-induced pattern deafness. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(49), 21001–21006. https://doi.org/10.1073/pnas.0908963106

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199. https://doi.org/10.1016/j.cogpsych.2010.05.001

Endress, A. D., & Johnson, S. P. (2021). When forgetting fosters learning: A neural network model for statistical learning. *Cognition*, *213*, 104621. https://doi.org/10.1016/j.cognition.2021.104621

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353. https://doi.org/10.1016/j.tics.2009.05.005

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology-General*, *143*(2), 548–565. https://doi.org/10.1037/a0033934

Endress, A. D., & Siddique, A. (2016). The cost of proactive interference is constant across presentation conditions. *Acta Psychologica*, *170*, 186–194. https://doi.org/10.1016/j.actpsy.2016.08.001

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology*, *63*(3), 141–171. https://doi.org/10.1016/j.cogpsych.2011.07.001

Feigenson, L. (2008). Parallel non-verbal enumeration is constrained by a set-based limit. *Cognition*, *107*(1), 1–18. https://doi.org/10.1016/j.cognition.2007.07.006

Fenlon, J., Denmark, T., Campbell, R., & Woll, B. (2008). Seeing sentence boundaries. *Sign Language & Linguistics*, *10*(2), 177–200.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(24), 15822–15826. https://doi.org/10.1073/pnas.232472899

Fitzpatrick, S. (2022). Simplicity in the philosophy of science. In *The internet encyclopedia of philosophy* Retrieved from https://iep.utm.edu/simplici/

Frank, M. C. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to endress. *Cognition*, *128*(3), 417–423. https://doi.org/10.1016/j.cognition.2013.04.010

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371. https://doi.org/10.1016/j.cognition.2010.10.005

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language Linguistic Theory*, *22*(1), 179–228.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1974). Behavioral regulation of the milieu interne in man and rat. *Science*, *185*(4154), 824–831.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1976). Flavor aversion studies. *Science*, *192*, 265–267.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science*, *4*(1), 123–124. https://doi.org/10.3758/BF03342209

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–197. https://doi.org/10.1038/72140

Gemberling, G. A., & Domjan, M. (1982). Selective associations in one-day-old rats: Taste-toxicosis and texture-shock aversion learning. *Journal of Comparative and Physiological Psychology*, *96*, 105–113.

Gemberling, G. A., Domjan, M., & Amsel, A. (1980). Aversion learning in 5-day-old rats: Taste-toxicosis and texture-shock associations. *Journal of Comparative and Physiological Psychology*, *94*, 734–745.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755. https://doi.org/10.1038/415755a

Gervain, J., & Endress, A. D. (2017). Learning multiple rules simultaneously: Affixes are more salient than reduplications. *Memory and Cognition*, *45*(3), 508–527. https://doi.org/10.3758/s13421-016-0669-9

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature*, *410*(6831), 930–933. https://doi.org/10.1038/35073582

Glymour, C. (2007). Bayesian Ptolemaic psychology. In W. Harper & G. Wheeler (Eds.), *Probability and inference: Essays in honour of Henry E. Kyburg, Jr.* London College Publications.

Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: The case of finnish vowel harmony. *Natural Language & Linguistic Theory*, *30*(3), 859–896. https://doi.org/10.1007/s11049-012-9169-1

Goodman, N. (1943). On the simplicity of ideas. *Journal of Symbolic Logic*, *8*(4), 107–121. https://doi.org/10.2307/2271052

Gould, S. J., Lewontin, R. C., Maynard Smith, J., & Holliday, R. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *205*(1161), 581–598. https://doi.org/10.1098/rspb.1979.0086

Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003

Gubernick, D. J., & Alberts, J. R. (1984). A specialization of taste aversion learning during suckling and its weaning-associated transformation. *Developmental Psychobiology*, *17*, 613–628. https://doi.org/10.1002/dev.420170605

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9066–9071. https://doi.org/10.1073/pnas.1003095107

Halle, M. (1962). Phonology in generative grammar. *WORD*, *18*(1–3), 54–72. https://doi.org/10.1080/00437956.1962.11659765

Hauser, M. D. (1997). Artifactual kinds and functional design features: What a primate understands without language. *Cognition*, *64*(3), 285–308.

Hauser, M. D., & Glynn, D. (2009). Can free-ranging rhesus monkeys (Macaca mulatta) extract artificially created rules comprised of natural vocalizations? *Journal of Comparative Psychology*, *123*(2), 161–167. https://doi.org/10.1037/a0015584

Heinz, J., & Idsardi, W. (2013). What complexity differences reveal about domains in language. *Topics in Cognitive Science*, *5*(1), 111–131. https://doi.org/10.1111/tops.12000

Hesiod. (1914). *The Homeric hymns and Homerica* (Vol. 57). Harvard University Press Retrieved from http://data.perseus.org/texts/urn:cts:greekLit:tlg0020.tlg001

Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and vowels: Different roles in early language acquisition. *Developmental Science*, *14*(6), 1445–1458. https://doi.org/10.1111/j.1467-7687.2011.01089.x

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. V. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, *155*, 67–76. https://doi.org/10.1016/j.cognition.2016.06.011

Howe, C. Q., & Purves, D. (2005). The Müller-Lyer illusion explained by the statistics of image-source relationships. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1234–1239. https://doi.org/10.1073/pnas.0409314102

Huang, C.-T. J., & Roberts, I. (2016). Principles and parameters of universal grammar. In I. Roberts (Ed.), *The Oxford handbook of universal grammar* (pp. 307–354). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573776.013.14

Hyams, N. (1986). *Language acquisition and the theory of parameters*. D. Reidel.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 69–88; discussion 188–231. https://doi.org/10.1017/S0140525X10003134

Kahneman, D. (2011). *Thinking, fast and slow*. London Penguin books.

Katzir, R. (2015). A cognitively plausible model for grammar induction. *Journal of Language Modelling*, *2*(2), 213. https://doi.org/10.15398/jlm.v2i2.85

Keidel, J. L., Jenison, R. L., Kluender, K. R., & Seidenberg, M. S. (2007). Does grammar constrain statistical learning? Commentary on Bonatti, Peña, Nespor, and Mehler (2005). *Psychological Science*, *18*(10), 922–923. https://doi.org/10.1111/j.1467-9280.2007.02001.x

Kelly, K. T. (2007a). How simplicity helps you find the truth without pointing at it. In M. Friend, N. B. Goethe, & V. S. Harizanov (Eds.), *Induction, algorithmic learning theory, and philosophy* (pp. 111–143). Springer. https://doi.org/10.1007/978-1-4020-6127-1_4

Kelly, K. T. (2007b). A new solution to the puzzle of simplicity. *Philosophy of Science*, *74*(5), 561–573.

Keynes, J. M. (2010). Newton, the man. In *Essays in biography* (pp. 363–374). Palgrave Macmillan. https://doi.org/10.1007/978-1-349-59074-2_35

Kibbe, M. M., & Feigenson, L. (2014). Developmental origins of recoding and decoding in memory. *Cognitive Psychology*, *75*, 55–79. https://doi.org/10.1016/j.cogpsych.2014.08.001

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 5241–5245. https://doi.org/10.1073/pnas.0608222104

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. University of Chicago Press.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134–1140. https://doi.org/10.1177/0956797610376652

Kutrib, M., & Pighizzini, G. (2013). Recent trends in descriptional complexity of formal languages. *Bulletin of the European Association for Theoretical Computer Science*, *111*, 70–86.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Lévi-Strauss, C. (1962). *La pensée sauvage*. Plon.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*, 748–759. https://doi.org/10.1016/j.tics.2016.08.001

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 19751–19756. https://doi.org/10.1073/pnas.0704452104

Manaster-Ramer, A. (1986). Copying in natural languages, context-freeness, and queue grammars. In *Proceedings of the 24th annual meeting on association for computational linguistics* (pp. 85–89). Association for Computational Linguistics. https://doi.org/10.3115/981131.981145

Manzini, M. R., & Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguistic Inquiry*, *18*(3), 413–444.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*, 2351–2360. https://doi.org/10.1177/0956797613495418

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science*, *18*(5), 387–391. https://doi.org/10.1111/j.1467-9280.2007.01910.x

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*(1), 1–27. https://doi.org/10.1016/0010-0285(84)90002-1

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, *17*(2), 207–263.

McCarthy, J. J., & Prince, A. (1999). Faithfulness and identity in prosodic morphology. In R. Kager, H. van der Hulst, & W. Zonneveld (Eds.), *The prosody morphology interface* (pp. 218–309). Cambridge University Press.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9014–9019. https://doi.org/10.1073/pnas.1105040108

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 298–305.

Meyer, A. R., & Fischer, M. J. (1971). Economy of description by automata, grammars, and formal systems. In *12th annual symposium on switching and automata theory (swat 1971)*. IEEE. https://doi.org/10.1109/swat.1971.11

Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech perception: Evidence for early compensation. *Perception & Psychophysics*, *65*, 956–969.

Moravcsik, E. (1978). Reduplicative constructions. In J. H. Greenberg (Ed.), *Universals of human language: Word structure* (Vol. 3, pp. 297–334). Stanford University Press.

Morgan, J. L. (1986). *From simple input to complex grammar*. MIT Press.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule learning by rats. *Science*, *319*(5871), 1849–1851. https://doi.org/10.1126/science.1151564

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223. https://doi.org/10.1111/j.1551-6709.2011.01212.x

Neiworth, J. J. (2013). Chasing sounds. *Behavioural Processes*, *93*, 111–115. https://doi.org/10.1016/j.beproc.2012.11.009

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(42), 16598–16603. https://doi.org/10.1073/pnas.0703913104

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology*, *20*(2), 139–156. https://doi.org/10.1002/acp.1178

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*, 258–278.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology. General*, *146*, 1761–1780. https://doi.org/10.1037/xge0000318

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244–247. https://doi.org/10.1016/j.cognition.2009.07.011

Pepperberg, I. M. (1987). Acquisition of the same/different concept by an African Grey parrot (Psittacus erithacus): Learning with respect to categories of color, shape, and material. *Animal Learning & Behavior*, *15*(4), 423–432. https://doi.org/10.3758/BF03205051

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*(7), 1299–1305. https://doi.org/10.3758/MC.36.7.1299

Pinker, S. (1998). *How the mind works*. Allen Lane. Penguin.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology. Human Perception and Performance*, *20*, 421–435.

Pons, F., & Toro, J. M. (2010). Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition*, *116*(3), 361–367. https://doi.org/10.1016/j.cognition.2010.05.013

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*(3), 303–343. https://doi.org/10.1016/S0364-0213(02)00064-2

Ramachandran, V. (1991). Interactions between motion, depth, color and form: The utilitarian theory of perception. In C. Blakemore, K. Adler, & M. Pointoin (Eds.), *Vision: Coding and efficiency* (pp. 346–360). Cambridge University Press. https://doi.org/10.1017/CBO9780511626197.033

Rasin, E., & Katzir, R. (2016). On evaluation metrics in optimality theory. *Linguistic Inquiry*, *47*(2), 235–282. https://doi.org/10.1162/LING_a_00210

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*(3), 429–447.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338–342. https://doi.org/10.1006/ccog.1999.0386

Rissanen, J. (2008). Minimum description length. *Scholarpedia*, *3*(8), 6727. (revision #91501). https://doi.org/10.4249/scholarpedia.6727

Rose, S., & Walker, R. (2011). Harmony systems. In J. Goldsmith, J. Riggle, & A. C. Yu (Eds.), *The handbook of phonological theory* (2nd ed., pp. 240–290). John Wiley & Sons. https://doi.org/10.1002/9781444343069.ch8

Rubino, C. (2013). Reduplication. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology Retrieved from http://wals.info/chapter/27

Russo, L. (2000). *The forgotten revolution. How science was born in 300 BC and why it had to be reborn*. Springer.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680. https://doi.org/10.1016/j.cognition.2006.11.004

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory and Cognition*, *36*(6), 1057–1065. https://doi.org/10.3758/MC.36.6.1057

Shine, R. (1999). Why is sex determined by nest temperature in many reptiles? *Trends in Ecology & Evolution*, *14*(5), 186–189. https://doi.org/10.1016/s0169-5347(98)01575-4

Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology*, *25*, 256–260. https://doi.org/10.1016/j.cub.2014.11.063

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.003

Sperber, D., & Wilson, D. (1987). Précis of relevance: Communication and cognition. *Behavioral and Brain Sciences*, *10*(4), 697–710.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell.

Spierings, M. J., & ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. *Proceedings of the National Academy of Sciences*, *113*(27), E3977–E3984. https://doi.org/10.1073/pnas.1600483113

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Standing, L., Conezio, J., & Haber, R. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*(2), 73–74.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, *41*(S4), 638–676. https://doi.org/10.1111/cogs.12416

Sugiyama, L. S., Tooby, J., & Cosmides, L. (2002). Cross-cultural evidence of cognitive adaptations for social exchange among the Shiwiar of Ecuadorian Amazonia. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11537–11542. https://doi.org/10.1073/pnas.122352999

Tauzin, T., & Gergely, G. (2019). Variability of signal sequences in turn-taking exchanges induces agency attribution in 10.5-mo-olds. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 15441–15446. https://doi.org/10.1073/pnas.1816709116

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(48), 19156–19159. https://doi.org/10.1073/pnas.0700271104

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640 discussion 652–791.

Thatcher, J. W. (1973). Tree automata: An informal survey. In A. V. Aho (Ed.), *Currents in the theory of computing* (pp. 143–172). Prentice-Hall.

Todd, P., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*(5), 727–741.

Toro, J. M., Bonatti, L., Nespor, M., & Mehler, J. (2008). Finding words and rules in a speech stream: Functional differences between vowels and consonants. *Psychological Science*, *19*, 137–144.

Toro, J. M., Shukla, M., Nespor, M., & Endress, A. D. (2008). The quest for generalizations over consonants: Asymmetries between consonants and vowels are not the by-product of acoustic differences. *Perception & Psychophysics*, *70*(8), 1515–1525. https://doi.org/10.3758/PP.70.8.1515

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 195–202.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*(10), 717–726. https://doi.org/10.1038/35094573

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J., Koroshetz, W., & Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, *9*, 266–276.

van Heijningen, C. A. A., Chen, J., van Laatum, I., van der Hulst, B., & ten Cate, C. (2013). Rule learning by zebra finches in an artificial grammar learning task: Which rule? *Animal Cognition*, *16*(2), 165–175. https://doi.org/10.1007/s10071-012-0559-x

Versace, E., Spierings, M. J., Caffini, M., Ten Cate, C., & Vallortigara, G. (2017). Spontaneous generalization of abstract multimodal patterns in young domestic chicks. *Animal Cognition*, *20*, 521–529. https://doi.org/10.1007/s10071-017-1079-5

Vroomen, J., Tuomainen, J., & de Gelder, B. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language*, *38*(2), 133–149.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. https://doi.org/10.1006/cogp.1995.1016

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*(3), 470–477. https://doi.org/10.1162/jocn.2008.20040

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297. https://doi.org/10.1111/j.1467-7687.2007.00590.x

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Yamazaki, Y., Suzuki, K., Inada, M., Iriki, A., & Okanoya, K. (2012). Sequential learning and rule abstraction in Bengalese finches. *Animal Cognition*, *15*, 369–377. https://doi.org/10.1007/s10071-011-0462-x

Yates, F. A. (1966). *The art of memory*. Routledge & Kegan Paul.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Endress, A. D. (2022). In defense of epicycles: Embracing complexity in psychological explanations. *Mind & Language*, 1–30. https://doi.org/10.1111/mila.12450