Bayesian learning and the psychology of rule induction

Ansgar D. Endress
Universitat Pompeu Fabra, Barcelona, Spain
City University, London, UK
Massachusetts Institute of Technology, Cambridge, MA

Draft of November 29, 2012. Please do not quote without permission.

In recent years, Bayesian learning models have been applied to an increasing variety of domains. While such models have been criticized on theoretical grounds, the underlying assumptions and predictions are rarely made concrete and tested experimentally. Here, I use Frank and Tenenbaum's (2011) Bayesian model of rule-learning as a case study to spell out the underlying assumptions, and to confront them with the empirical results Frank and Tenenbaum (2011) propose to simulate, as well as with novel experiments. While rule-learning is arguably well suited to rational Bayesian approaches, I show that their models are neither psychologically plausible nor ideal observer models. Further, I show that their central assumption is unfounded: humans do not always preferentially learn more specific rules, but, at least in some situations, those rules that happen to be more salient. Even when granting the unsupported assumptions, I show that all of the experiments modeled by Frank and Tenenbaum (2011) either contradict their models, or have a large number of more plausible interpretations. I provide an alternative account of the experimental data based on simple psychological mechanisms, and show that this account both describes the data better, and is easier to falsify. I conclude that, despite the recent surge in Bayesian models of cognitive phenomena, psychological phenomena are best understood by developing and testing psychological theories rather than models that can be fit to virtually any data.

To recognize the taste of an apple, do we automatically think about the tastes of oranges as well as all other foods before we can know that we are eating an apple? According to a growing literature of Bayesian models, we make inferences (e.g., the kind of food we are tasting) by considering all possible situations (e.g., tasting apples, oranges etc.) in addition to the situation we actually face, and then decide which of these situations is the most likely one. Bayesian inference models have been claimed to account for an impressive variety of cognitive phenomena, including visual grouping (Orbán, Fiser, Aslin, & Lengyel, 2008), action understanding (Baker, Saxe, & Tenenbaum, 2009), concept learning and categorization (Anderson, 1991; Goodman, Tenenbaum, Feldman, & Griffiths, 2008), (inductive) reasoning (Goodman, Ullman, & Tenenbaum, 2011; Griffiths & Tenenbaum, 2009; Kemp, Perfors, & Tenenbaum, 2007; Kemp & Tenenbaum, 2009; Kemp, Tenenbaum, Niyogi, & Griffiths, 2010; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Oaksford & Chater, 1994; Téglás et al., 2011), judgment about real-world quantities (Griffiths & Tenenbaum, 2006), word learning (Frank, Goodman, & Tenenbaum, 2009; Xu & Tenenbaum, 2007), word segmentation (Frank, Goldwater, Griffiths, & Tenenbaum, 2010), and grammar acquisition (Perfors, Tenenbaum, & Wonnacott, 2010; Perfors, Tenenbaum, & Regier, 2011).

Despite this growing literature, various authors have criticized Bayesian approaches on theoretical grounds (Altmann, 2010; Bowers & Davis, 2012; Fitelson, 1999; Jones & Love, 2011; Marcus, 2010; Sakamoto, Jones, & Love, 2008), and where Bayesian approaches have been explicitly compared to psychological models (e.g., in the case of causal inference), the non-Bayesian approaches typically explained the data better (e.g., Bes, Sloman, Lucas, & Raufaste, 2012; Fernbach & Sloman, 2009). Here, I add to this literature by taking a model in a domain that appears particularly suitable for Bayesian learning — rule induction, spell out its underlying assumptions as well as their predictions, and confront them with empirical data. Specifically, Frank and Tenenbaum (2011) recently proposed that infants acquire rules in Bayesian, optimal ways. I will compare this approach with an account of rule-learning based on

This research was supported by NIH grant MH47432, grants CONSOLIDER-INGENIO-CDS-2007-00012 and PSI2012-32533 from the Spanish Ministerio de Economía y Competitividad, SGR-2009-1521 from the Catalan government, and Marie Curie Incoming Fellowship 303163-COMINTENT. I am grateful to L. Bonatti, M. Frank and Á. Kovács for helpful comments on an earlier draft of this manuscript.

simple, psychologically grounded mechanisms, and show that the latter approach provides a principled explanation for the data.

Bayesian approaches to cognition: what is optimal?

On a conceptual level, Bayesian inference is straightforward. For example, if we encounter an individual with a Red Sox cap, we conclude that she is more likely to come from Boston than from, say, New York. However, to draw this conclusion, we use our knowledge that the likelihood of somebody wearing a Red Sox cap is higher in Boston than in New York. Bayesian calculations allow us to turn the likelihood that somebody who is in Boston wears a Red Sox cap into the likelihood that somebody who wears a Red Sox cap is from Boston. Moreover, such calculations make "optimal" use of the available information.

Despite its conceptual simplicity, Bayesian inference is tremendously useful in domains from statistics (e.g., Gill, 2008; O'Hagan, 1994) to evolutionary biology (e.g., Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001; Pagel, 1994). Further, natural selection can be formulated as a Bayesian optimization problem; as a result, Bayesian inference has given us important insights into the evolution of our mental abilities. For example, some researchers have shown that perceptual and cognitive mechanisms might be well adapted to the statistics of our natural environment (e.g., Brunswik & Kamiya, 1953; Elder & Goldberg, 2002; Geisler & Diehl, 2002, 2003; Sigman, Cecchi, Gilbert, & Magnasco, 2001; Weiss, Simoncelli, & Adelson, 2002).

However, when it comes to Bayesian models of learning and cognition, environmental statistics are generally lacking, forcing such models to be much more speculative and hard to verify. This problem follows directly from Bayesian claims to make "optimal" use of information in the environment, and our lack of understanding of what has been optimized over the course and under the constraints of evolution. In fact, not all behavioral traits are optimal, but some might simply be accidents of how a species has evolved. For example, in some monogamous animals such as Zebra finches, females seek extrapair copulations although this behavior is maladaptive for females. However, extrapair mating behavior might be selected for in females because it might be affected by an allele that is shared with males, for whom siring extrapair offspring is adaptive (Forstmeier, Martin, Bolund, Schielzeth, & Kempenaers, 2011). Hence, the seemingly maladaptive behavior might be due to the accidents of how this trait is encoded genetically, suggesting that it is extremely difficult to assess whether our cognitive mechanisms are optimal and, if so, what they have been optimized for.

An overview over Frank and Tenenbaum's (2011) models

Frank and Tenenbaum's (2011) model is representative of a large number of similar models, and is applied to a domain that is arguably well-suited to Bayesian approaches. (Frank and Tenenbaum (2011) present in fact three different models, but I will present the differences between these models as they become relevant for the current purposes.) They raise the question of how young infants learn rule-like patterns based on repetitions. For example, syllable triplets like *ba-li-li* follow an ABB pattern, where the last syllable is repeated; syllable triplets like *ba-ba-li* follow an AAB pattern, where the first syllable is repeated. Following Marcus, Vijayan, Rao, and Vishton's (1999) seminal demonstration that young infants can learn such patterns, repetition-patterns have become an important testing ground for rule-learning, both in humans (e.g., Dawson & Gerken, 2009; Endress, Dehaene-Lambertz, & Mehler, 2007; Endress, Scholl, & Mehler, 2005; Frank, Slemmer, Marcus, & Johnson, 2009; Gerken, 2010; Gómez & Gerken, 1999; Kovács & Mehler, 2008, 2009a; Marcus, Fernandes, & Johnson, 2007; Saffran, Pollak, Seibel, & Shkolnik, 2007) and in nonhuman animals (e.g., Giurfa, Zhang, Jenett, Menzel, & Srinivasan, 2001; Hauser & Glynn, 2009; Murphy, Mondragon, & Murphy, 2008).

According to Frank and Tenenbaum's (2011) model, infants try to figure out the "best" rule describing the stimuli they perceive. To do so, they come equipped with an innate inventory of elementary rules, and check whether what they hear (or see) is compatible with *all* of the rules in their inventory. For example, if they hear AAB triplets, they would not only think about AAB patterns, but also about ABB patterns and all other patterns Frank and Tenenbaum (2011) incorporated into their model, even if they never hear any of these alternative patterns. To choose a rule, Frank and Tenenbaum (2011) propose that infants assume that the probability that a stimulus has been generated by a rule is inversely proportional to the total number of stimuli that can be generated by the rule (equations 2 and 3 in their first model; the other models make similar assumptions); this strategy has been called the *size principle* by Tenenbaum and Griffiths (2001).

Concretely, infants might encounter the triplets *pu-li-li* and *ba-pu-pu*, both following an ABB pattern. Hence, they encounter a total vocabulary of three syllables (i.e., *pu*, *li* and *ba*). According to Frank and Tenenbaum (2011), infants know (i) that the three syllables allow for a total of $3 \times 3 \times 3 = 27$ triplets; (ii) that 6 of these triplets follow an ABB pattern; and (iii) that 3 of these triplets follow an AAA pattern (where all three syllables are identical), even though infants have never heard any AAA triplets; infants know the number of triplets that are compatible with any other conceivable rule.

As a result, irrespective of any Bayesian computations, infants know that AAA patterns are a priori more

unlikely than *ABB* patterns, because there are fewer potential *AAA* triplets than *ABB* triplets. According to Frank and Tenenbaum (2011), infants use this knowledge to infer patterns. That is, if they hear stimuli that are equally consistent with multiple patterns, they opt for the pattern that is a priori more unlikely, and harder to conform to. Below, I will refer to patterns that are harder to conform to as the more "specific" patterns.

An alternative view: rule-learning based on perceptual or memory primitives

Before reviewing Frank and Tenenbaum's (2011) models in detail, I will briefly outline an alternative approach to rule-learning, based on perceptual or memory primitives (e.g., Endress et al., 2005, 2007; Endress, Nespor, & Mehler, 2009; see also Marcus, 2008, for a similar approach). Specifically, previous empirical work suggests that humans (and other animals) are equipped with a "repetition-detector" that is sensitive to repeated elements in a sequence. For example, in a sequence such as "*pulili*," this detector would note the repetition of the syllable "li." Presumably, this detector works best when the repeated elements are adjacent (as in *ABB* patterns), is still operative with one intervening item between the repeated items (as in *ABA* patterns, even though such patterns are harder to learn than *ABB* patterns; Gervain, Macagno, Cogoi, Peña, & Mehler, 2008; Kovács & Mehler, 2009b), and might fail to detect the repeated items when the intervening items are too numerous (e.g., in **A***BCDEFGHIJ*...**A** patterns). However, it is still unknown how repetition-detection depends on the number of intervening items or the intervening time between two repeated items.

Further, humans are equipped with a second mechanism that allows them to learn the sequence elements that occur in the edges of sequences; for example, it is easy to note that "*pulidi*" and "*ranodi*" both end with "*di*." Of course, humans are endowed with many other mechanisms, but these two mechanisms suffice to explain most of the data below.

In addition to these mechanisms, I make the following assumptions. First, when learning occurs over time, learning performance will generally be better with more exposure or more opportunities to learn than with less exposure or fewer opportunities to learn (e.g., Ebbinghaus, 1885/1913). Second, when participants are more interested in stimuli, and attend more to them, they might learn better. One way to make stimuli more interesting might be to use species-specific vocalizations, i.e., speech.

Third, when a stimulus is compatible with multiple rules that each can be learned in isolation, participants will learn all of them, and expect items to conform to them. However, some rules will be more salient and easier to learn than others; as a result, violations of these rules might be more salient than violations of less salient rules. However, which rules are salient and easy to learn is an empirical question (an assumption that is shared with some Bayesian models; see e.g., Frank & Goodman, 2012), and provides important constraints on our rule-learning abilities as well as on the underlying mechanisms.

Fourth, to explain Kovács and Mehler's (2009a) data, I simply refer to Kovács and Mehler's (2009a) own interpretation that bilinguals have enhanced executive function compared to monolinguals (e.g., Bialystok, 1999; Bialystok & Martin, 2004; Bialystok, Craik, Klein, & Viswanathan, 2004; Bialystok, Craik, & Luk, 2008; Bialystok & Craik, 2010; Costa, Hernández, & Sebastián-Gallés, 2008; Kovács & Mehler, 2009b; Kovács, 2009). While this issue is most likely orthogonal to rule-learning per se, I will discuss it below because Frank and Tenenbaum (2011) claimed that their rule-learning models provided an alternative interpretation to Kovács and Mehler's (2009a) data as well as for performance differences between monolinguals and bilinguals in the Stroop task.

Fifth, there are developmental differences on which this account (and, for that matter, Frank and Tenenbaum's (2011) account) is completely silent. For example, young infants can detect repetition-patterns in musical stimuli (Dawson & Gerken, 2009), lose this ability a few months later (Dawson & Gerken, 2009; Marcus et al., 2007), and detect such patterns again in adulthood (Endress et al., 2007). However, the experiential or maturational processes responsible for this pattern of results are unclear, and might be related to the development of language, music cognition or other cognitive faculties (see Dawson & Gerken, 2009, for discussion).

Some general problems of Frank and Tenenbaum's (2011) model

Before reassessing whether Frank and Tenenbaum's (2011) models account for the data they proposed to simulate, I will provide some general criticisms of their models that are shared by many other Bayesian models of cognition. First, as acknowledged by Frank and Tenenbaum (2011), their models are not psychological plausible. Second, in contrast to their claims, they cannot be considered ideal-observer models either. Rather, these models are implementations of specific hypotheses about specific mechanisms of the mind; these mechanisms, however, are largely speculative. Third, while Frank and Tenenbaum's (2011) models are based on the assumption that human learners generally choose more specific patterns over less specific ones, I show empirically that this hypothesis is not generally true. Frank and Tenenbaum's (2011) assumption, is, therefore, unfounded. After these points, I will turn to the specific experiments simulated by Frank and Tenenbaum (2011), and assess whether their models provide an adequate

account for the data.

Are Frank and Tenenbaum's (2011) models psychologically plausible?

Taking Frank and Tenenbaum's (2011) model at face value, they claim that, once infants enter an experimental room, they keep track of all syllables they have heard in the experiment, and while comfortably seated on their parent's lap, contemplate all possible sequences that can be formed with these syllables, as well as all possible rules with which each of these hypothetical sequences might or might not be consistent. As I will estimate below, Frank and Tenenbaum's (2011) model thus assumes that infants can process up to 900 hypothetical and counterfactual triplets per second. Such an account of rule-learning appears implausible, and, to my knowledge, is not supported by empirical evidence.

Are Frank and Tenenbaum's (2011) models ideal observer models?

Frank and Tenenbaum (2011) offer a defense against the psychological implausibility of their models that is often used by Bayesian modelers, and claim in Footnote 1 that "this approach to modeling learning is also sometimes referred to as a 'computational level' analysis, after Marr (1982), because it describes the computational structure of the task rather than the algorithms or mechanisms necessary to perform it. Models at the computational level [...] compute normative statistical inferences." Frank and Tenenbaum (2011) further argue that the "models are ideal observer models: they provide a description of the learning problem and show what the correct inference would be, under a given set of assumptions. [...] On this approach, the ideal observer becomes a baseline from which predictions about human performance can be made. When performance deviates from this baseline, researchers can make inferences about how the assumptions of the model differ from those made by human learners."

After having reviewed their models, I will discuss whether Frank and Tenenbaum (2011) adhered to these goals. Specifically, I will ask (i) whether their models were used to make predictions as opposed to fitting them to existing data, and (ii) whether the models were used to detect non-normative behavior. Further, I will provide several general reasons for which Frank and Tenenbaum's (2011) models cannot be considered computational-level models, but rather make crucial implementational assumptions.

More specific rules are not learned more easily

As mentioned above, the critical assumption of Frank and Tenenbaum (2011) model is that humans choose some rules over others because some rules are a priori more specific and harder to conform to. As I will show below, this assumption underlies all of the models' alleged successes. Hence, I will start by evaluating it empirically.[1]

One prediction of this account is tested in the experiments presented in Appendix A. Specifically, human adults were familiarized with ABB triplets carried by speech syllables. Hence, they could discover two rules. The more "specific" rule stated that triplets followed an ABB pattern; the less specific rule stated that triplets were carried by human speech syllables. The rule that all triplets follow an ABB is more specific than the rule that triplets are carried by syllables, because the latter rule is true of all possible triplets, while the former rule is true only of a subset of the triplets

Following this familiarization, participants had to choose between AAB triplets carried by other speech syllables, and ABB items carried by rhesus monkey vocalizations. Hence, they had to choose between a triplet conforming to the more specific rule (i.e., the repetition-pattern) and violating the less specific rule (i.e., being carried by speech syllables), and one triplet conforming to the less specific rule and violating the more specific rule.[2] (I performed an analogous experiment where participants were familiarized with AAB triplets.)

As shown in Appendix A, most participants found that the triplets carried by speech syllables were more like the familiarization items compared to the triplets carried by rhesus vocalizations, even though the former violated the repetition-pattern. A control experiment showed that participants readily detected repetition-patterns in rhesus vocalizations. This contradicts Frank and Tenenbaum's (2011) account, because, as mentioned above, a rule of the form "all items are syllables" will inevitable receive lower probability scores than the ABB pattern, and, as discussed below, a rule of this kind is a critical component of Frank and Tenenbaum's (2011) model. Hence, the specificity of a rule does not predict how easily it is learned.

---

[1] In natural language acquisition, an acquisition strategy assuming that infants learn the most restrictive grammar consistent with what they hear (or with what they see in the case of sign languages) is known as the subset principle (Hyams, 1986; Manzini & Wexler, 1987). However, in contrast to Frank and Tenenbaum (2011), these authors did not make unsupported assumptions about our rule-learning abilities, but rather proposed that humans evolved to acquire language following a sequence of acquisition steps that is consistent with the subset principle, using specific "triggers" to move from a more restrictive grammar to a more permissive one.

[2] The repetition-patterns remain more specific even if participants anticipate that they will hear rhesus vocalizations, and base their specificity computations on a corpus of all items occurring during familiarization and test. If so, there are 8 familiarization syllables, 2 test syllables, as well as 2 rhesus vocalizations, leading to a total of $12^3 = 1728$ possible triplets, of which $10^3 = 1000$ contain only syllables, and of which $12 \times (12-1) = 132$ conform to the repetition-pattern.

Of course, one can argue that the contrast between speech syllables and rhesus vocalizations is much more salient than the contrast between the repetition-patterns, and that the experiments reported in Appendix A are, therefore, an unfair test of Frank and Tenenbaum's (2011) model. While the contrast between speech syllables and rhesus vocalizations is likely much more salient, this is exactly the point: some rules are much more salient than others, and this constrains how humans (and other animals) learn rules. However, the relative salience of the rules is by no means predicted by the learning situation, nor by Frank and Tenenbaum's (2011) models.

Another potential criticism of these experiments is that they do not rule out a role of specificity in rule-learning, but rather show that other factors might be more important. However, it is impossible to provide evidence for the absence of a role of specificity, and it is possible that learners might, in some situations, prefer more specific patterns (or, for that matter, patterns that happen to be more specific even if learners do not consider specificity at all). At minimum, however, these results fail to support the predictions of Frank and Tenenbaum's (2011) model and demonstrate that a preference for more specific patterns cannot be taken for granted. After all, given that all of Frank and Tenenbaum's (2011) alleged modeling successes critically depend on the models' ability to choose more specific rules, one would expect actual humans to show at least some sensitivity to this principle. Hence, it seems plausible to conclude that there is no evidence from rule-learning studies for a role of specificity in rule-learning, and the experiments in Appendix A suggest that, if such a role exists, it is not strong enough to drive rule-learning in general.

Put differently, while a bias to choose more specific rules has sound computational justifications (e.g., Hyams, 1986; Manzini & Wexler, 1987; Tenenbaum & Griffiths, 2001), there is no evidence at all that learners follow this bias in artificial language learning studies. In fact, one could offer the "explanation" of the rule learning results below that infants try to make the experimenters happy; after all, published studies are generally consistent with the experimenters' hypotheses, and humans have a tendency to be helpful (e.g., Warneken & Tomasello, 2006). However, just as this "explanation" fails to provide an account of how infants might possibly know what would make the experimenters happy, and does not assess whether infants actually consider experimenter happiness at all, Frank and Tenenbaum (2011) do not explain how infants can possibly know which patterns are more specific if they encounter the patterns exclusively in laboratory studies, and do not assess whether infants actually consider rule specificity at all. Below, I will, therefore, assume that the role of psychological theories is not psychologically agnostic data-fitting. Hence, even when Frank and Tenenbaum's (2011) models fit the data, I will conclude that they fail to provide an adequate account if the fit is exclusively due to unsupported assumptions wired into the models.

A re-examination of the studies modeled by Frank and Tenenbaum (2011)

In this section, I will consider the experiments Frank and Tenenbaum (2011) simulated, and ask whether they provide an adequate account of these experiments.

Marcus et al. (1999)

Marcus et al. (1999) showed that seven-month old infants can learn repetition-patterns such as AAB and ABB. Given that Frank and Tenenbaum's (2011) models have innate repetition-detectors, it is perhaps unsurprising that they can learn repetition-patterns.

However, a more detailed look at Frank and Tenenbaum's (2011) results raises the question of whether their model really learned the repetition-patterns. In fact, the model learned two rules. The repetition-patterns, and a rule to which all triplets conform automatically (dubbed "(.,.,.)" by Frank & Tenenbaum, 2011). In the context of most of the experiments considered by Frank and Tenenbaum (2011), the rule "all items are made of syllables" would be true of all triplets.

As acknowledged by Frank and Tenenbaum (2011), the repetition-pattern is preferred *exclusively* due to the assumption that learners prefer more specific rules that are "harder" to conform to; without this assumption, the model could not choose between the repetition-pattern, and the rule to which all triplets conform automatically. However, as mentioned above, this assumption is not supported by the data from Appendix A.

Of course, the model preferred ABB to AAB triplets when familiarized with ABB triplets, but this result is unsurprising given that a repetition-detector, a sensitivity to positions in sequences and the possibility to combine repetitions and positions, and, therefore, the very possibility of discriminating between AAB and ABB has been explicitly wired into the model. Crucially, however, the assumption that allowed the model to reject inappropriate grammars is not supported by human behavior. As a result, the model fails to account for Marcus et al.'s (1999) data.

Endress et al. (2007)

Endress et al. (2007) attempted to provide evidence that repetitions are particularly salient patterns, and that their saliency does not result from any obvious formal or statistical factors. In their experiments, they used piano tones to contrast two kinds of patterns. Some participants had to learn the repetition-based patterns ABB and ABA. Others learned what Endress et al. (2007) called two "ordinal" patterns. The tones in these triplets were ordered either as "lowest-highest-middle"

(LHM), or as "middle-highest-lowest" (MHL; see Figure 1). Results showed that participants readily learned the repetition-patterns; in contrast, they were much worse on the ordinal pattern, and remained close to chance performance even after hundreds of trials with feedback.

Frank and Tenenbaum (2011) propose an alternative account of these findings. Regarding the learning of repetition-patterns, Frank and Tenenbaum's (2011) model preferentially learns the repetition-patterns over the rule that is true of all items due to the assumption that learners prefer more specific patterns to less specific ones. As discussed above, this assumption is not supported by the experiments presented in Appendix A.

Regarding the participants' difficulty with ordinal patterns, the model had problems learning the ordinal rules because multiple rules conformed to the triplets. For example, *LHM* triplets conform to many different rules, including: (i) the first tone is lower than the third one, (ii) the first tone is lower than the second one, (iii) the second tone is higher than the third one, (iv) the first tone is lower than the second one *and* the third one, and so on. The model thus has to "choose" the most relevant of these patterns. According to Frank and Tenenbaum (2011), participants have difficulties learning ordinal patterns because the model cannot decide between the multiple rules that are consistent with the triplets.

This account makes a prediction that is highly implausible: people should be unable to discriminate patterns consisting of rising vs. falling melodies. Specifically, as shown in Figure 1, rearranging the tones in *LHM* and *MHL* patterns leads to *LMH* and *HML* patterns, that is, simply to rising and falling contours. Frank and Tenenbaum (2011) predict that people should have problems learning rising and falling contours, because these melodies are consistent with the same number of spurious rules as those melodies used by Endress et al. (2007). Hence, the model would fail to learn rising vs. falling contours any better than Endress et al.'s (2007) patterns.

Unsurprisingly, the experiment shown in Appendix B demonstrates that people readily discriminate rising from falling contours: after a familiarization with falling triplets or rising triplets, most participants are at ceiling discriminating rising from falling triplets, using the same tones as Endress et al. (2007). In contrast to Frank and Tenenbaum's (2011) claims, the number of spurious rules compatible with melodic patterns is, therefore, irrelevant to the success of actual humans in learning such patterns. As a result, Frank and Tenenbaum (2011) fail to provide an account of Endress et al.'s (2007) data.

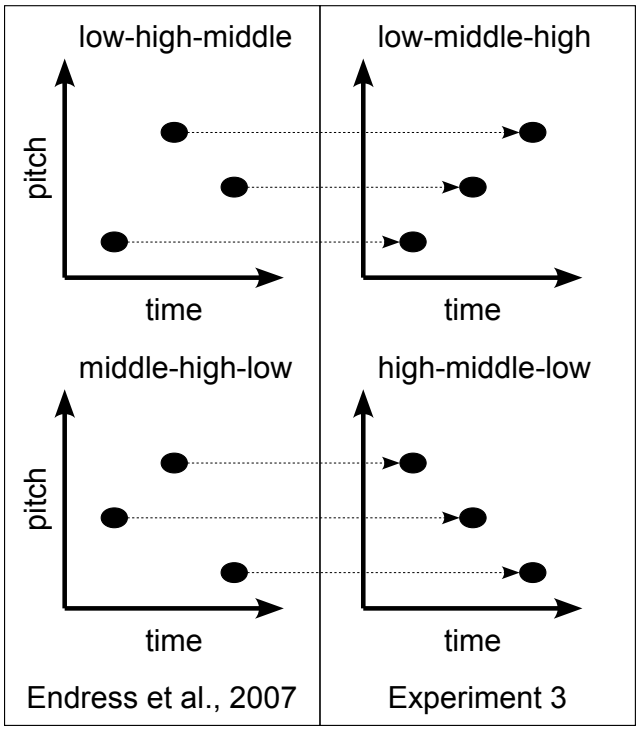

Figure 1. In Endress et al.'s (2007) experiments, participants had to learn the "ordinal" patterns low-high-middle and middle-high-low. Rearranging the tones in the pattern yields the patterns low-middle-high (rising) and high-middle-low (falling). Frank & Tenenbaum's (2011) model predicts that it should be equally difficult to learn ordinal patterns as to learn rising and falling patterns. Experiment 3 (presented in Appendix B) demonstrates that this is not the case.

Frank, Slemmer, et al. (2009)

Frank, Slemmer, et al. (2009) proposed that 5-months-old infants are better at learning repetition-patterns when these patterns are presented in two modalities simultaneously (i.e., looming shapes accompanied by syllables) compared to unimodal conditions where triplets were composed of either shapes or syllables.

Frank and Tenenbaum (2011) provide two explanations, both of which are problematic. The first explanation relies on their assumption that learners prefer more specific, harder-to-conform-to rules. That is, infants might generate all 262,144 multimodal triplets, and find that the triplets conforming to the repetition-patterns are 64 times less numerous than the triplets that conform to the rule that is automatically true of every triplet. In the unimodal condition, in contrast, the ratio between these triplet types is only 8 rather than 64. Hence, the specificity advantage is more pronounced in the multimodal condition than in the unimodal condition; according to Frank and Tenenbaum (2011), this explains why infants are better at learning multimodal rules.

While the experiments in Appendix A show that actual learners do not necessarily prefer more specific, harder-to-conform-to rules, it is instructive to

take Frank and Tenenbaum's (2011) explanation at face value. If infants access all 262,144 triplets, and verify all possible rules in a 5-min experiment, they have to check about 900 triplets per second. It seems reasonable to conclude that such a model requires further empirical backing. Of course, Frank and Tenenbaum (2011) claim that their model is an ideal observer model (but see below), but even in this case, one might ask how the infant mind might know that multi-modal rules are more specific than uni-modal rules, and, if infants have innate knowledge of the relative specificity of rules, why they might have such knowledge in the first place.

Frank and Tenenbaum's (2011) second explanation of Frank, Slemmer, et al.'s (2009) data is based on the assumption that infants might continuously consider the possibility that they might have misperceived or misremembered a triplet or whatever they happen to perceive. Frank and Tenenbaum (2011) assume that infants have a certain probability of misremembering or misperceiving triplets, and that they can adjust the probabilities of the different patterns accordingly. According to Frank and Tenenbaum (2011), infants believe that they are more likely to misremember or misperceive unimodal stimuli compared to multimodal stimuli. When infants misperceive or misremember items, they randomly pick a rule for these items. At first sight, Frank and Tenenbaum's (2011) second account just seems to be a description of the experiments rather an explanation: by changing the parameter controlling what the model considers its misperception/misremembering probability, it becomes easier to learn multimodal patterns compared to unimodal patterns.

However, this account raises an important problem. If infants keep track of what they might (mis-) remember, then they are batch learners, and learn patterns by faithfully remembering all of the triplets and then evaluating them. However, there is no evidence that infants remember any of the triplets they have heard. Moreover, there is no need to remember any triplets either: to learn the patterns, infants just need to remember the *patterns* of the triplets, but not the triplets themselves. In the complete absence of evidence for a batch learning model, it appears psychologically implausible, and requires empirical evidence.

While Frank and Tenenbaum's (2011) models can be easily transformed into an online learner, their account would raise the question of whether infants track misperceptions at all, and mentally replace them with a randomly picked rule. In fact, the most natural assumption seems to be that infants simply ignore the subset of the stimuli they do not remember. As a result, they would perceive them less often if they are presented multiple times. However, such a noise parameter would have only a limited effect on learning performance in Frank and Tenenbaum's (2011) models. Hence, to fit their model to the data, Frank and Tenenbaum (2011) posit that "misremembered/misperceived" items are associated with a randomly picked rule. Unfortunately, Frank and Tenenbaum (2011) do not provide any evidence in favor of this account. In any case, the model fails to account for the learning of repetition-patterns in the first place, because humans might not prefer more specific rules of less specific ones.

Gerken (2006)

Gerken (2006) investigated the generality of the rules that infants can learn. In one condition, infants were familiarized with AAB or ABA triplets, roughly as in Marcus et al.'s (1999) experiments. In the other condition, infants were again familiarized with AAB or ABA triplets. Crucially, however, the B syllable was always /di/, yielding patterns of the form AAdi and Adi$A$. In both conditions, infants were then tested on triplets that did *not* contain the syllable /di/.

When familiarized with AAB or ABB triplets, infants discriminated consistent from inconsistent items; in contrast, when familiarized with AAdi or Adi$A$ items, infants failed to discriminate between these items, although the conditions did not differ significantly. However, when tested on novel AAdi or Adi$A$ items, infants discriminated inconsistent items from consistent ones.

In the two conditions where infants succeeded, Frank and Tenenbaum's (2011) model succeeds as well because the "winning" rule is more specific and harder-to-conform-to than alternative rules (e.g., the rule that is automatically true of all triplets). However, as mentioned above, Frank and Tenenbaum's (2011) hypothesis that more specific rules are learned preferentially is not supported by the data presented in Appendix A.

While Frank and Tenenbaum's (2011) model fails to explain why repetition-rules can be learned in the first instance, it also fails to explain why infants fail when familiarized with AAdi or Adi$A$ items and tested on items that do not contain /di/. Specifically, Frank and Tenenbaum (2011) assume that infants familiarized with AAdi triplets maintain three distinct rules: (i) triplets start with a repetition; (ii) triplets end with /di/; (iii) triplets start with a repetition *and* end with /di/. That is, while the third rule is the conjunction of the first two, infants are claimed to maintain it separately. Given that the last rule is the most specific one and the hardest to conform to, the model prefers it.[3] However, as mentioned above, the assumption that more specific rules are generalized preferentially is not supported by the experiments presented in Appendix A.

In fact, Frank and Tenenbaum's (2011) explanation of Gerken's (2006) data does not only rely on the unsupported assumption that more specific rules are learned

[3] If there are N syllables used to construct the triplets, the first rule generates N^2 triplets: the first two syllables are a repetition of any of the N syllables, while the last syllable can again be any of the N syllables. The second rule yields N^2 triplets for similar reasons. The third rule, in contrast, yields only N triplets, and is, therefore, more specific than the other two.

preferentially, but also on further assumptions for which there is no evidence. Specifically, Frank and Tenenbaum (2011) assume that infants link rule (i) and rule (ii), and combine them into a conjunction rule. However, there is no evidence that infants actually link the two rules. If they maintain both rules independently, neither rule should be more specific. As a result, the model should not prefer either rule, suggesting again that the models fail to provide an account of Gerken's (2006) data.

Gerken (2010)

Gerken (2010) asked whether very limited experience would allow infants to show evidence of learning of the repetition-patterns when familiarized with *AA*di or *A*di*A* triplets. As in Gerken's (2006) experiments, infants were familiarized with *AA*di or *A*di*A* triplets, and then tested on *AAB* or *ABA* triplets that did not contain /di/. Crucially, however, she added five additional familiarization triplets, three of which conformed to the same repetition-pattern as the other familiarization triplets, but did not contain /di/. Strikingly, this minimal change allowed infants to discriminate the two repetition-patterns even if the test items did not contain /di/. In a crucial control condition, Gerken (2006) showed that replacing the *AA*di or *A*di*A* triplets with music (and keeping the last 5 familiarization triplets) did not allow infants to discriminate the repetition-patterns, suggesting that infants did not just use the last five trials to learn the repetition-pattern.

As in the simulations reviewed so far, Frank and Tenenbaum's (2011) model explains the learning success by the model's preference for more specific rules, which is not supported by the data presented in Appendix A.

Further, Frank and Tenenbaum's (2011) model makes a prediction that has not been tested but that appears highly implausible. Specifically, Frank and Tenenbaum's (2011) equations (1) and (2) show that the model predicts that, no matter for how long infants are familiarized with *AA*di or *A*di*A* items, a *single* item not containing /di/ leads to the rejection of all /di/ rules.[4] For example, if human adults are familiarized with 10,000 *AA*di triplets, and then shown a single *AAB* triplet not containing /di/, they should forget the *AA*di pattern, even if 9,999 out of 10,000 triplets were consistent with it. This prediction appears implausible.

Importantly, this is not an unfair test of Frank and Tenenbaum's (2011) model. Given that Frank and Tenenbaum (2011) consider a virtue of the model that it can learn "with only a small amount of evidence" (p. 366), it seems reasonable to conclude that the flipside of this ability, namely to unlearn "with only a small amount of evidence," is an equally crucial and central feature of the model. Hence, the feature of Frank and Tenenbaum's (2011) model that allows them to fit Gerken's (2010) data makes incorrect predictions, suggesting that it fails to provide an account of the data.

Marcus et al. (2007)

Marcus et al. (2007) asked whether infants preferentially learn repetition-patterns in some stimulus modalities than others. In a nutshell, they showed that infants readily learn repetition-patterns when these are implemented with speech syllables; however, they show no significant learning when the triplets are implemented using pure tones, timbres, or animal sounds during both familiarization and test.

In a marked contrast, when infants are familiarized with speech triplets conforming to a repetition-pattern, they successfully discriminate the pattern they have heard from unfamiliar patterns — even when tested on tones, timbres or animal sounds. Marcus et al. (2007) concluded that "infants may analyze speech more deeply than other signals because it is highly familiar or highly salient, because it is produced by humans, because it is inherently capable of bearing meaning, or because it bears some not-yet-identified acoustic property that draws the attention of the rule-induction system" (p. 390).

Although this conclusion is plausible enough to be taken as an accurate description of Marcus et al.'s (2007) results, Frank and Tenenbaum (2011) took issue with it, but failed to provide an adequate alternative account. First, as in all other experiments reviewed so far, their model fails to provide an account of why infants can learn repetition-patterns in the first place, because the model's success relies on the assumption that infants prefer the most specific, hardest-to-conform-to rule; as mentioned above, this assumption is not supported by the results reported in Appendix A.

Second, Frank and Tenenbaum (2011) explain Marcus et al.'s (2007) data by speculating that infants misperceive or misremember more non-speech items than speech items, and that they randomly pick a rule for the misperceived or misremembered items. Importantly, however, Frank and Tenenbaum (2011) assume that these perceptual problems are specific to the familiarization phase, while infants have perfect perception in the test phase irrespective of the type of stimuli they are exposed to.[5] With this assumption, it is unsurprising that

[4] The probability that a non-/di/ triplet has been generated by a rule involving /di/ is zero; as this probability appears in the product used to calculate the probability of the /di/ rules, the posterior probability of all /di/ rules is necessarily zero as well.

[5] From the middle part of their Figure 2, it is apparent that Frank and Tenenbaum (2011) believe that it is reasonable to assume that 80% of the non-speech items (and 10% of the speech items) are misremembered or misperceived (but only during familiarization, with perfect memory and perception during test); the left part of their Figure 2 reveals that, for the model to exhibit an advantage for speech items, one needs to assume that infants misperceive at least 50 or 60% of the non-speech items.

patterns implemented in speech are learned better than patterns implemented with non-speech items; after all, infants are hypothesized to misremember or misperceive them.

However, there is no reason to assume that infants perceive or remember the very same stimulus differently depending on whether it appears in a familiarization or a test phase. If one assumes that infants have the same perceptual or memory difficulties during test as during familiarization, they will perform much worse when tested on non-speech material, even after a familiarization with speech items.

Frank and Tenenbaum (2011) acknowledge this problem, and mention in Footnote 9 that, if the same memory or perception problems are assumed during familiarization and during test, they find "an appreciable gap in performance between speech and [non-speech]". However, this prediction is refuted by Marcus et al.'s (2007) data: in their experiments, the discrimination between consistent and inconsistent items yielded an effect size of .886 in the speech condition, and of .745 in the condition where infants were familiarized with speech items and tested on tones. Using a unit-normal approximation to the effect sizes, the two effect sizes are well within a 12% confidence interval of each other, and, therefore, do not differ significantly. This, however, contradicts Frank and Tenenbaum's (2011) model.

In sum, Frank and Tenenbaum (2011) do not provide an adequate account for Marcus et al.'s (2007) data, both because their model cannot account for the learning of repetition-patterns in the first place, and because their account of the differences between the speech and non-speech conditions makes predictions that are inconsistent Marcus et al.'s (2007) data.

Saffran et al. (2007)

Saffran et al. (2007) showed that infants can learn repetition-patterns of simultaneously presented dogs. As with the other experiments reviewed so far, Frank and Tenenbaum's (2011) model fails to account for this finding, because the assumption that learners prefer the most specific rule is not supported by the data presented in Appendix A.

Saffran et al. (2007) also showed that infants who were (according to parental report) "very interested" in dogs performed better than infants who were only "interested." Frank and Tenenbaum (2011) explain this result by claiming that infants who are only "interested" in dogs are more likely to misperceive or misremember them; then, they randomly pick a rule for misremembered or misperceived items, instead of simply ignoring them. This leads to a negative correlation between the probability that the model misremembered or misperceived items and its rule-learning performance.

However, there are two reasons that make Frank and Tenenbaum's (2011) interpretation of the effects of interest in dogs implausible. First, the left part of their Figure 2 shows that the biggest differences in rule-learning performance arise when unreasonably large probabilities of misremembering or misperceiving items are assumed. (This can be seen by holding the value on the x-axis constant, and varying what Frank and Tenenbaum (2011) call the α_{NS} parameter.) For example, performance is essentially unchanged if the misremembering/misperceiving probability is 0, 10, 20 or 30%, respectively; in contrast there are large performance differences when large misremembering/misperceiving probabilities of more than 40% are assumed. As a comparison, Frank and Tenenbaum (2011) assumed in the context of Marcus et al.'s (2007) experiments that a misremembering/misperception probability of 10% would be reasonable. This is especially troublesome if Frank and Tenenbaum's (2011) model is located at the computational level; after all, there is no difference in the "computational structure" of Marcus et al.'s (2007) and Saffran et al.'s (2007) experiments that justifies a fourfold increase in the misremembering probability, suggesting again that Frank and Tenenbaum (2011) do not provide computational-level, ideal-observer models but rather make detailed implementational assumptions.

Hence, Frank and Tenenbaum (2011) fail to provide and adequate model of Saffran et al.'s (2007) data, both because their model fails to learn repetition-patterns in the first place, and because their explanation of the effects of infants' interest in dogs relies on changing an ad-hoc parameter that they know a priori to correlate with rule-learning performance.

Gómez (2002)

Gómez (2002) investigated the role of variability for learning dependencies between non-adjacent items, both in adults and in infants. In the experiments with adults (those modeled by Frank & Tenenbaum, 2011), participants listened to triplets of the form *aXd*, *bXe* and *cXf*. *a*,*b*, *c*, *d*, *e* and *f* were specific non-words; *X* came from classes with 2, 6, 12 or 24 members. The size of the classes was varied across participants. Importantly, Gómez (2002) equated the number of occurrences of the $a \ldots f$ words in the different class-size conditions. That is, each triplet was presented 72, 24, 12 and 6 times for the class-sizes 2, 6, 12 and 24, respectively. Following this familiarization, participants were presented with test items, and had to choose whether they had heard them. These items were either items they had actually heard, or foils where the regularity between the first and the last word in a triplet was broken (i.e., foils had the form *aXe*, *bXf* or *cXd*). When the *X* words came from classes of 2, 6, or 12 elements, participants discriminated correct triplets from foils only at low, marginally significant levels of performance. In contrast, when *X* was taken from a set of 24 elements, performance was excellent.

To account for these data, Frank and Tenenbaum (2011) first modified their model to enable it to learn

multiple rules simultaneously; that is, they modified it so that it could learn all three dependencies between initial and final words. Second, they postulated that participants misremember or misperceive *exactly* 60% of the triplets they heard. (As shown in their Figure 3, if the model misremembers or misperceives fewer triplets, the performance of smaller set sizes becomes too high; and when it misremembers or misperceives more triplets, the overall performance becomes too low.)

This account is problematic for three important reasons. First, as in all other simulations reviewed so far, the model selects the appropriate rules by choosing the most specific ones compatible with the input; Appendix A shows that this assumptions is unsupported. Second, to fit the model to the data, Frank and Tenenbaum (2011) have to set a parameter to a specific value although it is unclear why, according to Frank and Tenenbaum (2011), "the computational structure" of the problem dictates a forgetting rate of *exactly* 60%.

Third, and crucially, granting that Frank and Tenenbaum's (2011) misremembering parameter has psychological meaning, their assumption that it is constant for the different class-size conditions is most likely incorrect. As mentioned above, Gómez (2002) kept the number of tokens in each class-size condition constant; for example, each triplet was played 72 times when X items were taken from a set of two words, and 6 times X items were taken from a set of 24 words. Hence, one would expect the misremembering likelihood to be higher when X items are taken from a set of 24 words than when they are taken from a set of two words.

In Figure 2, I replotted the results from Frank and Tenenbaum's (2011) Figure 3, but taking into account that triplets are repeated more often in the low variability condition than in the high variability conditions. Specifically, I assumed that the misremembering probability is lowest in the low variability condition, and that it decreases proportionally to the logarithm of the number of repetitions of each triplet. As shown in Figure 1, the results directly contradict Gómez's (2002) data: while participants performed best for the high variability condition, the model performed best for the low variability condition.

There is also a second way of using the misremembering/misperception parameter in the context Gómez's (2002) data. After all, she asked how likely participants were to endorse items they had heard, and items they had not heard and violated the non-adjacent dependency. Hence, Gómez's (2002) experiments might be seen, at least in principle, as testing the memory of the items participants had heard, even though it is clear that participants' performance was not driven by memory for complete triplets. (After all, their performance was worse for triplets they had heard 72 times than for triplets they had heard 6 times.) For the sake of completeness, I plotted in Figure 2 the model predictions assuming that participants' actual performance was reflected in the misremembering/misperception pa-

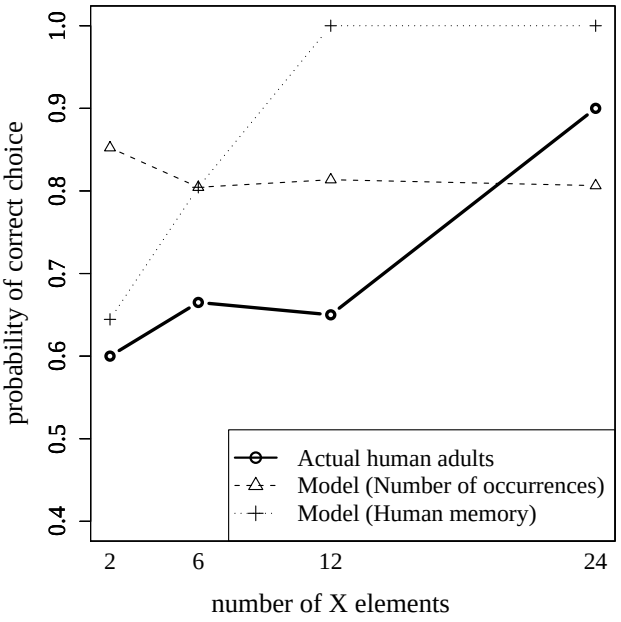

Figure 2. Results of Gómez's (2002) data (solid line with circles), and two predictions of Frank & Tenenbaum's (2011) model. (Dashed line with triangles) In Gómez's (2002) experiments, triplets with 2, 6, 12 and 24 X items were repeated 72, 24, 12 and 6 times, respectively. Assuming that the number of repetitions increases the memory strength of items (Ebbinghaus, 1885/1913), Frank & Tenenbaum's (2011) misremembering parameter should decrease with the number of repetitions. Choosing the smallest misremembering parameter in Frank & Tenenbaum's (2011) Figure 3 as the misremembering probability for 72 repetitions, and then scaling the misremembering parameter proportionally to the logarithm of the number of repetitions yields misremembering probabilities of 10, 30, 50 and 60% corresponding to 72, 24, 12 and 6 repetitions, respectively. Under these assumptions, Frank & Tenenbaum's (2011) model predicts that participants should be best with two X items, while participants actually are best with 24 it X items. (Dotted line with crosses) A different way of interpreting Frank & Tenenbaum's (2011) misremembering parameter is to consider Gómez's (2002) experiments as memory experiments (although they clearly are not), and to consider the percentage of incorrect responses in each of the condition as Frank & Tenenbaum's (2011) misremembering parameter. Choosing the misremembering probabilities in Frank & Tenenbaum's (2011) Figure 3 that are closest to the percentages of incorrect responses in Gómez's (2002) experiment shows that the model predicts participants to be at ceiling from set-size 12 onwards. This result replicates that the misremembering parameter correlates with performance. Importantly, however, the pattern of results does not fit the behavior of Gómez's (2002) participants.

rameter; that is, for each set size, I used the misremembering/misperception probability that was closest to the participants' probability of *incorrect* responses.

It should be noted that using the misremembering/misperception parameter in this way is circular; given that Frank and Tenenbaum (2011) report in the context of Saffran et al. (2007) data that the misremembering/misperception probability is negatively correlated with performance, one would expect this to be case here as well. Hence, the model should perform better for larger class-sizes for this reason alone. Importantly, however, the model's results did not fit those of Gómez's (2002) participants, as it performed at ceiling for all class-sizes from 12 onwards.

In sum, Frank and Tenenbaum's (2011) model fails to account for Gómez's (2002) data, both because it makes assumptions unsupported by the results reported in Appendix A, and because the model's results clash with those of actual participants.

Kovács and Mehler (2009a)

Kovács and Mehler (2009a) investigated how easy it is for infants to learn two patterns simultaneously. In their experiments, two patterns (e.g., *AAB* and *ABA*) predicted visual rewards on two different locations on a screen; they measured whether, upon hearing a pattern, infants would show anticipatory looks to the location where the reward would appear. They showed that infants from bilingual households learned both rules, but monolinguals learned only one. Kovács and Mehler (2009a) proposed that the bilingual advantage was due to bilinguals' well known advantage in executive function (e.g., Bialystok & Craik, 2010), which they had shown to be present already in infancy (Kovács & Mehler, 2009b).

Frank and Tenenbaum's (2011) explanation of Kovács and Mehler's (2009a) results is to introduce an additional parameter controlling how likely the model is to postulate multiple rules, and to show that the model is more likely to posit multiple rules when the parameter is set to allow for multiple rules. They conclude that bilingual infants have "a more permissive prior on the number of regularities infants assume to be present in a particular stimulus. In practice this may be manifest via better executive control, as hypothesized by Kovács & Mehler." In other words, Frank and Tenenbaum (2011) found that a model that is designed to be more likely to admit more than one regularity is indeed more likely to learn more than one regularity, and conclude that bilinguals are somehow designed to be more likely to admit more than one regularity as well.

However, Frank and Tenenbaum's (2011) simulations fail to provide an account of Kovács and Mehler's (2009a) data. First, as in all other simulations reviewed so far, the model fails to provide an account of the learning of repetition-patterns because the underlying assumption is not supported by the data reported in Appendix A: human learners do not necessarily prefer more specific, harder-to-conform-to rules.

Second, even though Frank and Tenenbaum's (2011) extra-parameter allowed them to fit their model to the data, their conclusion completely ignores the substantial literature on the effects of bilingualism on executive function. It is well established that bilingual adults and children have better executive function in a variety of tasks that are entirely unrelated to learning multiple regularities. These tasks include dimensional card sorting tasks (e.g., Bialystok, 1999; Bialystok & Martin, 2004), the Simon task (e.g., Bialystok et al., 2004), the Stroop task (e.g., Bialystok et al., 2008), and the Flanker task (Costa et al., 2008). In the case of reversal learning, the executive advantage can be observed even in early infancy (Kovács & Mehler, 2009b). Further, the effects of the bilingual advantage are seen in domains such as Theory of Mind (Kovács, 2009) that have no obvious relation to regularity learning either.

It thus seems reasonable to conclude that Frank and Tenenbaum's (2011) additional parameter has no relation at all to the data presented by Kovács and Mehler (2009a), on top of the fact that their model does not account for the learning of repetition-patterns in the first instance.

Did Frank and Tenenbaum (2011) use their models as ideal observer models after all?

As discussed in the introduction, the goal of Frank and Tenenbaum's (2011) "ideal observer" models, as well as of many other Bayesian models of cognition, is (i) to make predictions, and to (ii) detect non-normative behavior. Both would be important contributions to our knowledge. As a result, it is crucial to assess to what extent their models achieved these goals.

Regarding the first goal, the review of Frank and Tenenbaum's (2011) models suggests that they fitted their models to existing data, but did not make any novel predictions. Even in cases where their models make novel predictions, Frank and Tenenbaum (2011) do not discuss them. For example, they do not discuss why their model predicts that humans should be unable to discriminate rising from falling melodies; nor do they discuss why their account of Gerken's (2010) data predicts that learning should suffer catastrophically from a single counter-example, irrespective of how much exposure is given. They do not evaluate the models' predictions for Gómez's (2002) data either.

Regarding the second goal, Frank and Tenenbaum (2011) did not appear to attempt to detect non-normative behavior, but simply made additional assumptions to fit their models to the data. For example, they note that their first model cannot account for the learning differences as a function of the infants' interest in the stimuli (Saffran et al., 2007), and add additional assumptions to the model as a result. A natural as-

sumption would be that infants are simply more likely to ignore the stimuli they are not interested in, leading infants to perceive each training item less often. However, such an assumption would not allow Frank and Tenenbaum's (2011) model to fit the learning differences. As a result, they opt for a different assumption that does allow them to fit the data, namely that infants attribute a randomly chosen pattern to some randomly chosen familiarization items that they do not even remember. (Similar accounts are offered for Frank, Slemmer, et al.'s (2009) and Marcus et al.'s (2007) data.) Despite its *prima facie* implausibility, no further predictions of this account are offered.

Frank and Tenenbaum's (2011) focus on data fitting rather than on detecting non-normative behavior is also evident in their account of Marcus et al.'s (2007) experiments. To fit their models to the data, they assume that infant perception has markedly different properties depending on whether infants hear familiarization items or test items, even if nothing distinguishes the two types of items. It is hard to see how such an account would follow from the computational structure of the learning problem, or how it could be justified otherwise.

Likewise, to account for the ability of bilingual but not monolingual infants to learn several rules simultaneously, Frank and Tenenbaum (2011) postulate that bilinguals are more likely to posit multiple rules in general, completely ignoring Kovács and Mehler's (2009a) natural explanation based on the well-established executive function advantage in bilinguals. Rather, Frank and Tenenbaum (2011) opinionate that the bilinguals' increased propensity to postulate multiple rules "may be manifest via better executive control [in practice]." However, they fail to provide an account why the propensity to learn multiple rules might be important for, say, the Stroop task, nor do they make any predictions from this account. It seems plausible to conclude that a model that wires in the data it attempts to reproduce is not well suited for detecting non-normative behavior.

That being said, Frank and Tenenbaum (2011) did modify their model in response to data that the simpler versions of the model could not fit; however, according to Frank and Tenenbaum (2011), all versions of their model count as ideal-observer models, raising the question of what, if any, data would constitute evidence against the ideal observer view of rule-learning, or whether a model with an arbitrary number of assumptions is still "a useful baseline for future work on rule learning."

In addition to not using their models to make predictions or to detect non-normative behavior, Frank and Tenenbaum's (2011) highly detailed implementational assumptions suggest that their models cannot be considered ideal observer models. For example, they assume that learners come innately equipped with a repetition-detector, a detector for rising tone intervals, one for falling tone intervals, a mechanism that can link specific syllables to positions in triplets, and a mechanism that can combine all of these rules into one.

These implementational assumptions have a profound impact on the model predictions. For example, without a repetition-detector, the models could not learn the repetition-patterns. However, while Frank and Tenenbaum (2011) claim that their model describes the "computational structure" of the learning problem, the computational structure does not imply that the ability to notice repeated items is implemented using a dedicated repetition-detector. In fact, it is possible to know that the sequence "pupu" contains two identical syllables by noticing that the number of syllable tokens does not match the number of syllable types. Likewise, to know that two tones are identical, we can notice that the first tone is neither higher nor lower than the second one. Importantly, these are not just ad-hoc arguments against Frank and Tenenbaum's (2011) model, but, on both computational and psychological grounds, there is no a priori reason to take a repetition-detector for granted.[6] Hence, the "computational structure" of rule-learning implies by no means the existence of a repetition-pattern. But without this assumption, Frank and Tenenbaum's (2011) results would be profoundly changed.[7] (While both the account sketched below and Frank and Tenenbaum's (2011) models share the assumption of a repetition-detector, Frank and Tenenbaum (2011) reject the evidence for such a mechanism (see the review of their simulations of Endress et al.'s (2007) data above). Hence, by Frank and Tenenbaum's (2011) own conclusions, the assumption that humans are endowed with a repetition-detector would be entirely unsubstantiated.)

[6] From a computational point of view, there are computer architectures without equality operators but only "greater than" and "smaller than" operators. (On such architectures, an instruction to check the equality of two numbers is internally translated to checking that one number is neither greater nor smaller than the other.) Such architectures would not have a repetition-detector. From a psychological perspective, domains without repetition-detectors are well documented as well. For example, it is much easier to notice that two lines of contour are symmetrical than that they are repeated (i.e., that the two lines are translations of one another; Baylis & Driver, 1994, 1995, 2001); it is much harder or even impossible to process repetitions on consonants than on vowels (Pons & Toro, 2010; Toro, Bonatti, Nespor, & Mehler, 2008; Toro, Shukla, Nespor, & Endress, 2008); and human adults are unable to learn repetition-patterns over syntactic categories (Endress & Hauser, 2009).

[7] If repetitions are detected as a combination of two rules, for example as the negated conjunction of two "difference detectors," they necessarily become less accessible, because the probabilities of the two component rules must be multiplied at some point, resulting in a lower overall probability. Of course, it is possible to "patch" such a model, but this would be just another implementational assumption that is wired into the model.

Frank and Tenenbaum (2011) do not only make strong implementational assumptions about the available psychological mechanisms, but even about their detailed inner workings. For example, to fit their models to the various experiments, Frank and Tenenbaum (2011) sometimes assume that the forgetting rate is 10%, for other experiments 40%, and for still other experiments 80% — even though the computational problem is exactly the same. It is hard to see how such assumptions can be justified by the computational structure of the learning problem.[8]

It thus appears that, rather than making predictions or attempting to detect non-normative behaviors, Frank and Tenenbaum (2011) just attempted to fit their models to available data, using additional assumptions with no clear justification or further predictions. Hence, it seems plausible to conclude that Frank and Tenenbaum's (2011) models do not address any of the goals they attribute to ideal observer models.

Put differently, one might ask what the role of the Bayesian computations are for the models' successes. In fact, what Frank and Tenenbaum (2011) propose is a model positing that, among other things, (i) more specific rules are learned more readily; (ii) infants know, presumably innately, which patterns are more specific, even if they encounter the patterns exclusively during experiments; (iii) the same items are perceived/remembered differently depending on whether they appear during familiarization or during test; (iv) forgetting rates can be arbitrarily set, sometimes to 10%, sometimes to 60%, and sometimes to 80%; (v) there is catastrophic forgetting due to single counterexamples. Had these assumptions not been swept under the Bayesian carpet, it is hard to see how they could follow from the computational structure of the learning problem.

An account based on common-sense psychology

While Frank and Tenenbaum's (2011) models do not appear to provide an adequate account of any of the experiments reviewed above, a model based on perceptual or memory primitives might fare much better. I will now discuss these basic psychological explanations of each of the experiments reviewed above.

Marcus et al. (1999), Endress et al. (2007)

The view that humans are equipped with a repetition-detector that makes repetitions relatively salient patterns provides a straightforward interpretation of Marcus et al.'s (1999) and Endress et al.'s (2007) data, because these experiments show that humans can learn repetition-patterns, and that such patterns are learned better than arbitrary patterns for which no such detector exists.

Frank, Slemmer, et al. (2009)

As before, the ability to learn repetition-patterns can be explained by the existence of a repetition-detector. Likewise, there are numerous straightforward explanations of Frank, Slemmer, et al.'s (2009) finding that multi-modal rules are learned better than uni-modal rules. One possible interpretation is that infants received more instances of the repetition-pattern. That is, in the multimodal condition, infants were presented with visual and auditory triplets simultaneously and, therefore, with twice as many triplets as in either unimodal conditions. A second possible interpretation relates to variability of the stimuli. In the multimodal condition, triplets were more variable, as infants perceived both visual and auditory triplets; if variability helps learning (Gómez, 2002), infants might learn better in the multimodal condition as well. A third possible interpretation is provided by the fact that infants have more opportunities to attend to the stimuli in the multimodal condition; for example, if they are distracted by a sound, they might still attend to the visual stimuli, and if they are distracted by other visual objects, they might still attend to the sounds. A fourth account relies on the fact that infants have more opportunities to recognize the familiar pattern during test: if they do not attend to the sound of a test item, they might still attend to its visual component, and vice versa. It is easy to come up with further explanations of Frank, Slemmer, et al.'s (2009) data relying on basic psychological considerations.

Gerken (2006), Gerken (2010)

Gerken's (2006) and Gerken's (2010) experiments can be explained if, in addition to being sensitive to repetitions, humans (and other animals) track items in the edges of sequences (e.g., Endress & Mehler, 2009; Endress, Carden, Versace, & Hauser, 2010; Seidl & Johnson, 2006), and if they expect test items to conform to all regularities they have heard. That is, infants might consider triplets as a violation if *any* of the rules is violated. For example, when familiarized with *AAB* triplets (where the last syllable is not systematically /di/), infants should be sensitive to violations of the

[8] It might be argued that there is a long-standing distinction between the assumptions of a model and its free parameters in machine learning, and that reliance of Frank and Tenenbaum's (2011) model on specific parameter values should not be considered an assumption. However, the goals of machine learning and cognitive science are rather different. In fact, from a cognitive point of view, the purpose of a modeling enterprise is not merely to obtain model fits that are agnostic about psychological considerations; rather, parameters such as memory retention rates have specific psychological meaning, and should be treated as such. Moreover, and as mentioned above, it is unclear whether a model that relies on specific parameter values should be considered an ideal observer model.

repetition-pattern, because this is the only regularity present in the data. In contrast, when familiarized with *AA*di triplets, both *AAB* and *ABB* triplets are violations, since they do not conform to the /di/ regularity. Hence, infants might "expect" triplets to be consistent with *all* of the patterns they have picked up. If so, the role of the five additional familiarization triplets in Gerken's (2010) studies might be to familiarize infants with items not containing /di/, which, in turn, would allow them to reveal their learning of the repetition-pattern in the subsequent test phase, without being "surprised" to hear triplets not containing /di/.

Alternatively, items in edge positions might be more salient than repetitions (Gervain & Endress, in preparation), and, therefore, more likely to drive behavior compared to repetition-patterns. Note that rule saliency is not a placeholder to make the perceptual or memory primitives view consistent with the data. Rather, the more salient rule is empirically defined as the rule which participants choose when both rules are pitted against each other. (Such an empirically defined concept of saliency is well accepted among Bayesian modelers as well; see e.g. Frank & Goodman, 2012). If so, the role of the five additional familiarization triplets in Gerken's (2010) experiments would be to familiarize infants with violations of the more salient pattern, allowing them to reveal their sensitivity to the less salient one.

Marcus et al. (2007), Saffran et al. (2007)

Marcus et al.'s (2007) experiments can be explained with Marcus et al.'s (2007) account: humans preferentially attend to speech as opposed to non-speech items (e.g., Peña et al., 2003; Vouloumanos & Werker, 2004); as speech items are also sounds, the learning of repetition-patterns over speech-items might spill over to allow the recognition of repetition-patterns in other sounds. Likewise, Saffran et al.'s (2007) results follow from the truism that infants attend more to what they are interested in.

Gómez (2002), Kovács and Mehler (2009a), Dawson and Gerken (2009)

Like Frank and Tenenbaum's (2011) models, the perceptual or memory primitives account does not provide a good explanation of Gómez's (2002) data; nor does it provide any insight into why infants' ability to process repetition-patterns over musical stimuli changes over development (Dawson & Gerken, 2009). Likewise, I simply refer to Kovács and Mehler's (2009b) suggestion that bilingual infants might learn multiple rules due to their enhanced executive function.

In sum, it seems that most of the data simulated by Frank and Tenenbaum (2011) can be explained based on simple psychological mechanisms. However, like Frank and Tenenbaum's (2011) models, this account fails to provide an explanation of Gómez's (2002) and Kovács and Mehler's (2009b) data; unlike Frank and Tenenbaum's (2011) models, however, it is much harder to "patch" this account with further assumptions to fit it to the data, making it more verifiable and, therefore, more useful for discovering the nature of human rule-extracting capacities.

Conclusions

In 2002, Daniel Kahneman was awarded the Nobel Memorial Prize in Economic Sciences for showing that human behavior is not rational as assumed in most models of economic choice. Kahneman's work triggered the creation of behavioral economics, studying how actual humans make choices. Frank and Tenenbaum's (2011) paper is an example of a growing trend moving into the opposite direction, and arguing that humans are rational learners after all, making "optimal," Bayesian, decisions.

While several authors have criticized such models on theoretical grounds (Altmann, 2010; Bowers & Davis, 2012; Fitelson, 1999; Jones & Love, 2011; Marcus, 2008, 2010; Sakamoto et al., 2008), the assumptions of these models are rarely spelled out in plain English, which makes it hard to evaluate and test their psychological predictions. Here, I present a case study of a model in a domain that is arguably well-suited to Bayesian approaches, spell out its assumptions and predictions, and confront them with empirical data.

Frank and Tenenbaum (2011) attempt to account for various experimental results in the domain of rule-learning. According to Frank and Tenenbaum (2011), their models are psychologically implausible, but constitute computational-level, ideal-observer models of such results. However, while the analyses of the models confirmed that the underlying assumptions are psychologically implausible indeed, closer examination of the models revealed several reasons for which the models are difficult to accept as ideal-observer models. First, Frank and Tenenbaum (2011) fall short of their own modeling goals; neither do they make novel predictions nor do they detect any non-normative behavior. Second, there is no more support for the assumption underlying Frank and Tenenbaum's (2011) models — that humans learn the most specific, hardest-to-conform-to rule — than for an arbitrary "theory", such as that infants try to make experimenters happy. Third, Frank and Tenenbaum (2011) make a large number of crucial implementational assumptions that sometimes go as far as claiming that the infant perceptual system has different properties for the same stimuli, depending on whether they are presented during a familiarization or a test phase.

Accordingly, Frank and Tenenbaum's (2011) models do not provide an account of any of the results they attempted to model, unless the goal of psychological theory is to obtain psychology-agnostic data fits. The problematic model predictions are summarized in Table 1. The models' ability to learn any rule at all relies on the empirically unsupported assumption that hu-

mans preferentially learn the most specific, hardest-to-conform-to rules. Leaving aside this problem, their explanation of Endress et al.'s (2007) experiments predicts that humans cannot discriminate rising from falling contours (see Appendix B); their account of Marcus et al.'s (2007) data predicts that the infant perceptual system has different properties for the same stimuli, depending on whether they are presented during a familiarization or a test phase; and, when Frank and Tenenbaum's (2011) model parameters are taken seriously, their simulation results do not even reproduce the basic pattern of Gómez's (2002) results. Of course, it is possible to "patch" Frank and Tenenbaum's (2011) models to account for each of the experiments reviewed above. However, a model where the underlying representational assumptions need to be changed for every data point fails to generate scientific insight.

It should be noted that these problems are not specific to Frank and Tenenbaum's (2011) models. In fact, there are very few models that adhere to the modeling goals stated by Frank and Tenenbaum (2011). For example, Bayesian models of cognitive phenomena are usually not constructed to make predictions, but rather to fit existing data, giving modelers the opportunity to adapt their models to the experiments they attempt to fit (but see e.g. Orbán et al., 2008 for an exception). Moreover, these models typically have straightforward alternative explanations in terms of basic psychological mechanisms.

Further, Bayesian models of cognition are generally not used to detect non-normative behavior; in fact, except in the decision making literature and, in some cases, in the reasoning literature (e.g., Bes et al., 2012; Fernbach & Sloman, 2009), it is extremely rare that a Bayesian model of a cognitive phenomenon is published because the model does *not* account for the phenomenon. Moreover, even if a Bayesian model did not account for a cognitive phenomenon, it would be unclear whether this was due to unfounded assumptions (e.g., that learners prefer the most specific rules), or rather to genuine suboptimality. It would thus seem that Bayesian models of cognitive phenomena are rarely used as ideal observer models, and often share implausible assumptions similar to Frank and Tenenbaum's (2011) .

In contrast, and as discussed above, there is a satisfactory alternative account of the experiments based on simple psychological principles. This account is more piecemeal and less systematic than Frank and Tenenbaum's (2011), but appears to explain the data. For example, humans might learn some rules using a repetition-detector (Endress et al., 2007; Gómez, Gerken, & Schvaneveldt, 2000; Tunney & Altmann, 2001) or by attending to elements in sequence-edges (Endress & Mehler, 2009), and these mechanisms might constrain each other (Endress et al., 2005) and be constrained by various other factors, including phonological information (Pons & Toro, 2010; Toro, Bonatti, et al., 2008; Toro, Shukla, et al., 2008), syntactic processes (Endress & Hauser, 2009), executive function (Kovács & Mehler, 2009b), and probably many other aspects of our mental life. Such a collection of processes might not be elegant, and elude understanding based on higher-order principles. However, it might well reflect the true nature of our mental machinery. In fact, various authors have characterized cognition as an agglomeration of heuristics (Gigerenzer, Todd, & The ABC Group, 1999), as a "bag of tricks" (Ramachandran, 1990), as a kludge (Marcus, 2008) or as using a collection of primitive operations (Endress et al., 2009), each of which evolved to solve a particular problem in an organism's environment (Cheney & Seyfarth, 2007; Gallistel, 1990, 2000; Hauser, 2000). Before attempting to construct computational-level theories of cognitive operations, we thus first need to understand what is computed, and how.

Appendix A
Experiments 1 and 2: Do humans prefer the most specific rule compatible with the input?

In Experiment 1, I ask whether actual human learners conform to the central assumption that allows Frank and Tenenbaum's (2011) model to learn any rule at all, namely that learners prefer more specific, harder-to-conform-to rules over less specific ones (see main text for more details). To test this hypothesis, I familiarized human adults with either an AAB pattern or an ABB pattern, both carried by speech syllables. Following this, participants who had been familiarized with an AAB pattern had to choose between triplets of rhesus monkey vocalizations following an AAB pattern, and triplets of new human syllables following an ABB pattern; likewise, participants familiarized with an ABB pattern had to choose between triplets of rhesus monkey vocalizations following an ABB pattern, and triplets of new human syllables following an AAB pattern. In other words, participants had to choose between the repetition-pattern they had been familiarized with, and the rule that all items are carried by speech syllables.

The repetition-pattern is clearly more specific than the rule that all items are carried by syllables; after all, the latter is true of all possible syllable triplets, while the former is true only of a subset of them. The repetition-patterns remain more specific when considering the vocabulary of all familiarization and test items. Specifically, there are 8 familiarization syllables, 2 test syllables, as well as 2 rhesus vocalizations, leading to a total of $12^3 = 1728$ possible triplets, of which $10^3 = 1000$ contain only syllables, and of which $12 \times (12 - 1) = 132$ conform to the repetition-pattern.

Experiment 2 is a control experiment to Experiment 1, showing that, as in Marcus et al.'s (2007) experiments, human learners can discriminate repetition-

Table 1
Experiments Frank & Tenenbaum (2011) allegedly reproduced, reasons for the models' success, predictions inconsistent with available data, as well as alternative explanations based on simple psychological principles.

Experiment	**Reason for modeling success**	**Deviations from predictions**	**Alternative interpretation**
Marcus et al. (1999)	Repetition-patterns are more specific than control patterns.	Humans do not choose more specific patterns; see Appendix A.	Humans are equipped with a repetition detector that makes repetitions salient.
Endress et al. (2007)	• Repetition-patterns are learned as in the Marcus et al. (1999) case. • Ordinal patterns are difficult because they are consistent with various rules.	• Humans do not choose more specific patterns; see Appendix A. • Humans can discriminate rising from falling melodies; see Appendix B.	Humans are equipped with a repetition detector that makes repetitions salient.
Frank, Slemmer, et al. (2009)	• Repetition-patterns are learned as in the Marcus et al. (1999) case. • Multi-modal rules are more "specific." • Multi-modal rules are perceived better.	Humans do not choose more specific patterns; see Appendix A.	• Multi-modal stimuli are more variable. • Multi-modal stimuli provide more opportunities to learn. • Multi-modal stimuli provide more opportunities for success during test. • ...
Gerken (2006)	• Repetition-patterns are learned as in the Marcus et al. (1999) case. • Infants maintain a conjunction rule of two rules. • The "winning" conjunction rule is more "specific."	• If the unsupported assumption that infants maintain conjunction rules is not valid, no rule would be more specific. • Humans do not choose more specific patterns; see Appendix A.	• Some rules are easier to learn than others. • People expect items to conform to all rules that have been learned.
Gerken (2010)	• Repetition-patterns are learned as in the Marcus et al. (1999) case. • Infants maintain a conjunction rule of two rules.	• Humans do not choose more specific patterns; see Appendix A. • Humans should unlearn rules based on a *single* counterexample after thousands of positive examples.	• Some rules are easier to learn than others. • People expect items to conform to all rules that have been learned.
Marcus et al. (2007)	• Repetition-patterns are learned as in the Marcus et al. (1999) case. • Non-speech items have a misperception probability of 80% during familiarization, and 0% during test.	• Humans do not choose more specific patterns; see Appendix A. • Perception does not have different properties for familiarization items and test items. • There are other performance differences that are not observed empirically.	Humans preferentially process speech; see Marcus et al. (2007).
Saffran et al. (2007)	• Repetition-patterns are learned as in the Marcus et al. (1999) case. • Infants are more than four times more likely to misremember/misperceive items than in Marcus et al.'s (2007) experiments.	• Humans do not choose more specific patterns; see Appendix A. • See main text.	Infants who are more interested in stimuli attend more to these stimuli.
Gómez (2002)	• The "winning" rule is more "specific." • Participants misremember *exactly* 60% of the items.	• Humans do not choose more specific patterns. • The predicted results deviate qualitatively from the empirical data.	?
Kovács & Mehler (2009b)	• Repetition-patterns are learned as in the Marcus et al. (1999) case. • Bilinguals are better at learning two rules simultaneously because they are wired to learn multiple rules simultaneously.	• Humans do not choose more specific patterns. • The model ignores work on the executive function advantage in bilinguals.	See Kovács & Mehler (2009b).

patterns when these are carried by animal vocalizations.

Experiment 1: Do human adults prefer more specific rules?

Materials and method.

Participants Fourteen (9 females, mean age 25.4, range 19–34) native speakers of English participated in Experiment 1. They were recruited from the MIT community and received monetary payment in exchange for their participation. Half of the participants were assigned to the AAB condition, and half to the ABB condition (see below).

Apparatus Stimuli were presented over headphones using Psyscope X (http://psy.ck.sissa.it). Participants were tested individually in a quiet room. Responses were collected from pre-marked keys on the keyboard.

Stimuli During familiarization, *A* syllables were ga, li, ni and ta; *B* syllables were gi, la, na, ti. All syllables were pronounced by a male native speaker of American English. These syllables had an average duration of 627 ms (range: 477 – 727 ms), and were combined into *AAB* or *ABB* triplets, depending on the condition. Syllables in triplets were separated by 200 ms silences. This yielded 16 familiarization triplets.

During test, *A* syllables were wo and du, while *B* syllables were ru and ko. All syllables were pronounced by the same speaker as the familiarization syllables. These syllables had an average duration of 616 ms (range: 496 – 750 ms). *A* vocalizations were an aggressive call and a harmonic arch; *B* vocalizations were a scream and a coo call. These vocalizations had an average duration of 617 ms (range: 485 – 942 ms). However, I used only the syllable triplets *wo-wo-ru*, *du-du-ko* (*AAB*), *wo-ru-ru*, and *du-ko-ko* (*ABB*), as well as the rhesus triplets *aggressive-aggressive-scream*, *harmonic arch-harmonic arch-coo* (*AAB*), *aggressive-scream-scream*, and *harmonic arch-coo-coo* (*ABB*).

Procedure Participants were informed that they would hear some sound sequences, and were instructed to listen to them. Following this, they were presented with all 16 familiarization triplets played once in random order, with a silence of 1 s between triplets. Half of the participants were familiarized with *AAB* triplets, and half with *ABB* triplets.

Following this familiarization, participants were informed that they would hear pairs of sound sequences, and that they would have to decide which sequence in each pair was like the sequences they had heard before. They were advised that there was no "trick" in the experiment, and that they should just make their choices if they seemed obvious to them. Following this, they were presented with pairs of triplets, and had to choose which one was 'like' the triplets they had heard.

Participants familiarized with *AAB* triplets had to choose between *AAB* triplets carried by rhesus vocalizations, and *ABB* triplets carried by speech syllables. There were four test pairs of triplets, presented twice in different item orders. For participants familiarized with *ABB* triplets, the test items were constructed similarly.

Results and discussion. As shown in Figure A1, 10 our of 14 participants chose the syllable triplets with the incorrect repetition-pattern, while four had the opposite preference. On average, participants significantly preferred the triplets violating the repetition-pattern (i.e., they chose the syllable triplets, percentage of correct responses: $M = 24.1\%$, $SD = 34.5\%$), $t(13) = 2.81$, $p = 0.015$, Cohen's $d = 0.75$, $CI_{.95} = 56.0\%$, 95.8%. The

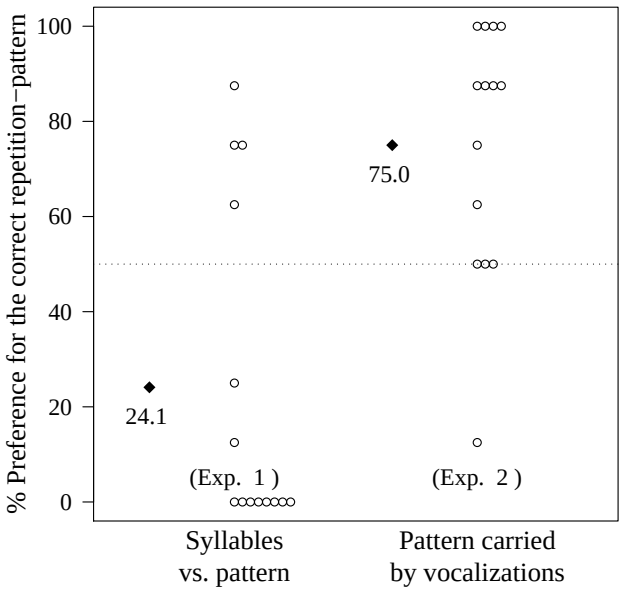

Figure A1. Results of Experiments 1 and 2. Circles represent individual participants, the diamonds the sample averages, and the dotted line the chance level of 50%. In Experiment 1, participants were familiarized with syllable triplets conforming to a repetition-pattern. Then, they had to choose between triplets of rhesus vocalizations conforming to the same pattern and triplets of syllables conforming to a different pattern. Most participants chose the syllable triplets even though they violated the repetition-pattern. In Experiment 2, participants were familiarized with the same triplets as in Experiment 1. Following this, they had to choose between triplets of rhesus vocalization that either had the same pattern as the familiarization items or a different pattern. They preferred the triplets with the familiar pattern.

results did not differ as a function of the familiarization pattern (i.e., *AAB* or *ABB*), $F(1,12) = 1.7$, $p = 0.221$, $\eta^2 = 0.122$. Hence, most actual human learners do not preferentially learn the most specific, hardest-to-conform-to rule, but rather whatever happens to be most salient to them.

However, before accepting this conclusion, it is necessary to establish that participants can detect repetition-patterns carried by rhesus vocalizations. This is tested in Experiment 2.

Experiment 2: Do human adults detect repetition-patterns in rhesus vocalizations?

Materials and method.

Participants Fourteen (7 females, mean age 25.6, range 19–34) native speakers of English participated in

Experiment 2. They were recruited from the MIT community and received monetary payment in exchange for their participation. Half of the participants were assigned to the *AAB* condition, and half to the *ABB* condition.

Procedure The familiarization was identical to that in Experiment 1. The test items were triplets of rhesus vocalizations that either conformed to an *AAB* pattern or and *ABB* pattern.

Results and discussion. As shown in Figure A1, participants readily choose the vocalization triplets with the correct repetition-pattern, (M = 75.0%, SD = 26.4%), $t(13) = 3.54$, $p = 0.004$, Cohen's $d = 0.95$, $CI_{.95}$ = 59.8%, 90.2%. The results did not differ as a function of the familiarization pattern (i.e., *AAB* or *ABB*), $F(1,12) < .01$, $p > .999$, $\eta_p^2 < .0001$. Hence, human adults readily discriminate repetition-patterns when they are carried by rhesus monkey vocalizations.

Appendix B
Experiment 3: Can humans discriminate rising from falling contours

Participants Fourteen (9 females, mean age 25.4, range 19–34) native speakers of English participated in Experiment 3. They were recruited from the MIT community and received monetary payment in exchange for their participation. Half of the participants were familiarized with *rising* contours, and half with *falling* contours.

Apparatus Stimuli were presented over headphones using Psyscope X (http://psy.ck.sissa.it). Participants were tested individually in a quiet room. Responses were collected from pre-marked keys on the keyboard.

Stimuli The stimuli were the same piano tones as used in Endress et al.'s (2007) Experiment 2. However, as shown in Figure 1, instead of being arranged into *lowest-highest-middle* and *middle-highest-lowest* triplets, they were arranged into rising contours (i.e., *lowest-middle-highest*) and falling contours (i.e., *highest-middle-lowest*).

There were 16 familiarization triplets and three test triplets for each pattern. The test triplets we combined into 9 test pairs.

Procedure The procedure and instructions were identical to Experiments 1 and 2. Participants were familiarized with all 16 familiarization triplets played once in random order, with a silence of 1 s between triplets. Half of the participants were familiarized with rising contours, and half with falling contours. Following this, they were presented with pairs of triplets, and had to choose which one was 'like' the triplets they had heard. In each trial, one triplet had a rising contour, and one a falling contour. The 9 test pairs were presented twice with different item orders. Test trials were presented in random order.

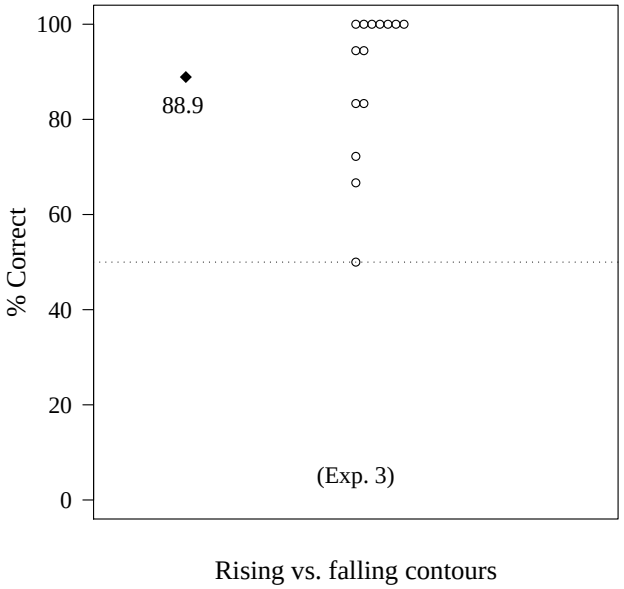

Figure B1. Results of Experiment 3. Circles represent individual participants, the diamond the sample average, and the dotted line the chance level of 50%. In contrast to the predictions of Frank & Tenenbaum's (2011) model, participants readily discriminate rising from falling contours.

Results and discussion. As shown in Figure B1, most participants were at ceiling discriminating rising from falling contours (percentage of correct responses: M = 88.9%, SD = 15.9%), $t(13) = 9.2$, $p < .0001$, Cohen's $d = 2.5$, $CI_{.95}$ = 79.7%, 98.1%. There was no difference between the familiarization condition (rising vs. falling), $F(1,12) = 0.5$, $p = 0.476$, $\eta^2 = 0.043$, ns. Hence, in line with much work in music perception, participants readily discriminate rising from falling contours.

Appendix C
References

Altmann, G. T. M. (2010). Why emergentist accounts of cognition are more theoretically constraining than structured probability accounts: comment on Griffiths et al. and Mcclelland et al. *Trends in Cognitive Sciences*, *14*(8), 340.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

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. (1995). Obligatory edge assignment in vision: The role of figure and part segmentation in symmetry detection. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(6), 1323–1343.

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.

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, E. (2012). Non-bayesian inference: causal structure trumps correlation. *Cognitive Science*, *36*(7), 1178–1203.

Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child Development*, *70*(3), 636–644.

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(4), 859–873.

Bialystok, E., & Craik, F. I. M. (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science*, *19*(1), 19-23.

Bialystok, E., Craik, F. I. M., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the simon task. *Psychol Aging*, *19*(2), 290–303.

Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. *Developmental Science*, *7*(3), 325–339.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414.

Brunswik, E., & Kamiya, J. (1953). Ecological cue-validity of 'proximity' and of other Gestalt factors. *The American Journal of Psychology*, *66*(1), 20–32.

Cheney, D., & Seyfarth, R. (2007). *Baboon metaphysics: the evolution of a social mind*. Chicago, IL: University of Chicago Press.

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: evidence from the ant task. *Cognition*, *106*(1), 59–86.

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.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University. (http://psychclassics.yorku.ca/Ebbinghaus/)

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision*, *2*(4), 324–353.

Endress, A. D., Carden, S., Versace, E., & Hauser, M. D. (2010). The apes' edge: positional learning in chimpanzees and humans. *Animal Cognition*, *13*(3), 483-495.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

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), 21001-6.

Endress, A. D., & Mehler, J. (2009). Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology*, *62*(11), 2187–2209.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology. General*, *134*(3), 406-19.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(3), 678-93.

Fitelson, B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66*, S362–S378.

Forstmeier, W., Martin, K., Bolund, E., Schielzeth, H., & Kempenaers, B. (2011). Female extrapair mating behavior can evolve via indirect selection on males. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(26), 10608–10613.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, *12*(4), 504–509.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371.

Gallistel, C. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Gallistel, C. (2000). The replacement of general-purpose learning models with adaptively specialized learning modules. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (2nd ed., pp. 1179–91). Cambridge, MA: MIT Press.

Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, *357*(1420), 419-448.

Geisler, W. S., & Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, *27*(3), 379 - 402.

Gerken, L. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74.

Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115*(2), 362-6.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(37), 14222-7.

Gigerenzer, G., Todd, P., & The ABC Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gill, J. (2008). *Bayesian methods: A social and behavioral sciences approach* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature*, *410*(6831), 930-3.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-6.

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109-35.

Gómez, R. L., Gerken, L., & Schvaneveldt, R. (2000). The basis of transfer in artificial grammar learning. *Memory and Cognition*, *28*(2), 253-63.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Hauser, M. D. (2000). *Wild minds: What animals really think*. New York: Henry Holt.

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), 161-7.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, *294*(5550), 2310–2314.

Hyams, N. (1986). *Language acquisition and the theory of parameters*. Dordrecht: 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*.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Kovács, A. M., & Mehler, J. (2008). *Regularity learning in 7-month-old infants under 'noisy' conditions: Adjacent repetitions vs. non-adjacent repetitions.* (Talk given at the 33rd Boston University Conference on Language Development)

Kovács, A. M., & Mehler, J. (2009a). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6556–6560.

Kovács, A. M., & Mehler, J. (2009b). Flexible learning of multiple speech structures in bilingual infants. *Science*, *325*(5940), 611–612.

Kovács, A. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science*, *12*(1), 48-54.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984.

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

Marcus, G. F. (2008). *Kluge: The haphazard construction of the human mind*. New York: Houghton Mifflin.

Marcus, G. F. (2010). Neither size fits all: comment on Mcclelland et al. and Griffiths et al. *Trends in Cognitive Sciences*, *14*(8), 346–347.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science*, *18*(5), 387-91.

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

Marr, D. (1982). *Vision*. San Francisco, CA: W.H.Freeman and Company.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule Learning by Rats. *Science*, *319*(5871), 1849-1851.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608 - 631.

O'Hagan, A. (1994). *Kendall's advanced theory of statistics* (1st ed., Vol. 2B: Bayesian inference). London, UK: Edward Arnold.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745–2750.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *255*(1342), 37-45.

Peña, M., Maki, A., Kovaci?, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., et al. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11702-5.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306-338.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607–642.

Pons, F., & Toro, J. M. (2010). Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition*, *116*(3), 361–367.

Ramachandran, V. (1990). Interactions between motion, depth, color and form: The utilitarian theory of perception. In C. Blakemore (Ed.), *Vision: Coding and effi-*

ciency (pp. 346–360). New York: Cambridge University Press.

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), 669-80.

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.

Seidl, A., & Johnson, E. K.(2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, *9*(6), 565-573.

Sigman, M., Cecchi, G., Gilbert, C., & Magnasco, M.(2001). On a common circle: Natural scenes and Gestalt rules. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(4), 1935-40.

Tenenbaum, J. B., & Griffiths, T. L.(2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–40; discussion 652-791.

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 and Psychophysics*, *70*(8), 1515–1525.

Tunney, R., & Altmann, G. T.(2001). Two modes of transfer in artificial grammar learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(3), 614-39.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059.

Vouloumanos, A., & Werker, J. F. (2004). Tuned to the signal: the privileged status of speech for young infants. *Developmental Science*, *7*(3), 270–276.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H.(2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245 - 272.