Is there evidence for the size principle? A critical review

Ansgar D. Endress
Universitat Pompeu Fabra, Barcelona, Spain
City University, London, UK

Draft of September 13, 2013. Please do not quote without permission.

According the size principle, learners choose the less likely of two hypotheses if it is compatible with the data (e.g., Tenenbaum & Griffiths, 2001). While the size principle has played important roles in recent theorizing, there is limited evidence in support of it. Frank (2013) provided a list of experiments that are presumably the strongest cases for the size principle (Denison, Reed, & Xu, 2013; Gweon, Tenenbaum, & Schulz, 2010; Navarro, Dry, & Lee, 2012; Xu & Tenenbaum, 2007a, 2007b). I critical review these experiments, and show that they either rely on extraneous assumptions without which they would not fit the data, sometimes do not fit the data in the first place, and that, when they fit the data, the data often have more plausible interpretations that do not appeal to the size principle. I conclude that there is no support for the hypothesis that learners use the size principle to decide among hypotheses, and that basic psychological principles provide a better account of these data.

Introduction

When making an inference, we need to decide among competing hypotheses based on limited data. This raises the problem of how we choose the correct hypothesis although other hypotheses are just as consistent with the data (e.g., Goodman, 1955; Hume, 1739/2003; Wittgenstein, 1953). Empirically, humans seem to be fairly successful at choosing appropriate hypotheses. For example, infants make the correct "inferences" to learn their native language, and, over longer time periods, our inferences about the natural world seem to be correct enough for cars, computers, CT scans and so on to work. In other cases, however, our inferences are fairly poor (Kahneman & Tversky, 1996; Tversky & Kahneman, 1974). The question thus is why we are so good at making some inferences but not others, and what the psychological principles are that let us make such inferences.

Following Tenenbaum and Griffiths (2001), a subset of such inference problems have been proposed to be solvable using the "size principle." If we have to choose between two hypotheses that are equally consistent with the examples we have seen, we choose the hypothesis that is a priori more unlikely (see Tenenbaum & Griffiths, 2001, for a justification). In fact, a related strategy has been proposed in the context of natural language acquisition. Specifically, some scholars proposed that infants learn the most restrictive grammar consistent with what they hear (or with what they see in the case of sign languages; Hyams, 1986; Manzini & Wexler, 1987). In some reading of this proposal, these authors proposed that humans evolved to acquire language following a sequence of acquisition steps that is consistent with the most restrictive grammar given the input, using specific "triggers" to move from a more restrictive grammar to a more permissive one. The underlying idea is that the triggers allow learners to "conclude" that their current grammars are not general enough, and to adjust them appropriately, while it is unclear how they could even notice that they started out with a grammar that is too general.

While this idea seems plausible in the domain-specific case of language acquisition, it has been suggested in the literature following Tenenbaum and Griffiths (2001) that the size principle can be used for arbitrary inferences (at least judging from the cases it has been applied to). In fact, the size principle has been applied to domains from the basic probabilistic inferences of young infants to language acquisition to social cognition (see below for references). For example, Frank and Tenenbaum (2011) modeled how infants might learn certain grammar-like regularities. They proposed that, when infants have to choose between multiple regularities that are consistent with examples they have heard, they choose the one that has fewer potential items conforming to it. According

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.

to their model, infants might encounter a total of three syllables. Before encountering any syllable triplet, infants know that the three syllables allow for a total 27 triplets, that 6 of these triplets follow an ABB pattern (e.g., *pu-li-li*), that 3 of these triplets follow an AAA pattern (where all three syllables are identical), as well as the number of triplets that would conform to any conceivable rule. They then use the number of triplets that are consistent with each rule to choose among possible generalizations. However, they did not provide any suggestion for how infants might know the number of triplets that conform to each generalization.

I (Endress, 2013) analyzed these models in detail, and showed that they did not account for the data they were designed to account for (see Frank (2013) and Endress (under review) for discussion). One of the conclusions that followed from my analysis as well as from novel experiments was that there is no convincing evidence that we use the size principle to make arbitrary inferences. However, Frank (2013) proposed that several papers provide evidence for the size principle (Denison et al., 2013; Gweon et al., 2010; Navarro et al., 2012; Xu & Tenenbaum, 2007a, 2007b). Below, I critically review these papers, and show that they provide no evidence for the size principle.

The objective of this review is twofold. First, given the importance of the size principle in recent theorizing, it is critical to assess whether there is really convincing evidence in favor of this hypothesis. Second, the size principle is intimately linked to Bayesian approaches to cognition, which, in turn, have been proposed to account for a growing variety of data (see e.g., Bowers & Davis, 2012; Endress, 2013; Jones & Love, 2011, for critical reviews). However, I have argued that, when Bayesian models of cognitive phenomena are reported, they often rely on extraneous assumptions without which they would not fit the data, sometimes do not fit the data in the first place, and that, if they fit the data, the data often have more plausible interpretations from common-sense psychology (e.g., Endress, Dehaene-Lambertz, & Mehler, 2007; Endress, Nespor, & Mehler, 2009). Hence, I will use the aforementioned papers to provide another case study for the superiority of explanations based on common-sense psychology compared to Bayesian models of cognition.

Probabilistic inferences in infancy

Before discussing the aforementioned papers in more detail, it is important to address the relation between the size principle and probabilistic abilities in early infancy. Indeed, infants have remarkable knowledge of probabilistic principles. For example, Téglás, Girotto, Gonzalez, and Bonatti (2007) presented infants with displays in which two kinds of objects moved randomly inside an enclosure. Importantly, the enclosure had a hole where the objects could leave. One kind of object had a higher cardinality than the other. Téglás et al. (2007) showed that infants are less surprised to find an object outside the enclosure when it belonged to the more numerous class than when it belonged to the less numerous one, which is consistent with the inferences one would draw if the objects were randomly picked from the enclosure. Subsequent work has supported similar conclusions(e.g., Téglás et al., 2011; Xu & Garcia, 2008; Xu & Denison, 2009). Further, infants can use these abilities to detect non-random behavior in agents (Kushnir, Xu, & Wellman, 2010).

While such results attest to the impressive probabilistic abilities of young infants, and show that infants can, in some situations, form expectations about outcomes based on a priori considerations, they do not provide any evidence that infants (or adults for that matter) choose hypotheses based on the number of items that are compatible with them, especially if most of these items are never presented. In fact, in these experiments, there simply are no hypotheses infants could choose based on the number of items they are compatible with.

Xu and Tenenbaum (2007b), Navarro et al. (2012)

One of the strongest potential pieces of evidence in favor of the size principle comes from Xu and Tenenbaum's (2007b) experiments. They asked how learners assign meaning to novel nouns, and under what condition they would choose a meaning at the subordinate category level (e.g., "Dalmatian"), at the basic-level category level (e.g., "dog"), or at the superordinate category level (e.g., "animal").

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

The test screen comprised 8 items of each of three superordinate categories (i.e., animals, vegetables and vehicles). Within each category, there were 2 examples of the same subordinate category (e.g., two other Dalmatians), 2 examples of the same basic-level category (e.g., two non-Dalmatian dogs), and 4 examples of the same superordinate category (e.g., four non-dog animals). If participants concluded that fep meant "animal", they should choose all eight pictures of that category; if they concluded that it meant "dog," they should choose the four corresponding pictures; and if they concluded that it meant "Dalmatian," they should choose only the other two Dalmatians.

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

These results supposedly support the size principle. Indeed, there are more animals than there are dogs, and there are more dogs than there are Dalmatians. Hence, based on the size principle, one would expect a preference for inferences at the subordinate level, because subordinate categories contain the fewest possible referents. Below, however, I will show that this conclusion is empirically unwarranted.

An explanation based on language acquisition research

While Frank (2013) and Xu and Tenenbaum (2007b) take these results as evidence of the size principle, and suggest that language learners might deploy similar computations in the service of word learning, it is questionable whether such an approach would be viable for natural language acquisition. In fact, natural language learners are rarely shown test screens explicitly providing them with the 9 possible meanings of a novel word as well as with the number of elements of each category; rather, learners acquire a words whenever they encounter a situation that is conducive for inferring its meaning (Medina, Snedeker, Trueswell, & Gleitman, 2011). This raises the question of how learners could possibly estimate the number of elements of a category.

Even if they had access to this information, it is questionable whether young infants could process it. In fact, there are an estimated 75 million dogs in the United States (American Humane Society U.S. Pet Population Fact Sheet, http://www.americanhumane.org/assets/pdfs/pets-fact-sheet.pdf, retrieved on 9/10/13). It is an entirely open question whether infants can process numbers of this magnitude, or which other information they might possibly exploit according to Xu and Tenenbaum (2007b). As a result, it is questionable whether Xu and Tenenbaum's (2007b) account based on the size principle would scale up to actual language acquisition.

However, prior research in language acquisition might provide a solution to such problems. For example, it has been suggested that certain "triggers" might lead to changes in the hypotheses learners entertain. If so, learners might assume by default that word meanings correspond to a basic-level category. This might not be the case when the exemplar is a "bad" example of a category; for example, people consider peppers much less of a vegetable than, say, carrots (e.g., Armstrong, Gleitman, & Gleitman, 1983), and learners might infer that the meaning corresponds to the subordinate category instead. Further, what learners consider good and bad examples of a category might change with development, and as a function of their experience, which, in turn, might influence the kinds of inferences learners make about words.

If learners preferentially assign labels at the basic level, how might they infer labels at other levels than the basic category one? Possibly, when they notice that the basic-level interpretation is incorrect (e.g., because a cat is an animal but not a dog such that the interpretation "dog" is no longer tenable), they infer that the meaning corresponds to a superordinate category. In contrast, when the variability of the exemplars is less than expected from a basic-level category (e.g., because learners are not shown the best examples), they might conclude that the label corresponds to a subordinate category. (This model still faces the problem that learners need to figure out that the relevant distinction is, say, between dogs and animals rather than between other concepts associated with the stimuli that learners experience (Quine, 1960). However, this problem is shared by Xu and Tenenbaum's (2007b) account.) Leaving aside the problem that infants would need to figure out the relevant inferences in the first place, this model thus accounts for Xu and Tenenbaum's (2007b) data, without making any use of the size principle or Bayesian computations.

For completeness, I will now show that, even assuming the rest of Xu and Tenenbaum's (2007b) formalism, their results do not provide any evidence for the size principle.

An explanation based on Xu and Tenenbaum's (2007b) assumptions about the information learners consider

As mentioned above, language learners are typically not shown test screens that explicitly provide them with the 9 possible meanings of a novel word as well as with the number of elements of each category before they make an inference about a word's meaning. For completeness, I will now show that Xu and Tenenbaum's (2007b) data do not provide any evidence for the size principle even espousing their assumption that learners make rational inferences based on the examples and possible referents they are presented with.

In fact, their results can be account for by a very simple hypothesis: participants might choose a category that is (i) consistent with the examples they have seen, and (ii) where the examples are most similar to the other items in the category.

The first assumption prevents learners from considering, say, the meaning "Dalmatian" when presented with non-Dalmatian dogs, because the label would not fit the examples. To illustrate the second assumption, I constructed a simple similarity score, using the number of shared category levels of two items as a proxy of their similarity (i.e., subordinate, basic and superordinate). That is, two Dalmatians have a similarity score of 3 (because they share all 3 levels), a Dalmatian and a non-Dalmatian dog have a similarity score of 2, and a Dalmatian and a cat have a similarity score of 1. This similarity score reflects the intuition that items from the same subordinate category (e.g., two poodles) tend to be more similar than two items from the same basic-level category (e.g., a poodle and a Labrador), which,

Examples		Inference: "fep means"	Similarity to other dalmatians (N=2)	non-dalmatian dogs (N=2)	non-dog animals (N=4)	average
Dalmatians		Dalmatian	3			3
		Dog	3	2		2.5
		Animal	3	2	1	1.75
3 Dogs	For dalmatian (N=1)	Dog	3	2		2.5
		Animal	3	2	1	1.75
	For other dogs (N=2)	Dog	2	2		2
		Animal	2	2	1	1.5
	Average	Dog	2.333333333	2		2.166666667
		Animal	2.333333333	2	1	1.583333333

Figure 1. Similarity between training exemplars and other category members in Xu & Tenenbaum's (2007b) experiments. Here, I use the number of category-levels that two items share as a proxy of similarity.

in turn, tend to be more similar than two items from the same superordinate category (e.g., a poodle and a bear).

As shown in Figure 1, this simple model predicts Xu and Tenenbaum's (2007b) results, without using the size principle at all. When the examples are consistent with a subordinate-level category, the similarity is highest for items within such a category; and when they are consistent only with a basic-level category, the similarity is highest in such a category as well. Hence, Xu and Tenenbaum's (2007b) data present no evidence for the size principle.

Inferences as a function of the number of examples

Xu and Tenenbaum (2007b) report another result that, at first sight, seems to provide strong evidence for the size principle. Specifically, they show that, when participants are familiarized with a single Dalmatian, they conclude that fep means "dog;" in contrast, when familiarized with three Dalmatians, they conclude that fep means "Dalmatian." There are at least two straightforward explanations for this. First, participants might have a tendency to use basic level categories (see Xu & Tenenbaum, 2007b for discussion), and conclude that the lone Dalmatian is an exemplar of the "dog." When shown three Dalmatians, however, they might be surprised that all of the dogs are Dalmatian (e.g., because the examples do not correspond to the most prototypical dogs), and change their inference accordingly.

Second, Xu and Tenenbaum's (2007b) formal explanation does not provide any evidence for the size principle either. According to their model, the likelihood of each category given all exemplars is the product of the likelihoods of each category given the individual exemplars. Importantly, one of the factors in the individual likelihoods comes from the size principle, and is inversely proportional to the number of items in the category. As a result, with more exemplars, the influence of the set size is more pronounced, and should favor smaller categories (e.g., subordinate categories if these are consistent with the data). However, this is true for any probability smaller than 1 that is raised to a power corresponding to the number of exemplars.

For example, if the aforementioned similarity score is converted to a probability score, the same qualitative predictions follow. That is, the a priori likelihood of a meaning is largest when the similarity between the training examples and the other items to which the meaning applies is largest; as such, this probability score favors small (i.e., subordinate) categories. Hence, when it is raised to the third power (due to the three examples), the preference for the subordinate category will be more pronounced. As a result, what guarantees the narrowing of the generalizations between one and three exemplars is not the size principle but rather the rest of Xu and Tenenbaum's (2007b) formalism.

A similar conclusion applies to the results reported by Navarro et al. (2012). "Narrowing" of the inferences occurs simply due to raising probabilities to a power corresponding to the number of examples, but not due to the size principle per se. As such, neither Xu and Tenenbaum's (2007b) nor Navarro et al.'s (2012) results provide evidence for computations involving the size principle.

Xu and Tenenbaum (2007a)

Like Xu and Tenenbaum (2007b) , Xu and Tenenbaum (2007a) investigated the conditions under which learners assume that a verbal label refers to a subordinate category or a basic-level category, respectively. Participants (adults and 3-to 4 year olds) were presented

with a display showing items from two categories. The categories were defined by their shapes, and were spatially grouped together. Within each category, there were three subordinate categories of 4 items each. Items in different subordinate categories shared their shape, but differed in texture.

Xu and Tenenbaum (2007a) presented two conditions. In the teacher-driven condition, the experimenter pointed to 3 items of the same subordinate category, and labeled them with a novel word (e.g., a blicket). In the learner-driven condition, the experimenter labeled only 1 object, and then encouraged participants to point to two other blickets. Results showed that all but one participant pointed to two other items from the same subordinate category.

Following this, participants were pointed to 5 more objects, and asked whether these were blickets as well. Results showed that participants in the teacher-driven condition were more likely to infer that "blicket" refered to the subordinate category than participants in the learner-driven condition. Xu and Tenenbaum's (2007a) explain the preference for the subordinate category in the teacher-driven condition as follows. Since the teacher knows the meaning of the word, she will choose example objects to which this meaning applies. Hence, due to the size principle, if all examples are consistent with a subordinate category, subordinate category should be preferred, because the number of items in that category is lower than that in a basic-level category. In the learner-driven condition, in contrast, the learner does not know the meaning of the word; hence, the size principle does not apply to favor smaller categories, such that the learner should be less likely to infer a subordinate-level meaning than in the teacher-driven condition.

It should first be noted that Xu and Tenenbaum's (2007a) do not in fact provide a test of the size principle, because they do not manipulate the number of items that are consistent with a hypothesis; rather the size principle is, according to these authors, applicable in the teacher-driven condition, but not in the learner-driven condition. Crucially, however, Xu and Tenenbaum's (2007a) results are inconsistent with their conclusions. As mentioned above, all but one participant in the learner-driven condition selected items from the same subordinate category when asked to find other blickets; given that Xu and Tenenbaum's (2007a) model predicts that participants in the learner-driven condition should favor a basic-level interpretation, and that there are more than twice as many candidate blickets from different subordinate categories, one would expect them to preferentially choose items from different subordinate categories.

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

Gweon et al. (2010), Denison et al.'s (2013)

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

The critical manipulation was whether the majority of the balls inside the box was blue or yellow, and how many balls the experimenter picked from the box. In some conditions, the experimenter extracted three blue balls. Results showed that infants squeezed the yellow ball more often when the three squeaky blue balls had been extracted from a population with 75% blue balls, than when they had been extracted from a population with 25% blue balls. In contrast, when only one blue ball was extracted from a population of 25% blue balls, infants squeezed the ball as much as when three blue balls were extracted from 75% blue balls. In a crucial control condition, three blue balls were ostensibly drawn by chance from the box with 25% blue balls. In that condition, infants did not suppress squeezing the yellow ball.

Is Gweon et al.'s (2010) model consistent with the data?

To explain their data, Gweon et al. (2010) propose that infants consider the four possibilities spanned by two factors: (i) Is the teacher is cooperative and picks the balls only from the squeaky ones, or is she nasty, and picks from all balls irrespectively of squeakiness? (ii) Are all balls squeaky, or only the blue ones? Infants would then compute likelihoods of the results of

the experimenter's actions according to all four possible scenarios, and compare these likelihoods to decide whether or not to squeeze the yellow ball.[1] This likelihood ratio is given by, with β being the proportion of blue balls and α being a parameter that is irrelevant for the current purposes:

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

While this model is extremely complex, and postulates important processing abilities that might or might not be available to infants, it is inconsistent with the data, for two reasons. First, the model always concludes that it is more likely that only blue balls are squeaky. In fact, it is easy to see that $L_\alpha = 1$ for $\alpha = 0$ or $\beta = 1$, and that $L_\alpha < 1$ for all $\alpha > 0$ and $\beta < 1$. Hence, Gweon et al.'s (2010) model predicts that infants should never squeeze the yellow ball at all. Further, it is easy to see that, irrespective of the proportion of blue balls, this effect should be more pronounced when more balls are drawn from the container, and that, eventually L_α goes to 0.[2] As a result, it seems fair to conclude that Gweon et al.'s (2010) model does not account for the fact that infants squeeze the yellow ball in the first place, and rather predicts that infants should never squeeze it.

Second, the model assumes that infants desire to find *blue* squeaky balls, as opposed to just desiring to find balls that squeak irrespective of their color. Plausibly, however, what infants really care about when squeezing a ball is whether it squeaks. As shown in Appendix A, an improved version of Gweon et al.'s (2010) model that assumes that infants care about squeaky balls irrespective of their color predicts that infants should be more likely to squeeze the yellow ball when only 25% of the balls are blue, which is just the situation where infants are *less* likely to squeeze it. (However, in contrast to Gweon et al.'s (2010) model, the improved model accounts for the fact infants squeeze the yellow ball in the first place).

It should be noted that the failure of the improved model does not favor Gweon et al.'s (2010) original account; after all, the alternative model correctly predicts that infants should squeeze yellow ball, while the original one does not. Rather, the different versions of the model illustrate that the model behavior is not driven by the Bayesian machinery or the size principle, but rather by extraneous assumptions about what infants are most interested in.

An account based on common-sense psychology

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

Conclusions

Following Tenenbaum and Griffiths (2001), the size principle has been used to model a variety of cognitive phenomena. These models are intimately linked to Bayesian approaches to cognition. However, another growing literature suggests that models based on psychological considerations might provide a better account for empirical data (e.g., Bowers & Davis, 2012; Endress, 2013; Jones & Love, 2011).

Here, I review the strongest evidence for the size principle so far, and show that these experiments do not provide any support for it. Rather, they have alternative explanations in basic psychological considerations. As such, it appears preferable to use psychological explanations of psychological phenomena rather than formalisms that fit the data only due to implausible assumptions.

Appendix A
An improved version of Gweon et al.'s (2010) model

The "success" of Gweon et al.'s (2010) model critically depends on an assumption about what infants care about when they see a squeaky ball. Gweon et al. (2010) assume that infants estimate the likelihood of observing squeaking *blue* balls. However, it is plausible that they are mostly interested the squeakiness of the balls, irrespective of their color. In this case, Gweon et al.'s (2010) model predicts the opposite behavior.

[1] Frank presumably claims that Gweon et al.'s (2010) data supports the size principle because infants appear to know that blue balls are more likely to be drawn out of a box with a majority of blue balls than out of a box with a majority of yellow balls (Téglás et al., 2007). However, this ability is arguably unrelated to the ability postulated by FT, namely to generate all events consistent with a hypothesis to choose the more unlikely one.

[2] Indeed, the partial derivative $\partial_n L_\alpha = \frac{\alpha\beta^n \ln\beta}{(\alpha+(1-\alpha)\beta^n)^2}$ is strictly smaller than 0 since $\ln\beta$ is smaller than 0 for $\beta < 1$. Further, $\lim_{n\to\infty} L_\alpha = 0$.

Given a proportion β of blue balls, one can derive different likelihoods for the teacher picking three squeaky blue balls. These likelihoods are presented in Figure A1. In the two middle columns, I present the likelihoods from Gweon et al.'s (2010) model, assuming that infants desire *blue* squeaky balls. In the two right-most columns, I present an improved model holding that infants desire squeaky balls, but that they do not care about color. The likelihood of picking n squeaky blues balls can be obtained by averaging across the choice strategies of the teacher (see Gweon et al., 2010, for a justification of this average).

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

$$L = \frac{2\beta^n}{1+\beta^n}. \quad (2)$$

(Compared to equation (1), and following Gweon et al. (2010), I set α to .5) When 75% of the balls are blue, and the teacher picks 3 balls, the ratio is .59; when 25% of the balls are blue, the ratio is .03. Hence, leaving aside the fact that the model predicts that infants should never squeeze yellow balls, they should be more likely to squeeze yellow balls when 75% of the balls are blue.

The improved model, where infants care only about the squeakiness of the balls, but not their color, reverses the predictions. As can be seen from Figure A1, the corresponding likelihood ratio is:

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

With Gweon et al.'s (2010) α parameter, the likelihood ratio would be given by:

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

It is easy to see that $L'_\alpha = 1$ for $\alpha = 1$ or $\beta = 1$, and $L'_\alpha > 1$ for $\alpha < 1$ and $\beta < 1$. Hence, this model accounts for the fact that infants have a tendency to squeeze balls irrespective of their color.[3]

However, when 75% of the balls are blue, the ratio is 1.41, while it is 1.97 when 25% of the balls are blue. Hence, the improved model (incorrectly) predicts that infants should be more likely to squeeze yellow balls when 25% of the balls are blue.

As mentioned above, the failure of the improved model does not support Gweon et al.'s (2010) original model, because the original model fails to account for the fact that infants squeeze the yellow ball. Rather, this models illustrate that the predictions are not driven by Bayesian computations or the size principle, but rather by extraneous assumptions that are also part of their models.

Appendix B
References

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

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

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

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

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

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

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

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176.

Endress, A. D. (under review). How are bayesian models really used? *Cognition*.

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

Frank, M. C. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to. *Cognition*, *128*(3), 417–423.

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

Goodman, N. (1955). *Fact, fiction and forecast*. Cambridge: Harvard University Press.

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

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

Hume, D. (1739/2003). *A treatise of human nature*. Project Gutenberg.

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

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

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

Figure A1. Likelihoods of the teacher picking n balls of interest out of a box with a proportion of β blue balls. The middle two columns present Gweon et al.'s (2010) model, in which infants seek blue squeaky balls. The likelihood ratio in favor of the hypothesis that all balls are squeaky is given by $2\beta^n/(1 + \beta^n)$. The rightmost columns present an alternative model, according to which infants just care about the squeakiness of the balls, irrespective of color. In that case, the likelihood ratio is $2/(1 + \beta^n)$.

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*.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*(3), 582-91; discusion 592-6.

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

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

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

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.

Quine, W. v. (1960). *Word and object*. Cambridge, MA: MIT Press.

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

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*(4481), 1124-31.

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

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.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*(1), 97–104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5012–5015.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in bayesian word learning. *Developmental Science*, *10*(3), 288–297.

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