# Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013)

Michael C. Frank

Department of Psychology, Stanford University

Many thanks to Ali Horowitz, Noah Goodman, Chigusa Kurumada, Gary Marcus, Ellen Markman, Jay McClelland, Josh Tenenbaum, and Dan Yurovsky for valuable feedback on previous drafts.

Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall, Jordan Hall (Building 420), Stanford, CA 94305, tel: (650) 724-4003, email: mcfrank@stanford.edu

**Abstract**

A recent probabilistic model unified findings on sequential generalization via independently-motivated principles of generalization (Frank & Tenenbaum, 2011). Endress (this issue) critiques this work, arguing that learners do not prefer more specific hypotheses (a central assumption of the model), that "common-sense psychology" provides an adequate explanation of rule learning, and that Bayesian models imply incorrect optimality claims but can be fit to any pattern of data. Endress's response raises valuable points about the importance of mechanistic explanation. Nevertheless, the specific critiques of our work are not supported, and Endress undervalues the importance of formal models. I argue that probabilistic modeling provides a powerful framework for describing cognition but should not be used as evidence for optimality claims.

## Introduction

How do you reverse engineer an alien computer? Figuring out how it works requires moving back and forth between what you can learn about its individual parts and broader hypotheses about its function and governing principles. The general theory of computation leads to questions about the artifact's inputs, outputs, and methods for storing information (Hopcroft, Motwani, & Ullman, 1979). But since computational systems can store their state in processes as diverse as symbols on a tape or weights between neurons (McCulloch & Pitts, 1943), a high-level understanding of the device provides only general constraints on lower-level hypotheses. In Marr's (1982) terms, a *computational* level understanding of the system needs to be integrated with both a model of the system's sub-components (the *algorithmic* level) and, critically, an understanding of the individual units of the system (the *implementational* level). Each of these levels of representation contributes to the ability to repair, duplicate, and extract general insights from the artifact.

Reverse engineering the human mind requires the same attention to multiple levels of abstraction. A wide range of theorists have recognized that insights into the workings of complex systems like perception, memory, and language require an understanding of the general operating principles of the system (e.g., Chomsky, 1995; Anderson, 1990; Marr, 1982). Probabilistic models, which use tools from Bayesian statistics and machine learning to describe such systems, represent a promising framework for exploring high-level descriptions of cognitive processes (Chater, Tenenbaum, & Yuille, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).[1]

Although probabilistic models have grown tremendously in popularity in recent years, they have also attracted significant criticism (Jones & Love, 2011; Bowers & Davis, 2012). Chief

[1]I use the terms "probabilistic" and "Bayesian" synonymously. I prefer "probabilistic," as it better describes the key virtue of these models: that they use probability as a single framework for integrating across widely varying tasks, representations, and constraints.

among these criticisms is that these models imply a claim that the mind itself is *rational* or even *optimal*. A claim of optimality entails that a particular cognitive process provides the best possible solution relative to some problem. The weaker claim of rationality suggests that the process provides a logical, well-designed solution to a problem, perhaps relative to limitations on cognitive resources like memory or computation. These claims—especially the optimality claim—strike many authors as unsupported and unfalsifiable, given that the particular problem being solved and the assumptions of the model solving it are rarely specified independently. Endress's (2013) article echoes these criticisms of optimality claims, applying them to Frank and Tenenbaum's (henceforth, FT; 2011) models of sequential rule learning and providing additional theoretical and empirical arguments.

## "Rule learning" and Endress's critique

FT used probabilistic models to describe infants' and adults' ability to learn sequential regularities in auditory stimuli, a learning ability that may be linked to language acquisition (Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Peña, Bonatti, Nespor, & Mehler, 2002; Marcus, Fernandes, & Johnson, 2007). In "rule learning" paradigms (Marcus et al., 1999), learners hear strings of syllables like "wo fe fe," instantiating simple regularities (e.g., in this case $ABB$, or "last syllable repeats"). They are then tested on their ability to generalize these regularities to novel syllable strings. Experiments across a variety of ages, modalities, and rule types provide a rich body of data that can be explored for insights about how infants and adults make such generalizations (e.g., Marcus et al., 1999; Endress, Scholl, & Mehler, 2005; Gerken, 2006; Marcus et al., 2007; Johnson et al., 2009).

In FT, we created a set of three probabilistic models that made predictions about learners' performance across a wide range of empirical results. All three of these models were based on the assumption that learners prefer more specific hypotheses (the "size principle" of Tenenbaum & Griffiths, 2001), but they varied in their complexity. Model 1, the simplest, made inferences

directly from the input data, but it always learned the correct rule perfectly. Model 2 added a single free parameter that controlled noise in memory or perception, allowing the model to produce quantitative predictions. Model 3 introduced the possibility that more than one rule could be operating simultaneously. Despite the apparent diversity of results in the literature, these simple models sufficed to capture a wide range of empirical data.

Endress (2013) argues that our models do not provide a good account of rule learning behavior, contesting both the general framework we used and the specifics of our simulations. For purposes of space I will focus on several primary points made in this critique:

1. Learners prefer more salient rules rather than more specific hypotheses,

2. "Common-sense psychology" provides an adequate explanation of rule learning,

3. The use of probabilistic models implies an optimality claim, which is problematic as these models (and FT in particular) can be fit to any pattern of data, and

4. The use of free parameters is inappropriate in cognitive modeling more generally.

Endress's article raises important questions about how computational principles can be instantiated in human minds, and I am in agreement that claims of optimality on the basis of probabilistic modeling are rarely appropriate (indeed, FT did not make such a claim). Nevertheless, as I will describe below, the criticisms of FT are not valid: The empirical evidence given against specificity is not compelling, the proposed alternative account is circular, and FT did not overfit the data.

## Do learners prefer more specific rules?

At the heart of Endress's critique is the claim that "humans do not prefer more specific patterns." This claim is important because the size principle (Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007b)—the principle that hypotheses are weighted proportional to their specificity—was the major explanatory assumption in FT's models.

An independent body of evidence supports the use of the size principle as a description of

word learning and categorization (Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007b, 2007a). For example, in the word learning tasks used by Xu and Tenenbaum (2007b), adults and children saw either one or three examples of a category and were asked to make judgements that revealed the specificity of their generalization. Presented with one example, they showed gradient generalization, but after seeing three examples, their judgments were consistent with the most specific category that fit the data they observed. In addition, in an analysis of a large-scale similarity judgment dataset, Navarro and Perfors (2010) found that adults' ratings were well-fit by a model that used specificity to weight features. These datasets both provide strong evidence for the importance of the size principle, but are not discussed by Endress.

Instead, in support of the claim that humans do not prefer more specific rules, Endress conducted an experiment in which participants were familiarized with human speech syllables in an $AAB$ or $ABB$ pattern. At test they were asked to choose between pattern-incongruent human syllables, or pattern-congruent strings instantiated in rhesus monkey vocalizations, pitting consistency with the pattern regularity (e.g., $AAB$ vs. $ABB$) against consistency in the modality of presentation (human speech vs. monkey vocalizations). Participants largely preferred test trials consistent with the modality of presentation.

These data show the size principle is not the only factor affecting category judgments, but do not provide evidence against the size principle. Test trials in the experiment gave participants the opportunity to select either a modality match or a pattern match. The preference for the modality match suggests that modality was the stronger of the two cues in this particular case. Such a trend is not surprising, given the unexpected and striking nature of hearing monkey vocalizations in what participants might guess to be a language-related task. In fact, the probabilistic perspective provides a valuable tool for understanding how learners in Endress's experiment integrated the salience of a particular hypothesis and its specificity. For example, Frank and Goodman (2012) described a model of language comprehension that gave a probabilistic integration of these two factors. These same methods could easily be applied to the

models in FT as well.

In sum, Endress's data are consistent with probabilistic models and do not provide evidence against specificity, and Endress does not provide an alternate account of the additional evidence for specificity. Nevertheless, the critique raises an interesting question about how rule specificities could be computed during a short laboratory experiment. Endress rejects the proposal that learners enumerate the full set of strings consistent with each rule, but research on numerical cognition suggests that adults and infants need not enumerate to make quick and accurate judgments about the cardinality of sets (Xu, 2002; Whalen, Gallistel, & Gelman, 1999). While there are thus interesting future research directions in understanding this computation, the question "specificity or salience" is ill-posed. An adequate theory of rule learning must incorporate both, and the tools of probabilistic modeling are well-suited to capture the tradeoff between them.

## Common-sense psychology and the need for explicit theories

The "common-sense psychology" account given by Endress does not provide a suitable explanation of the rule learning phenomenon, and is not a replacement for more explicit theories. To illustrate this point, I focus here on Endress's account of Gerken's (2006) findings.

Gerken familiarized infants with strings that conformed to the regularity $AAx$, where $x$ represents a single syllable like $/di/$. These same strings were also consistent with the broader regularity $AAB$ (where $B$ represents any syllable), but this rule was more general, being consistent with strings that never appeared during training. At test, Gerken found that infants differentiated new $AAx$ examples from new $AxA$ examples, but failed to differentiate new $AAB$ examples from $ABA$ examples when $B$ was not an $x$ element.

Endress's acount of this phenomenon is as follows:

> Gerken's... experiments can be explained if, in addition to being sensitive to repetitions, humans (and other animals) track items in the edges of sequences... 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 repetition-pattern, because this is the only regularity present in the data. In contrast, when familiarized with $AAdi$ 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.

This explanation feels superficially compelling, but a closer look shows that it presupposes precisely the phenomenon being explained. In particular, it suggests that infants prefer strings that are consistent with the conjunction of "all the patterns they have picked up." But what are "all the patterns they have picked up"? Without an independent specification of this set, there is no explanation. The passage above assumes that infants make two generalizations from the available stimuli, one based on the ending syllable and one based on the consistent repetition. Why these rules and not another one, like "any string that ends in $/di/$, $/je/$, $/li/$, or $/we/$" or "any string with three or four elements"? Deriving predictions from Endress's proposal requires the same assumptions that FT made: about the nature of the space of rules and how they apply to strings. These are precisely the constructs that Endress claims are not present in the mind of

the learner.[2]

To posit an account in which learners discover "all those rules consistent with a stimulus," there must be some story about what the possible rules are. This was exactly the story that FT attempted to give. While our account was almost certainly incomplete, our goal—stated explicitly in the paper—was to create a starting point for future work, even if this required simplifying assumptions that would later be overturned. In the words of George Box, "all models are wrong, but some are useful" (Box & Draper, 1987).

### Probabilistic models, optimality, and fit to data

A common criticism of probabilistic models in cognitive science (Jones & Love, 2011; Bowers & Davis, 2012), taken up in Endress's article, is that they are used to make claims that particular cognitive processes are optimal, but they can be fit to any process or dataset. The combination in turn leads to optimality claims that are unwarranted but unfalsifiable. I will first discuss the relationship between probabilistic models and optimality and then the issue of model flexibility and fit.

[2]To explore Endress's proposal, I implemented it in the FT framework. I held all details of FT's Model 1 constant, but assumed that learners encode the set of rules that are consistent with all of the training strings, and test strings are given zero probability unless they are consistent with these rules. This model ("Model E") does not fit the empirical data as well as FT's Model 1. For example, Model E makes no distinction between musical rules of the forms $low - high - middle$ ($LHM$) and $low - middle - high$ ($LMH$), even though human learners can master only the latter (Endress, Dehaene-Lambertz, & Mehler, 2007; Endress, 2013). (In contrast, and contra Endresss claim, FTs Model 1 does in fact distinguish $LHM$ rules from $LMH$ rules, showing a larger difference in the probabilities of test stimuli under the simple $LMH$ rules than the more difficult $LHM$.) To keep Model E from learning in the tricky $LHM$ condition, we must restrict its hypothesis space so that it *cannot consider* rules that would uniquely specify such strings. But this kind of ad-hoc restriction of the hypothesis space without independent evidence is exactly the error that Endress wants to avoid in the first place.

*Claims of optimality for probabilistic models*

Consider a simple linear regression. A regression model can be fit to any dataset, with whatever predictors the modeler chooses (albeit with better or worse performance in predicting new data). The model's fit can then be compared both with other models within the regression framework— via the addition or subtraction of predictors—or with models of different frameworks—for example, models that do not make an assumption of linearity. In practice though, once the model is fit, it is the rare data analyst who declares that they have discovered a linear process.[3] Instead, the analyst asks what predictors carry most weight, how these predictors interact with one another, what data-points are best or worst fit by the model, and how this model compares to others with more or fewer predictors. This kind of exploratory model-checking and model-comparison approach is standard statistical practice (Gelman & Hill, 2006).

Probabilistic models are no different. Just as an analyst considering a regression model typically examines whether individual predictors should be included, modelers consider design decisions and their impact on overall model fit. The impact of these design decisions can lead to interpretable conclusions, just as a predictor receiving a large coefficient estimate in a regression model can lead to inferences about that predictor's relative importance.

Optimality claims about human behavior enter the picture via two routes. The first is via a conceptual confusion. Probabilistic models define a posterior distribution over hypotheses, which is then typically computed via a range of Bayesian inference methods, from exact computation to sampling methods like Markov chain Monte Carlo (MacKay, 2003). Probabilistic inference methods based on Bayes' rule come with *normative guarantees*: that these inference methods will (in the limit) converge to the correct posterior distribution. These guarantees are useful for the modeler: they mean that, if care is taken in designing the inference procedure, modelers can be

[3]Claims of linearity can certainly be supported by linear modeling (e.g., Shepard & Metzler, 1971), but it would be odd to suggest that this is their primary use!

relatively sure that they have correctly estimated the consequences of their design decisions.[4] These guarantees imply that Bayesian inference is "optimal" in the sense that it leads to the correct posterior distribution. This optimality is a property of the model, however, not of the data being modeled.

The second route to optimality is via the claim that human performance corresponds to the predictions of a model with such normative guarantees.[5] The standards for such a claim of optimality (e.g., a claim at the framework level) are far higher than those for a claim that one model within a framework fits better than another within that framework, and are almost never met. First, such a claim requires evidence that other modeling frameworks do not provide similarly powerful explanations. Second, an optimality claim requires rarely-given qualifications about the level of optimality that is assumed, and whether behavior is optimal in individual judgments or only when aggregated over multiple observations or even individuals.

For these reasons and others, FT did not make a claim of optimality.[6] We framed our models in terms of an alternative tradition from perception: the *ideal observer* tradition. In contrast to the probabilistic modeling tradition, where issues about optimality have had a complex history (Anderson, 1990; Oaksford & Chater, 1994), the ideal observer tradition has been more explicit about the use of models with normative guarantees to model non-normative human

[4]See Perfors, Tenenbaum, Griffiths, and Xu (2011) for further explanation of this topic and Goldwater, Griffiths, and Johnson (2009) for an example in which improper probabilistic inference led to a problematic interpretation.

[5]This claim is bound up in the tradition of *rational analysis*, which codified the idea of considering cognition as adapted to its situation (for an introduction to these ideas and their genealogy in functionalism and ecological validity, see Anderson, 1990). This tradition raises many rich (and problematic) issues, but a full discussion of rational analysis is beyond the scope of this manuscript.

[6]Indeed, the evidence suggests that human performance in sequence learning is far from conventional standards of optimality. To take examples from my own work, models of word segmentation performance provide extremely poor fit to human performance in segmentation tasks unless they are "handicapped" by the addition of severe memory constraints (Frank, Goldwater, Griffiths, & Tenenbaum, 2010), and human learners appear to make suboptimal use of contextual information in these tasks (Kurumada, Meylan, & Frank, in press).

behavior (Geisler, 2003). Such tools have been used both to provide evidence that early perceptual behavior makes optimal use of the available information in some domains (e.g., in light wavelength discrimination; Geisler, 1989) and that it is suboptimal in others (e.g., in contrast sensitivity; Banks, Sekuler, & Anderson, 1991).

*Free parameters, flexibility, and fit to data*

A model is fit to data when its free parameters are set so as to maximize some objective function. In the case of regression, this would be the step of estimating coefficient weights by minimizing the sum of squared prediction errors. In a probabilistic model, this might involve searching for the parameter setting that maximized the posterior probability of the data. While the quality of a fit can be captured using goodness-of-fit statistics like $r^2$, these measures do not account for the number of free parameters that were needed to achieve this fit (Roberts & Pashler, 2000; Hastie, Tibshirani, Friedman, & Franklin, 2005; Pitt & Myung, 2002). An individual model is *overfit* when its flexibility allows it to be tailored to idiosyncrasies of the current dataset, resulting in poor performance in generalizing to other datasets.

Endress criticized FT on the grounds that several of our models had free parameters that were fit to the data. Indeed, why fit cognitive models to the data at all? Although there is a large body of data on rule learning, experiments vary widely in the type of stimuli they use, the amount of exposure they give to learners, and the age of the learners, among other things. A model making quantitative predictions about the behavior of adults learning from three examples (Endress et al., 2007) must be different in some respect from a model making predictions about the behavior of seven-month-olds learning from dozens of examples. Unless the modeler has a complete theory of how adults and babies differ from one another, some ad-hoc modification is needed to distinguish the two. Unfortunately, sufficient data were not available to allow validation of our decisions on a held-out dataset.

To avoid overfitting in FT's models, therefore, we allowed ourselves a very small set of free parameters: none in Model 1, only one in Model 2, and two in Model 3 (even though this decision

lumped together distinct psychological constructs like noise in perception and noise in memory). These free parameters—in particular, the noise parameter introduced in Model 2—allowed us to compare datasets across widely varying populations and tasks.

Endress also questions the legitimacy of using different memory parameters for training and test items. When necessary, we distinguished the larger memory demands involved in maintaining a representation of training items across a long exposure period compared with an individual evaluating test items in the moment. These decisions gave us some flexibility in our predictions, but there are nevertheless infinitely many patterns of data that our models could not fit.[7]

Finally, the ability to fit a particular model to a dataset should be distinguished from the ability to construct a model that provides a good fit to a dataset. The charge that probabilistic models can fit any dataset is the second type of claim, not the first. It is not a claim of overfitting—it is a claim of framework flexibility. Flexibility in a modeling framework is a much looser, more intuitive notion and suggests merely that a model with good fit to a dataset could potentially be constructed within some framework. In contrast to the flexibility of an individual model—which promotes overfitting—the framework flexibility is a virtue: Linear regression is such an effective and widespread tool because it can be used effectively across a plethora of datasets.

To summarize: FT did not make an optimality claim. Absent this claim, the flexibility of probabilistic models is an important feature that allows them to be used to explore a wide variety of cognitive domains. Nevertheless, more data—especially from experiments that keep participant groups and paradigms constant across many rule types—are necessary to advance the study of rule learning and allow for more elaborated models.

[7]To take one example: If learners in Gerken's (2006) experiments had succeeded in distinguishing $AAB$ examples from only $AAx$ training but not $AAB$ training, no manipulation of our noise parameter would have produced this pattern of results.

## Conclusion

Endress's article raises important issues about the relationship between computational-level and algorithmic-level descriptions, and adds to critiques of optimality claims for Bayesian models. Nevertheless, the substantive contentions—that learners do not show a preference for more specific rules, and that psychological principles explain the extant body of results in rule learning—are not supported by the evidence.

More generally, Endress imputes that the goal of probabilistic modeling is to show that babies (or other learners) are Bayesian and hence optimal. Probabilistic models on this view are opposed to basic psychological principles such as salience or memory. On the contrary, I have argued here that probabilistic approaches—along with connectionist and other formal approaches to the cognitive modeling—are a tool for theorizing, for moving from "common sense" to formal theories that make quantitative predictions from well-understood and explicit assumptions. Ideal observer models posed at Marr's computational level, as ours were, represent one tool for such theorizing, while models at the algorithmic and implementational levels represent others. Reverse engineering the mind will require all the tools at our disposal.

## References


Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Banks, M., Sekuler, A., & Anderson, S. (1991). Peripheral spatial vision: Limits imposed by optics, photoreceptors, and receptor pooling. *JOSA A*, *8*, 1775–1787.

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

Box, G., & Draper, N. (1987). *Empirical model-building and response surfaces*. Oxford, UK: Wiley and Sons.

Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*, 287–291.

Chomsky, N. (1995). *The minimalist program* (Vol. 28). Cambridge Univ Press.

Endress, A. (2013). Bayesian learning and the psychology of rule induction. *Cognition*.

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

Endress, A., Scholl, B., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*, *134*, 406–419.

Frank, M. C., & Gibson, E. (2011). Overcoming memory limitations in rule learning. *Language, Learning, and Development*, *7*, 130–148.

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

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

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

Geisler, W. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*, 267.

Geisler, W. (2003). Ideal observer analysis. In *The visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Gerken, L. A. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*, 67–74.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21-54.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*, 83–85.

Hopcroft, J., Motwani, R., & Ullman, J. (1979). *Introduction to automata theory, languages, and computation.* Reading, MA: Addison-Wesley.

Johnson, S., Fernandes, K., Frank, M., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. (2009). Abstract rule learning for visual sequences in 8-and 11-month-olds. *Infancy*, *14*, 2–18.

Jones, M., & Love, B. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–188.

Kurumada, C., Meylan, S., & Frank, M. C. (in press). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge, UK: Cambridge University Press.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science*, *18*, 387–391.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: Henry Holt and Company.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *5*, 115–133.

Navarro, D., & Perfors, A. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, *133*, 256–268.

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

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

Perfors, A., Tenenbaum, J., Griffiths, T., & Xu, F. (2011). A tutorial introduction to bayesian models of cognitive development. *Cognition*, *120*, 302–321.

Pitt, M., & Myung, I. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 701-703.

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

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*, 223–250.

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

Xu, F., & Tenenbaum, J. (2007b). Word Learning as Bayesian Inference. *Psychological Review*, *114*, 245–272.