Cognition 128 (2013) 417–423

Contents lists available at SciVerse ScienceDirect

Cognition

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

Brief article

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

Michael C. Frank*

Department of Psychology, Stanford University, United States

ARTICLE INFO

Article history:
Received 16 January 2013
Revised 29 April 2013
Accepted 30 April 2013

Keywords:
Rule learning
Bayesian models
Cognitive modeling

ABSTRACT

A recent probabilistic model unified findings on sequential generalization ("rule learning") via independently-motivated principles of generalization (Frank and Tenenbaum, 2011). Endress 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 useful points about the importance of mechanistic explanation, but the specific critiques of our work are not supported. More broadly, I argue that Endress undervalues the importance of formal models. Although probabilistic models must meet a high standard to be used as evidence for optimality claims, they nevertheless provide a powerful framework for describing cognition.

1. 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 et al., 1979). But since computational systems can store their state in processes as diverse as symbols on a tape or weights between neurons (McCulloch and 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 (Anderson, 1990; Chomsky, 1995; 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 et al., 2006; Tenenbaum et al., 2011).[1]

Although probabilistic models have grown tremendously in popularity in recent years, they have also attracted significant criticism (Bowers and Davis, 2012; Jones and Love, 2011). Chief among these criticisms is that these models imply a claim that the mind itself is *rational* or even

* Tel.: +1 650 724 4003.
E-mail address: mcfrank@stanford.edu

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2013.04.010

[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.

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 (2011) models of sequential rule learning (henceforth, "FT") and providing additional theoretical and empirical arguments.

2. "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 et al., 2007; Marcus et al., 1999; Peña et al., 2002). 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 (Endress et al., 2005; Gerken, 2006; Johnson et al., 2009; Marcus et al., 2007; Marcus et al., 1999).

FT created 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 and 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, allowing the model to produce quantitative predictions. Model 3 learned multiple rules. Despite the apparent diversity of results in the literature, these simple models sufficed to describe a wide range of empirical data. Our explicit goal in FT was to provide "a baseline for future work that can be modified and enriched" as the data warranted.

Endress (2013) argues that our models do not provide a good account of existing data on rule learning, however, contesting both the general framework we used and the specifics of our simulations. In this response I will focus primarily on a set of critiques that have broad interest:

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 free parameters is inappropriate in cognitive modeling.
4. The use of probabilistic models implies an optimality claim.

In Appendix A, I briefly summarize responses to criticisms of specific simulations.

To summarize, I argue that Endress's criticisms 1–3 are not valid. Moving beyond a notion of "common sense" psychology to theories that make graded and quantitative predictions, we will need to use statistical tools to understand and evaluate the flexibility and specificity of our theories. Nevertheless, Endress's article raises useful questions about how computational principles can be instantiated in human minds and I am in agreement that there should be a high standard for claims of optimality on the basis of probabilistic modeling (indeed, FT did not make such a claim).[2]

3. 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 and Griffiths, 2001; Xu and Tenenbaum, 2007b)—the principle that hypotheses are weighted proportional to their specificity, as a consequence of how those examples are sampled—was the major explanatory assumption in FT's models.

A large, independent body of evidence supports the use of the size principle as a description of word learning and categorization (Navarro et al., 2012; Tenenbaum and Griffiths, 2001; Xu and Tenenbaum, 2007a, 2007b) and the sensitivity of even young infants to the sampling processes that result in the size principle (Denison et al., 2012; Gweon et al., 2010; Kushnir et al., 2010; Xu and Garcia, 2008; Xu and Denison, 2009). To take just one 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. This dataset and many others provide powerful evidence for the importance of strong sampling and 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

[2] This response represents my personal views. There is substantial variance in attitudes towards optimality claims in the probabilistic modeling literature, and the attitudes of many researchers have evolved with respect to this issue as more research has focused on "process level" explanations for cognitive phenomena (Chater et al., 2011; Griffiths et al., 2012; Sanborn et al., 2010; Vul et al., 2009).

vocalizations). Participants largely preferred test trials consistent with the modality of presentation.

These data show that 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 here.

Endress's data thus 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 are computed during brief experiments. Endress rejects as implausible 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 and Spelke, 2000; Whalen et al., 1999). While there are 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 provide a powerful method for capturing the tradeoff between them.

4. 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. Computational models provide a method for making theoretical assumptions explicit, and avoiding issues of vagueness and circularity. 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, the "common-sense" account 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"?

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, and deriving predictions from Endress's proposal requires precisely the same assumptions that FT's models made explicit: the nature of the hypothesis space of rules and how rules apply to individual strings. In addition, while our models were almost certainly incomplete, they had a virtue that any "common-sense" account necessarily lacks: the ability to make quantitative predictions. If psychologists are limited to "common-sense" theorizing, we will be unable to move beyond crude binary hypotheses to graded predictions about human behavior. In the words of George Box, "all models are wrong, but some are useful" (Box and Draper, 1987).

5. Probabilistic models, optimality, and fit to data

A common criticism of probabilistic models in cognitive science (Bowers and Davis, 2012; Jones and Love, 2011), 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.

5.1. 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 datapoints 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 and 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 we interpret the coefficients of predictors in a regression model.

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 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 are almost never met. First, such a claim requires evidence that other modeling frameworks cannot fit the data without making the same assumptions as the normative model. This type of framework-level evidence is almost impossible to provide. Second, an optimality claim requires rarely-given qualifications about (A) whether behavior is claimed to be optimal for individuals or at the population level and (B) whether it is optimal in single judgments or in the long-run average.

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 and Chater, 1994), the ideal observer tradition has been more explicit about the use of models with normative guarantees to model non-normative human behavior (Geisler, 2003). Such tools have been used both to provide evidence that early perceptual behavior makes effective use of the available information in some domains (e.g., in light wavelength discrimination; Geisler, 1989) and that it is clearly suboptimal in others (e.g., in contrast sensitivity; Banks et al., 1991).

5.2. 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 maximizes 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 (Hastie et al., 2005; Pitt and Myung, 2002; Roberts and Pashler, 2000). 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. Unless the modeler has a complete theory of, for example, how memory for sequentially-presented syllables (Gerken, 2006) differs from their memory for simultaneously-presented dog photos (Saffran et al., 2007), a free parameter is needed to distinguish the two.

To avoid overfitting in FT's models, 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, stimuli, and tasks. In fact, from a statistical point of view, the issue with FT's

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

[4] See Perfors et al. (2011) for further explanation of this topic and Goldwater et al. (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 et al., 2010), and human learners appear to make suboptimal use of contextual information in these tasks (Kurumada et al., 2013).

Table A.1
Responses to Endress's critiques of individual findings.

Experiment	Specific critique	Response
Marcus et al. (1999)	Finding depends on size principle	Independent evidence for size principle (see Section 3)
Endress et al. (2007)	Incorrect predictions about rising/falling melodies	Contra Endress, FT models discriminate rising/falling melodies better than LHM melodies (see Appendix)
Frank et al. (2009)	Computing rule cardinalities is psychologically implausible	Rule sizes can be estimated (see Section 3)
Gerken (2006)	Finding depends on size principle as well as details of hypothesis space	See Sections 3 and 4
Gerken (2010)	One token may not always be a strong counterexample	Issues in memory for types vs. tokens (see Appendix)
Marcus et al. (2007)	Overfitting of the memory parameter	See Section 5.2
Saffran et al. (2007)	Overfitting of the memory parameter	See Section 5.2
Gómez (2002)	Problems with use of memory parameter	Issues in memory for types vs. tokens (see Appendix)
Kovács and Mehler (2009)	Finding can be interpreted in terms of bilingualism and executive control	Psychological accounts are not opposed to model-based accounts (see Sections 4 and 6)

models was not the number of free parameters (which was small from any perspective) but instead the amount of data. One of the goals of our work was to illustrate the necessity of collecting quantitative data on rule learning so that more detailed models could be constructed.

Endress questions the use of free parameters to explore the effects of factors like modality differences or training/test asymmetries, but this type of exploration provided us with a method for understanding the degree to which simple issues of exposure or stimulus familiarity could have driven rule learning effects. For example, 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. This type of flexibility allowed us to investigate the role played by memory (and in one simulation we reported results both with and without memory noise at test). But this flexibility should not be misconstrued: it did not allow our model to fit any pattern of data[7] and it was in no way inconsistent with an ideal observer approach. On the contrary, investigating the dependence of predictions on assumptions about perceptual and memory noise is precisely the purpose of ideal observers (Geisler, 2003).

Finally, the possibility of *fitting* a particular model to a dataset should be distinguished from the possibility of *constructing* a model that provides a good fit to a dataset. Endress's charge is not a claim of overfitting: It is a claim of framework flexibility. But flexibility in a modeling framework is a good thing. Linear regression and probabilistic models are both effective and widespread tools precisely because they can be applied to a plethora of datasets. Being able to construct a model that fits a dataset is only a problem if the mere act of constructing such a model then somehow becomes evidence for an optimality claim.

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.

6. Conclusion

Endress's article raises important issues about the relationship between computational-level and algorithmic-level descriptions, but his substantive critiques of our work are not supported. 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" intuitions 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.

Appendix A. Specific critiques of simulations

In this section, I briefly summarize responses to specific critiques (Table A.1). I have referenced sections above whenever appropriate. Since all of FT's findings depend on the size principle and all of Endress's critiques require it, I have not repeated this point, but it applies to all of the experiments listed.

Endress et al. (2007). Endress writes that FT's models make the following prediction: Human learners should be no better at learning a *LMH* (low-middle-high, or rising) rule and discriminating consistent strings from *HML* (falling) strings than they are at learning *LHM* rules and discriminating consistent strings from *MHL* strings. Simulations from our Model 1 show that Endress's deriva-

[7] There are of course infinitely many patterns of data that our models could not fit. To take one important example from the perspective of the size principle: If learners in Gerken's (2006) experiments had succeeded in distinguishing *AAB* examples from only *AAx* training but not *AAB* training, there is no manipulation of our noise parameter that would have produced this pattern of results.

tion of this prediction is not correct. Rising/falling contours are consistent with the same number of rules as *LHM* rules in our model, but incorrect test items (e.g., a falling string after rising training) violate nearly all of these rules and lead to a difference in surprisal similar to that caused by violations of identity rules (5.44 for correct items, 15.78 for incorrect items, compare with Table 3 of FT). Thus, rising/falling contours are predicted to group with identity rules rather than *LHM*-type rules. This prediction is confirmed by Endress's Experiment 3.

Gerken (2010) and Gómez (2002). In his treatment of these two findings, Endress raises an issue that FT also noted: The memory noise parameter used in the simulations was assumed to operate over unique string types rather than individual tokens or some combination of the two. Of course, given that more tokens of an individual string type likely leads to it being remembered better, the assumption to model memory over tokens alone is a major simplification. Endress's examples (e.g., 9999/10,000 tokens being consistent with a rule) highlight this simplification and points to the necessity for better models of type/token generalization and the data to test them (Goldwater et al., 2006).

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. *Journal of the Optical Society of America 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., Goodman, N., Griffiths, T., Kemp, C., Oaksford, M., & Tenenbaum, J. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences, 34*, 194–196.

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*. Cambridge University Press.

Denison, S., Reed, C., & Xu, F. (2012). The emergence of probabilistic reasoning in very young infants: Evidence from 4.5- and 6-month-olds. *Developmental Psychology*, *49*, 243–249.

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

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., 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., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps five-month-olds learn abstract rules. *Developmental Science, 12*, 504.

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.

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115*, 362–366.

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.). *Advances in neural information processing systems* (Vol. 18, pp. 459–466). Cambridge, MA: MIT Press.

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

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431–436.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*, 263–268.

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*, 107.

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., et al. (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.

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

Kurumada, C., Meylan, S., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition, 127*, 439–453.

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

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

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.

Saffran, J., Pollak, S., Seibel, R., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition, 105*, 669–680.

Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 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.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (2009). One and done? Optimal decisions from very few samples. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 148–153).

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

Xu, F., & Spelke, E. S. ((2000). Large number discrimination in 6-month-old infants. *Cognition, 74*, B1–B11.

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

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

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.