Developmental Science

Developmental Science (2014), pp 1–16 DOI: 10.1111/desc.12083

PAPER

Rich analysis and rational models: inferring individual behavior from infant looking data

Steven T. Piantadosi, Celeste Kidd and Richard Aslin

Department of Brain and Cognitive Sciences, University of Rochester, USA

Abstract

Studies of infant looking times over the past 50 years have provided profound insights about cognitive development, but their dependent measures and analytic techniques are quite limited. In the context of infants' attention to discrete sequential events, we show how a Bayesian data analysis approach can be combined with a rational cognitive model to create a rich data analysis framework for infant looking times. We formalize (i) a statistical learning model, (ii) a parametric linking between the learning model's beliefs and infants' looking behavior, and (iii) a data analysis approach and model that infers parameters of the cognitive model and linking function for groups and individuals. Using this approach, we show that recent findings from Kidd, Piantadosi and Aslin (2012) of a U-shaped relationship between look-away probability and stimulus complexity even holds within infants and is not *due to averaging subjects with different types of behavior. Our results indicate that individual infants prefer stimuli of intermediate complexity, reserving attention for events that are moderately predictable given their probabilistic expectations about the world.*

Introduction

The 'blooming buzzing confusion' (James, 1890) of early childhood provides a substantial challenge to young learners. Not only must infants learn much about the structure and properties of the world, but before such learning can begin, infants must first attend to the right subset of experience – they must discover which information sources are useful. In particular, children must ignore both environmental noise, and stimuli from which they have nothing more to learn. Kidd *et al*. (2012; henceforth KPA) suggested that infants solve this problem by building implicit statistical models of observed stimuli, and directing attention according to the information-theoretic properties of these cognitive representations. Their work showed that infants stop attending to a stimulus that is either too predictable, or too *un*predictable, according to an idealized model of infants' implicit beliefs. Predictable stimuli leave very little left to learn, and unpredictable stimuli are potentially either noise or the result of incomprehensible complexity. In either case, cognitive effort is best spent elsewhere.

KPA showed this by presenting infants with controlled sequences of discrete events – objects appearing from behind boxes much like the game Whac-a-mole except that infants did not interact with the objects. Figure 1(a) shows one of their displays, in which three objects were each occluded behind a box.[1] In the experiment, a single object would appear from behind each box (an 'event') in some sequential order. If the three events are denoted $\{A, B, C\}$, an example event sequence would be $ABACAABAABA\ldots$, corresponding to box A's object appearing from behind box A, then box B's object appearing from behind B, then box A's object again, etc. Overall, there were 32 event sequences each 30 elements long, which were fixed across infants, but presented in random orders. This design allowed for collection of a substantial quantity of data on infants' probability of looking away at each event in each sequence. The sequences were designed by KPA to vary in their predictability and information-theoretic properties.

[1] Box locations and the objects pairings were randomized across trials.

Address for correspondence: Steven T. Piantadosi, Brain & Cognitive Sciences, Meliora Hall, RC Box 270268, University of Rochester, Rochester, NY 14627-0268, USA; e-mail: spiantadosi@bcs.rochester.edu

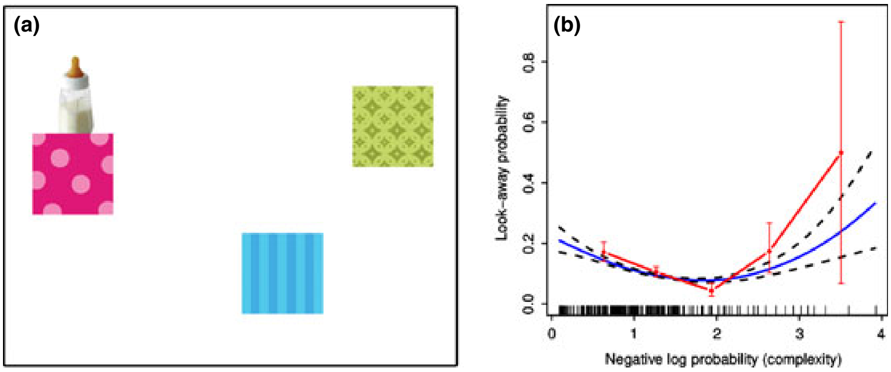

Figure 1 *1(a): An example display from KPA showing three boxes and an object (a bottle) appearing from behind one. 1(b): KPA's U-shaped relationship between look-away probability (y-axis), and log probability (x-axis), with raw binned data (red line) and a Generalized Additive Model (Hastie & Tibshirani 1990), both collapsing over subjects. [Reprinted with permission from Kidd, Piantadosi, & Aslin, (2012), distributed under the Creative Commons Attribution License].*

Building off studies demonstrating that infants have rich capabilities for statistical learning and inference (Saffran, Aslin & Newport, 1996; Saffran, Newport & Aslin, 1996; Saffran, Johnson, Aslin & Newport, 1999; Fiser & Aslin, 2002; Xu & Garcia, 2008; Xu & Denison, 2009; Dewar & Xu, 2010; Téglás, Vul, Girotto, Gonzalez, Tenebaum & Bonatti, 2011; L. Smith & Yu, 2008), KPA constructed an ideal observer model of these stimuli that quantified the degree to which infants *should* expect any particular event to appear next in the sequence. For instance, the sequence *ABACAABAABA* would lead infants to expect *A* as a relatively likely next event. Conversely, *C* is relatively unlikely since it has not appeared frequently in the past.

Infants watched these displays on a Tobii eye tracker, and the critical dependent measure was the *specific event* in each sequence when infants terminated their attention to the displays and directed attention away from the screen (e.g. to the room, their feet, their parent, etc.). KPA showed that infants' look-aways during these sequences were influenced by the predictability of each outcome according to their model of the previous events. Infants were significantly more likely to look away on events that were either *highly surprising* or *highly non-surprising* according to the idealized model. 'Surprise' was measured according to the *negative log probability* (see Shannon, 1948) of an event according to the statistical model. Negative log probability can be viewed as a measure of complexity on the scale of *bits of information*, corresponding to how many bits would be required for an ideal learner to store or process the current event. Figure 1(b) shows KPA's primary result: probability of looking away (*y*-axis) has a U-shaped relationship with the negative log probability of an event (*x*-axis). This plot shows binned raw data, and a smooth line for a Generalized Additive Model (Hastie & Tibshirani, 1990) with binomial link function, although KPA's primary data analysis tested significance using a kind of regression, a survival analysis, more suited to the experimental framework (Hosmer, Lemeshow & May, 2008; Klein & Moeschberger, 2003). In general, these results are suggestive that infant attentional mechanisms form an efficient gateway to other learning mechanisms, filtering stimuli using infants' own predictive expectations to primarily attend to learnable stimuli (Gerken, Balcomb & Minton, 2011). KPA's work therefore provides a formal quantification (see also Civan, Teller & Palmer, 2005; Kaldy, Blaser & Leslie, 2006) and test of the long informally-hypothesized theory that infants prefer intermediate stimulus complexity (Dember & Earl, 1957; Hunter & Ames, 1988; Kinney & Kagan, 1976; Roder, Bushnell & Sasseville, 2000; Rose, Gottfried, Melloy-Carminar & Bridger, 1982; Sokolov, 1963; Wagner & Sakovits, 1986).

KPA's model is rational in that the statistical model infants are hypothesized to create represents a good solution to the problem of determining the true distribution of experimental stimuli they observed (see Anderson & Schooler, 1991; Chater & Oaksford, 1999; Geisler, 2003). Further, KPA's model is a kind of 'pure' rational model which did not attempt to capture any limitations of learners such as memory decay, or individual variability. Such models are useful in part because both agreement with and deviations from the predictions of these models are substantially informative. Agreement reveals congruence with the expectations of an idealized view of learning; deviations likely reveal nontrivial cognitive limitations. However, downsides of 'pure' rational models are also clear: it is almost certain that infants do not have perfect memory for previous events, and that individuals may vary in their response patterns. Greater statistical power and accuracy is gained by modeling these components. More importantly, as we discuss in detail later, the U-shape KPA observe *could* be due to collapsing across two types of infants (for an example, see McMurray & Aslin, 2005) – some who prefer low complexity and some who prefer high. This would substantially change the interpretation of KPA's results.

Our primary goal in this paper is to develop richer methods for rational modelling in infant cognition – methods that can capture effects such as memory decay and individual subject variation and formalize an explicit linking function between infants' beliefs and behavior (Aslin, 2007; Yurovsky, Hidaka & Wu, 2012). We aim to demonstrate how rich modeling of data can be combined with formalized cognitive theories, to the benefit of both. Our analysis incorporates both an idealized statistical learning model posited to exist in infant cognition and a behavioral model of the responses collected in an experiment. The behavioral model uses *as input* the state of the ideal observer model: infants' actions at each time depend on their implicit beliefs about the world. The ideal observer model, in turn, uses as input the observed experimental stimuli. Both of these statistical models are formalized as Bayesian models (for tutorials, see Chater, Tenenbaum & Yuille, 2006; Griffiths & Yuille, 2008; Perfors, Tenenbaum, Griffiths & Xu, 2011). The power of this approach is that by combining cognitive modeling with sophisticated data analysis, we are able to make strong inferences about the components of infants' learning systems and distinguish a theoretically important range of possible hypotheses (see also Yurovsky *et al*., 2012; Piantadosi, Tenenbaum & Goodman, 2009; Piantadosi, 2011). For instance, this method is capable of discovering the (prior) assumptions of infants' own inferential models: KPA's model used a prior parameter, α, which controlled the degree to which learners expect previously observed events to be repeated. While KPA assumed a fixed value of $\alpha = 1$ in their rational analysis – corresponding to largely unbiased learners – it is much more interesting to determine what values of α 'best' explain infants' behavior. Do infants expect that previously observed events are more likely to re-occur ($\alpha \ll 1$) or do infants expect that all events are always equally likely to be seen ($\alpha \gg 1$)? Alternatively, infants might possess an even stronger form of unbiased rationality, perhaps bringing prior expectations to learning that are as unbiasing as possible ($\alpha = 0.5$) with respect to the probabilistic setup (Jeffreys, 1961). These are not the types of questions that are traditionally considered in the domain of infant cognition since they require both formalized models and sophisticated ways to relate models to behavior.

The approach we use is centered around Bayesian data analysis (Gelman, Carlin, Stern & Rubin, 2004), which has become increasingly popular in psychology (Kruschke, 2010a, 2010b). The advantages of Bayesian methods for data analysis are numerous (see Jaynes, 2003; Gelman *et al*., 2004; Kruschke, 2010a, 2010b) and include, most notably, an advantageous conceptualization of scientific inference.[2] Bayesian techniques have been applied most notably in the developmental literature by Lee and Sarnecka (Lee & Sarnecka, 2010a, 2010b; Sarnecka & Lee, 2009), who provide a Bayesian data analysis method for testing 'knower-levels' in number knowledge, while controlling for other factors like subject effects and an individual baseline distribution of responses. Their work formalized a much richer hypothesis space than previous studies of these developmental stages, and allowed for powerful inferences and theory comparisons. In the infant literature, Yurovsky *et al*. (2012) present a novel Bayesian analysis of looking-time data, as well as simulations showing that their approach can recover 'true' parameters with very high accuracy.

For our purposes, the important feature of Bayesian data analysis is its power: we are able to specify an arbitrarily complex probabilistic model of behavior, and make inferences about the parameters of that model using observed behavioral data. The outline of the present paper is as follows: the next section describes KPA's experiment and the Dirichlet-Multinomial (DM) models used by KPA. We describe modifications of KPA's model, including memory decay, and formalize a space of linking functions between the properties of the DM and look-away behavior. We then describe the Bayesian data analysis model, and present results showing that the U-shaped relationship reported by KPA is observed even in individual infants. Each step in the formalization of this model requires making assumptions about the mathematical form of the model and its statistical form. In general, we make these choices to conform to prior literature on this type of modeling. As many of the choices made in prior literature are motivated by conceptual or mathematical simplicity, these properties carry forward to our model. Our presentation focuses on providing an intuitive understanding of the techniques, with the hope that this work illustrates how cognitive modeling can be combined with state-of-the-art data analysis to produce fully probabilistic models that are capable of distinguishing critically different hypotheses.

[2] For discussion of Bayesian epistemology of science, see Godfrey-Smith (2003).

A rational model: Whac-a-mole

It is useful to first consider the behavior of an individual subject in KPA's task. In the general framework of a DM model used by KPA, the learner observes each event (A, B or C) some number of times, and uses this to infer an overall distribution of events.[3] For instance, in the sequence *ABABCCAA*, the learner will have observed four As, two Bs, and two *C*s, meaning that they should expect that it is more likely the next event is an A than a B or a C. The DM model allows one to take this set of counts and compute the expected probability that each event will occur in the future.[4] As mentioned above, the DM formalization we use has a single free parameter, α, that controls the degree to which the model expects a uniform distribution on events (large α, making all events equally likely), versus expecting future events to be exactly like previously observed ones (small α, primarily the data drives expectations). A simple way to interpret α is to imagine that α is the number of counts which are *assumed* by the learner to already have been observed on each event. For instance, if $\alpha = 1$, then the model acts *as though* it had observed A, B, and C each $\alpha = 1$ times more than were actually observed, exactly as in 'add one' smoothing (Chen & Goodman, 1999). A convenient feature of a DM model is that the expected probability of each event can be found by dividing the number of counts (observed and assumed) for an event, by the total number of counts. Thus, if we had observed a sequence like *ABBBABAA* and $\alpha = 1$, the expected probability of observing A next would be $4 + 1 = 5$, the observed plus assumed counts of A, divided by the total counts, $9 + 3$ (9 observed events plus $3\alpha = 3$ assumed counts, corresponding to α assumed counts for each of three event types), giving a probability of $5/12 \approx 0.42$. Alternatively, if we increase α, the distribution is pulled towards the uniform probability of ⅓. So if $\alpha = 10$, then the probability of A is $(4 + 10)/(9 + 3{\cdot}10) \approx 0.36$. Conversely, decreasing α moves the distribution towards the empirically observed distribution of 4/9 *A*s: when $\alpha = 0.1$, the probability of A is $(4 + 0.1)/(9 + 3{\cdot}0.1) \approx 0.44$.

While the DM setup looks simple, it is actually quite powerful. In the above example, even though C has not been observed, the parameter α performs a 'smoothing' role and assigns C a nonzero probability of $(0 + 1)/12$. As the number of data points gets large, the choice of α matters less and less: in the limit of data, the expected distribution is the empirically observed one, a desirable feature of a good learner. Also, the *degree* of strength of belief in uniformity depends on the magnitude of α. As noted earlier, if $\alpha = 10$, then all probabilities are pulled closer to 1/3, the uniform distribution and when $\alpha \ll 1$, the distribution tends strongly towards the observed counts. A priori, we may not know what a plausible value for an individual's α is, although reasonable choices exist such as $\alpha = 1$, corresponding to flat (unbiased) expectations about the distribution of events. Much better, though, would be to infer α from infants' behavior, and thus discover what assumptions *their* statistical model of the world makes.

In its standard form, the DM assumes that the events are independent and distributed according to some unknown multinomial distribution. But in the cognitive realization of such a model, it is likely that not all events are treated equally. In particular, learners likely have better memory for more recent events, a finding dating back to Ebbinghaus (1913).[5] As in adult work on rational statistical modeling with memory decays (Goodman, Mansinghka & Tenenbaum, 2007; Piantadosi, 2011; Piantadosi *et al*., 2009), we assume a *power law* decay in memory (Anderson & Schooler, 1991). Here, an event i items back is given an *effective* count of $(i + 1)^{-\lambda}$. This is less than the count of 1 that it would receive with no memory decay. For instance, if $\lambda = 1$, the previous event ($i = 1$) will have an effective count of ½, the event before will have an effective count of ⅓, etc. One way to visualize the influence of these parameters it to consider a simple example, such as the model's inferences after observing the sequence *AB*. Figure 2 shows the DM model's predicted distribution of events after observing *AB*, for various values of α and λ. In the sequence *AB*, both A and B have occurred once, but A occurred one item further back in the sequence, meaning that the difference between the probability of A and the probability of B represents effects of memory decay, parameterized by λ. In this '*AB*' sequence C has not been observed, so comparing the probability of C to A and B shows how much smoothing is provided by α, relative to the memory-decayed counts of A and B. Thus, when α is

[3] We note that KPA also tested a *transitional* model which learned dependencies between transitions of events, such as A being likely to follow B. This model showed stronger patterns than the non-transitional model which assumed independence between events, but is more complex. For simplicity, we focus here on the non-transitional model and leave transitional modeling for future work.

[4] Much neater, the DM model actually allows one to compute the probability that any hypothesized *distribution* on events is the correct generating one. For simplicity, here we discuss only the expectation of this full distribution, given some set of counts.

[5] It is also in principle possible to include a primacy bias, but this substantially complicates the model and our sense in watching the stimuli is that the early events are easily forgotten.

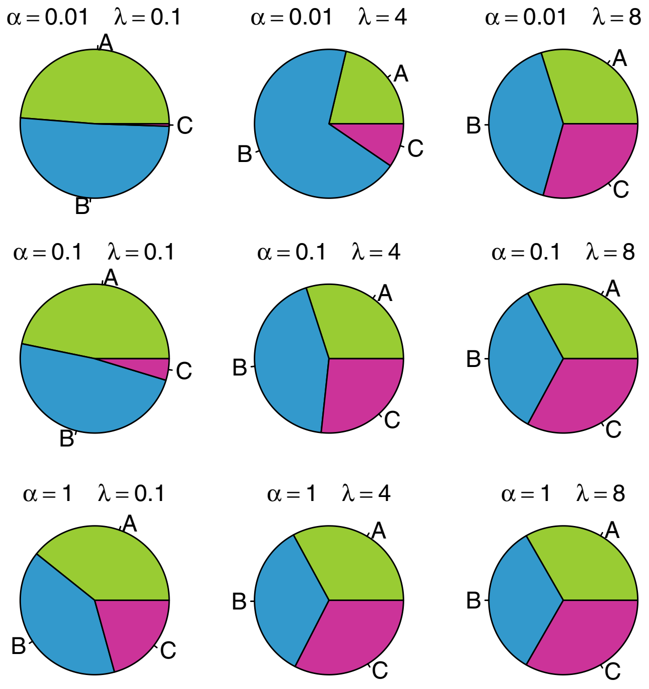

Figure 2 *Pie charts illustrating the relative influence of α and λ on the DM model's inferences. Each pie represents the learner's expected distribution on events after observing the sequence* AB. *The center plot corresponds to α = 0.1, λ = 4, the approximate values found for infants in the experiment (see results). This shows a moderate amount of decay and substantial smoothing.*

small (top row of Figure 2), the model assigns the unobserved event C very little probability unless the memory decay is very strong (e.g. $\lambda = 8$). When the memory decay is so strong, the model essentially 'forgets' all the data and expects all events to be equally likely. Similarly, when α is large (relative to the decay provided by λ), as on the bottom row, the smoothing term dominates the model's inferences, pushing predictions towards equal probabilities for the three events.

Linking the DM model to behaviour

The key finding from KPA is that infants' probability of looking away has a U-shaped (quadratic) relationship with the information provided by each event, given the previously observed events in the sequence. Following Shannon (1948), KPA measured information as the *negative log probability* of each event according to the infants' current model of the world. So for instance, if the event A was predicted to occur with probability ⅓ by the DM, this event would convey $-\log_2 ⅓ = 1.6$ *bits* of information. This is a measure of the amount of information processing an idealized infant would have to do in order to access (i.e. encode or remember) the event. KPA found that this information measure[6] has a quadratic relationship with infants' look-away probability, meaning that infants were significantly more likely to look away from an event sequence when the current event was either very high or very low in information.

As in KPA, we are interested in testing for the U-shape and so we must potentially allow for a broader range of relationships. Unlike KPA, our model and analysis focuses on the behavior of *individual* infants, rather than group-means. We assume that each individual infant's probability of looking away at each item in the event sequence is a parameterized function of the negative log probability (complexity) of each event:

$$P(\text{look-away at event i}) = \exp(\beta_0 + \beta_2(x - \beta_1) + \beta_3(x - \beta_1)^2), \qquad (1)$$

where x is the information (negative log probability) of event i, and β_0, β_1, β_2, and β_3 are parameters that characterize the relationship between information and look-away.[7] Here 'exp' is the exponential function $\exp(y) = e^y$. The information value, x, in (1) is what relates this equation to the statistical learning model: the DM model specifies the x value (information content) of each observed event, conditioned on all the previously-observed data. Equation (1) converts this idealized measure of information to a measure of *behavior*, giving infants' look-away probability as a function of the 'surprisingness' of the data. This equation says that infants' probability of looking away is an exponential of the sum of a baseline probability β_0, a linear term $\beta_2(x - \beta_1)$ and a quadratic term $\beta_3(x - \beta_1)^2$. The exponential is used because it is a standard linking function in the survival analysis used by KPA (e.g. Cox & Oakes, 1984; Klein, 1992; Hougaard, 2000; Ibrahim, Chen & Sinha, 2005); indeed, our model can be viewed as Bayesian framework very similar in spirit to KPA's original survival analysis, but one that makes several simplifying (and parametric) assumptions. The summing operation allows multiple influences (constant, linear, quadratic) to all simultaneously influence the outcome. While we think of this equation as specifying look-aways, it is important to point out that it also specifies *non*-look-aways when it predicts a probability of looking away close to 0. Thus, the data analysis model's

[6] In their formulation, this was computed by integrating over the distribution of events, not taking the expected value. These are closely related and using the expected value is a simplifying computational assumption here.

[7] We also upper bound the probability returned by this linking function at 1.0

inferences take into account both the places where infants do look away and the places where they do not.

Equation (1) formalizes a range of different linking functions as the B_i are allowed to vary.[8] Figure 3(a) illustrates several possible relationships between information and look-aways that our analysis could discover with this setup. β_0 characterizes the baseline look-away probability distribution: ignoring effects of information (i.e. $\beta_1 = \beta_3 = 0$), β_0 determines the raw probability of looking away at each time point. The green line in Figure 3(a) shows one line with *no* influence of information, with look-away probability constant over a wide range of information values. Next, β_2 and β_3, respectively, determine the *linear* and *quadratic* influence that information has on (log) look-away probability and B_1 shifts the curves left ($\beta_1 > 0$) and right ($\beta_1 < 0$). The black line in Figure 3(a) shows a linear influence of information (note the y-axis is logarithmically scaled). This would correspond to a subject who had a very high probability of looking away on high information events, and very low probability on low information events. An analogous line could give the opposite, a preference to look away on low information events compared to high information. The red line shows a quadratic influence with no linear influence: this would be a preference to look away on *either* highly surprising or highly non-surprising events. The purple line shows another possible relationship. Over the whole x-axis, this would be a parabola, but in the scale of information values in the plot (0 to 3), this curve gives a strongly asymmetric relationship that prefers only to look away on high information events. Analogously, one could choose the β_is to prefer to look away on *low* information events. The point illustrated by Figure 3(a) is therefore that the model allows many possible relationships in the data to be revealed by the analysis.

Figure 3(b) illustrates an important problem: by averaging the purple curve and its 'flip' – preferences for low and high information, respectively – one could observe a U-shape in the subject average (red dashed line in 3(b)). This means that the average-level U-shape found by KPA is not *necessarily* due to individual subjects preferring a particular information level. It could be that some subjects only look away to high information events, and some to low, and that in aggregate their behavior looks like a U. Teasing apart this possibility from a U-shaped relationship for individual infants is a major goal of the present paper.

Summary of variables

We have so far described several variables that are likely to influence each individual:

- α controls the learner's prior expectations that future events will look like previously observed ones.
- λ controls the learner's memory decay, the degree to which earlier events are discounted in predicting future events.
- β_0 controls the baseline probability of looking away at each event, or equally, the y (vertical) location of the U or linear curve.
- β_1 controls the x-offset, shifting the U curve or line horizontally.
- β_2 controls the *linear* influence of information on look-away probability.
- β_3 controls the *quadratic* influence of information on look-away probability.

It is important to note here that these variables are of very different types, in terms of cognitive theorizing. α controls the *assumptions* of the rational model posited to be in infants' mental representations – how much they think future events will be like past ones. λ controls the *limitations* of the rational model[9] – the degree to which imperfect memory for events influences their expectations about future events. Finally, the β_is control the *linking* between the rational model and behaviour – they parameterize the shape of each individual's response to events, whether they ignore information, prefer high or low information, or potentially show a quadratic relationship. It is also important to emphasize the difference between the structural assumptions of the model, and the parts of the model which are inferred from data. The model *assumes* that infants use a DM with some prior parameter α and memory decay λ, but infers their values from the behavioral data. The model also assumes that look-aways are *some* constant, linear, or quadratic function of the negative log probability of an event

[8] We note that Equation 1 is a nonstandard quadratic parameterization in that it contains more free parameters than are mathematically necessary. We could have chosen to fit equations of the form $\beta_0 + \beta_1 x + \beta_2 x^2$. We use Equation 1 because in it, similarity in coefficients corresponds directly to similarity in U-shape since the parameters (β_0, β_1) give the location of the bottom of the U and its shape (β_3). Without the extra parameters, similarities in U-shapes correspond to less obvious relationships between parameters since the location of the U depends on multiple parameters. Additionally, we have observed that the parameterization in (1) allows for more efficient inference (Markov Chain Monte Carlo), although we have not evaluated this quantitatively. However, our results-in particular the U-shapes observed in individuals-do not depend on which parameterization we use.

[9] At least to the degree in which imperfect memory can be considered a limitation—see Anderson and Schooler (1991).

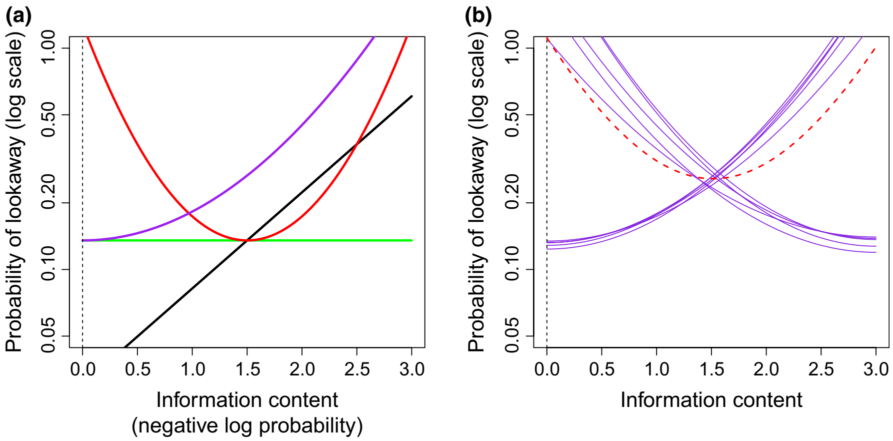

Figure 3 *Relation between information content and look-away probability for several types of hypothetical infants. In 3(a), the green line* ($\beta_0 = -2$, $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$) shows no influence of information on look-aways, corresponding to β_1 and β_2 not significantly different from zero. The black line ($\beta_0 = -2$, $\beta_1 = 1.5$, $\beta_2 = 1.0$, $\beta_3 = 0$) corresponds to a linear relationship between information and look-away log probability. This shows no quadratic effects. The red line, ($\beta_0 = -2$, $\beta_1 = 1.5$, $\beta_2 = 1.0$) shows a quadratic influence. Finally, the purple line ($\beta_0 = -2$, $\beta_1 = 0.0$, $\beta_2 = 0.3$) shows a response that is asymmetric over the range of value in the experiment, giving rise to primarily a preference to look away on high-information events. Averaging purple lines and their reflection-individual subject preferences for high and low information-can give rise to a U-shaped subject average, the dotted red line in (b).

according to this model, but leaves the specific function (specified by the β_i) to be determined from the data. The model is therefore constrained to make inferences about a class of DM models linked to look-away probability, but is free within that general class of models to discover different kinds of linking functions, priors, and memory decays.

Data analysis model: from individuals to groups

It is tempting to consider fitting the parameters summarized in the previous section simply on a by-subject basis. This approach is often used in psychophysics and is simple and intuitive. However, it is well-known that this type of analysis is not efficient since it does not share or 'pool' any information between subjects (see Gelman & Hill, 2007).[10] Ideally, we should use estimates about the group-level distribution of parameters to influence our estimates of individual subject parameters. This provides substantially better analysis, improving both the subject-level parameter estimates and allowing better inferences about the group-level parameters. In a simple example, if we are attempting to predict a driver's probability of an accident from, say, their gender, the best estimate will incorporate both information from the driver's previous driving record (observed data for the subject) and the priors given by their gender group average. This type of *partial pooling* is elegantly included in mixed effect or hierarchical Bayesian regression models (e.g. Gelman & Hill, 2007; Baayen, Davidson & Bates, 2008), which have become common in experimental psychology (e.g. Baayen *et al*., 2008).

Partial pooling is cached out in Bayesian data analysis by imagining that individual subject parameters are chosen according to some (unobserved) overall group distribution. The model is then hierarchical: the data (responses) are informative about individual subject parameters, and the individual subject parameters are themselves informative about the group distribution. By doing inference about the group distribution, such models effectively combine information from subjects and allow individual subject data to inform hypotheses about the group average in a statistically 'correct' way. Somewhat counter-intuitively, at the same time, the hypothesized group mean is informative about individual subjects and can improve our estimates of noisy individual subjects' parameters. So, with the statistical model described in the previous section, individual infants' β_3 (the quadratic effect) might be clustered around a group mean, here denoted $\hat{\beta}_3$, that equals, perhaps 0.5, with a standard deviation of 0.2. By doing inference over the location of this group mean $\hat{\beta}_3$ (and the group variance about this mean), we can test the

[10] Indeed, in exploratory work, analyses without pooling did not lead to robust results, despite the fact that infants provided 32 trials each.

statistical hypothesis that the group average quadratic effect is significantly nonzero, while controlling for the fact that individual infants may have different values for β_3. Formally, we use a scaled inverse-Wishart setup from Gelman and Hill (2007), which also captures the subject-wise variance and covariance between the β_i. This setup essentially fits a covariance matrix to individual subject coefficients in the population allowing the model to capture systematicities in individual subject parameters. We chose this because in regression it is a better choice than, for instance, assuming that each individual subject's random parameters are independent (since they are unlikely to be). Although we have not exhaustively explored other options, it is likely that our results do not depend strongly on this particular form of regression.

We also would like to bring the advantages of partial pooling to α and λ. Unlike the regression parameters β_i, α and λ are both required to be non-negative, and so treating them as normally- distributed (as with the β_is) is not appropriate. We instead assume that individual values of α and λ, here denoted α_s and λ_s, come from a group-level *gamma* distribution. The gamma distribution is a common distribution on non-negative real numbers that can control both the size and variability of the distribution of these parameters. In the full Bayesian setup, the parameters of this group-level gamma distribution require priors themselves. Since we do not know what these should be, we choose a form of prior which biases the model as little as possible, known as *reference priors* (from Yang & Berger, 1996; Sun & Ye, 1996). Such priors are – by definition – designed to allow the data to have the largest possible influence on the model (Bernardo, 1979; Berger, Bernardo & Sun, 2009), allowing us to avoid the influence of the priors as much as possible. We chose these over alternatives like an arbitrary parametric form specifically because we did not have strong expectations about these parameter values, so we wanted to allow the data to 'speak for itself' as much as possible. Effectively, what this means is that we 'build in' very few expectations about the group-level distribution of α and λ, and infer both the group distribution and the individual subject values of these parameters.

Methods

We have so far described the setup of the probabilistic model used in our data analysis. The logic of Bayesian data analysis (for reference, see Gelman *et al.*, 2004; Kruschke, 2010a, 2010b) is that we specify a probabilistic model, as above, and then do *inference* over the unobserved parameters given the observed data (behavioral responses). This inference will tell us what parameter values are likely given the data we have observed. Formally, Bayesian techniques provide a *posterior distribution* on the model parameters, which specifies how likely any particular parameter value is to be the true one, given the observed data (see, e.g. Griffiths & Yuille, 2008; MacKay, 2003; Kruschke, 2010b, for introductions). For instance, it will tell us likely ranges for each individual's β_i, α, and λ, as well as the likely ranges of the group distribution.

In general, it is often a substantial challenge to take the assumptions and observed data, and precisely determine the posterior distribution. Instead, a standard technique in Bayesian inference is to generate *samples* from the posterior distribution, which then can be used to compute relevant values such as parameter means, medians, and ranges. This technique, *Markov-Chain Monte-Carlo* (MCMC) (see chapter 29 of MacKay, 2003, for an introduction) essentially takes a biased random walk around the space of parameter values such that in the limit, it draws samples from the correct posterior distribution, P(variables | data), where the variables are the ones described above and the data are the observed behavioral look-away patterns. An intuitive way to think of MCMC is as a hill-climbing algorithm that tries to 'fit' the variables to best explain the data, except that sometimes the algorithm 'climbs' downhill (e.g. moving towards worse-fitting parameters) in order to more completely explore the space of parameter values. In fact, it achieves a perfect balance of uphill and downhill climbs such that in the long-term average, it correctly samples from the distribution. Thus, through MCMC, we are able to discover the range of likely parameter values under the assumption that the formal model we constructed is (at least close to) correct. A full review of this technique and the associated mathematics is beyond the scope of the present paper, so we refer readers to primary sources (e.g. MacKay, 2003; Andrieu, De Freitas, Doucet & Jordan, 2003; Gelman *et al.*, 2004; Griffiths & Yuille, 2008). We use a programming package for Bayesian inference called PyMC (Patil, Huard & Fonnesbeck, 2010), which is similar to BUGS (Gilks, Thomas & Spiegelhalter, 1994; Lunn, Thomas, Best & Spiegelhalter, 2000), but substantially more powerful.[11] Using PyMC, we ran MCMC for 100,000 steps of burn-in, and 150,000 steps of samples, drawing a sample every 100 steps. We assessed convergence using eight multiple chains, and confirming that each gave similar results despite starting from different initial conditions. This took approximately 48 hours on a 2.6 GHz computer. In order to prevent having to compute negative log prob-

[11] Running code is available from the first author.

ability values for every real-valued sample of α_s and λ_s, we discretized these variables by steps of 0.2 from 0.0 to 20.0 for α_s and 0.1 from 0.0 to 10.0 for λ_s. Inference was run over the discrete versions of these parameters, as an approximation to the full continuous model.

MCMC produces samples of the parameter values from the data analysis model's posterior distribution. These samples tell us where we should believe the true parameter values lie, given the observed data. The samples can be used to compute quantities such as the expectation of parameter values, range and distribution, and quantiles of the posterior distribution. That is, the output of the model gives the distribution of likely parameter values, given the behavioral data we see. When samples of individual parameters are examined, they correspond to *marginal* estimates of parameters, meaning that they integrate out the uncertainty in other parameters. This means that the model's samples for each parameter automatically (and optimally) account for uncertainty in the rest of the model. In particular, by marginalizing over the unknown cognitive parameters like α_s and λ_s, we are able to make inferences about the other parameters (the β_i) while correctly controlling for the cognitive parameters' values even though we do not know them exactly.

In the analysis, we primarily focus on the mean of each parameter's posterior distribution – a good guess for its value – and each variable's 95% *highest posterior density* (HPD) interval, which quantifies the range and precision of our parameter estimate.

Results and discussion

Figure 4 shows individual-level response curves for each infant, the primary result of the present paper. Here, the *y*-axis shows each infant's probability of looking away on an item of a given information content (*x*-axis). Since individual differences in α_s and λ_s lead to different information content ranges for each infant, each curve is plotted for the infant's information content range as determined by their estimated median α_s and λ_s. Thus, the *x*-axes differ slightly for each infant because each individual has a different α_s and λ_s, giving a different range of information content. Most individuals demonstrate a clear U-shaped relationship between information and look-aways.[12] The prevalence of U-shapes here indicates that the effect found by KPA was *not* an artifact of some infants preferring predictable events and some preferring unpredictable events. In that case, we would expect to observe primarily increasing- and decreasing-curves within individuals, not U-shapes. The group-level tendency towards U-shaped relationships can best be statistically evaluated by examining the posterior distribution of group-level coefficients, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ shown in Figure 5(a). This posterior distribution shows how much belief we should place in any particular estimate of these values, showing the range of plausible linking function parameters given infants' data. For our purposes, the most important of these coefficients is $\hat{\beta}_3$, which gives the *quadratic* influence of information content on look-away probability. This variable's 95% highest posterior density intervals are well away from 0, indicating that the quadratic group trend is statistically significant. The other variables have interesting distributions: the linear term $\hat{\beta}_2$ is substantially negative, indicating a strong decreasing tendency in the model, but which is offset by the intercept term $\hat{\beta}_0$ and the quadratic term $\hat{\beta}_3$. The interaction of these four variables can more easily be visualized by showing the group average linking function that they imply. Figure 5(b) shows a clear U-shaped tendency for look-aways centered on information content values near 1. We note that this shows a stronger pattern than reported in KPA. This is because, unlike KPA's analysis which collapsed over individuals, Figure 5(b) essentially subtracts out noise due to individual differences by fitting individual parameters. This leads to a more robust and clear estimation of the group curve.

In addition, the plots in Figure 4 have been colored if the left increase of the U-shape is more than 2 times higher than the right (red); none had the right more than 2 times the left. Only two of the 41 infants show this pattern, indicating that overall the trend is towards relatively symmetric U-shapes.

Note, though, that these individual curves cannot be estimated very reliably: the area between the dotted lines denotes the 95% posterior probability mass bounds for these curves, giving an estimate of our confidence in each. However, we should expect that if we had a more complete picture of infants' experience – both in the experiment and in real life – our ability to model individual curves might be substantially enhanced. Put another way, the uncertainty inherent in these curves may be due to the fact that our cognitive model is an extremely simple approximation of much more complex cognitive processing.

Using the model, we can also investigate the individual subject parameters for memory decays (λ_s) and priors (α_s). Figure 6 shows one row for each subject, with their estimated values for these two parameters,

[12] Note the location of the bottom of the U differs from KPA's results. This is because here we fit memory decay and alpha parameters, each of which alters the numeric range of information content values.

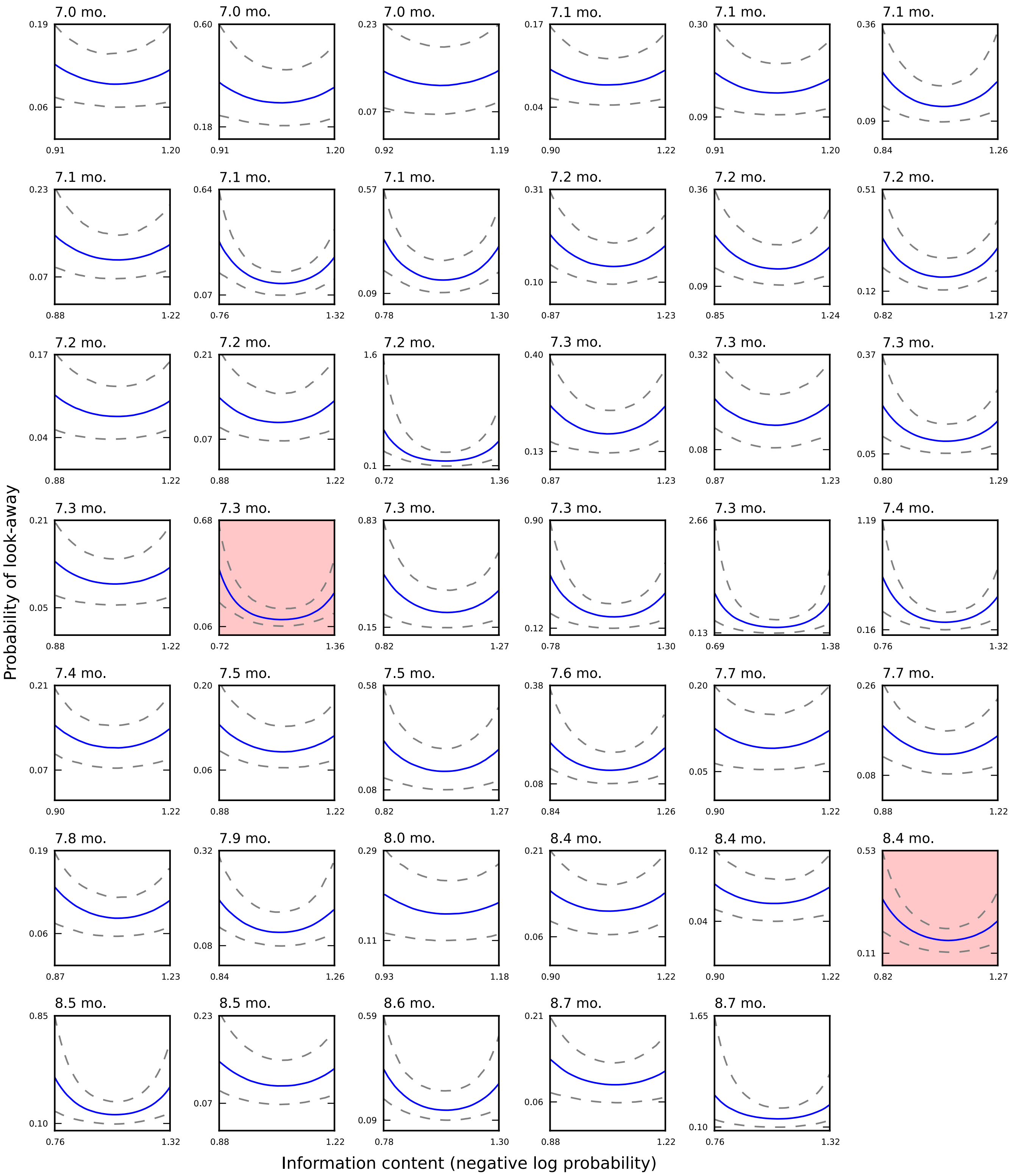

Figure 4 *Individual linking functions recovered by the data analysis model relating information content (x-axis) to probability of look-away (y-axis) for each infant (log scale). The area between the dotted error bars denotes 95% of the posterior probability mass for these curves. Red boxes show primarily decreasing curves, where the left upturn is more than 2x the right; no infants showed the reverse pattern of a right upturn more than 2x the left.*

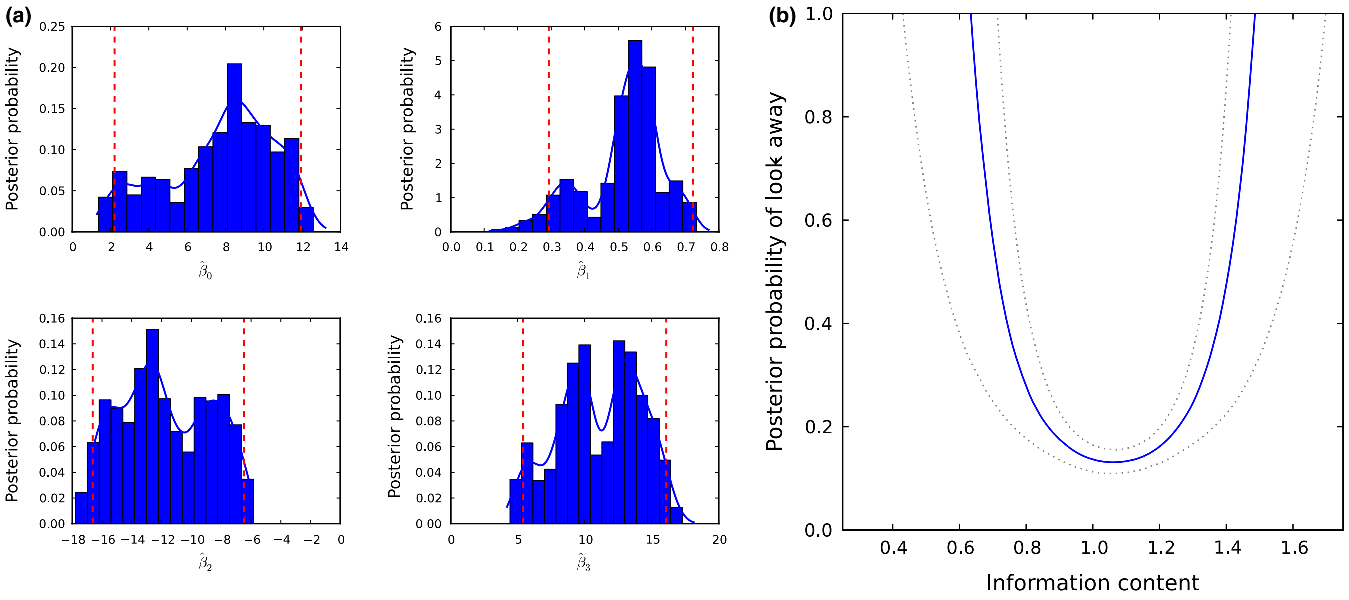

Figure 5 *(a): Group-level coeffi_cients for each of the parameters of the linking function, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. This shows the model's posterior estimates for these parameters, corresponding to how confident we should be in the group-level effects given the experimental data. The (red) dotted lines show 95% highest posterior density (HPD) intervals for each coefficient. (b): mean group-level curve relating information content to look-away probability, using the (full joint) distribution of the coefficients that are plotted in (a). The dotted line shows a 95% HPD interval for this line at each* x *location.*

and 95% highest-posterior density ranges (horizontal lines). This figure shows that infants tend to have values of the prior parameters, α_s, around 0.1, with a range in uncertainty of this estimate from close to 0.0 up to about 0.4. These numbers are only interpretable relative to the number of 'counts' provided by the actual data, which is determined by the memory decay parameters λ_s, also shown in Figure 6. λ_s values average about 4, with a range of about 2.5 to 5.5. As above, one simple way to visualize the effect of these parameter values is to imagine that a learner had observed the sequence *AB* and was then attempting to predict the next element. Following the range of values in Figure 6, we assume $\alpha = 0.1$ and $\lambda = 4$. These parameter values correspond to the center pie plot of Figure 2, and so can be understood intuitively by comparing the center plot to the others in Figure 2. Thus, in the example sequence *AB*, *A* would receive an effective count of $(2 + 1)^{-4} \approx 0.01$, *B* would receive an effective count of $(1 + 1)^{-4} \approx 0.06$, and *C* would have a count of 0 since it does not occur. These would then be 'smoothed' by $\alpha = 0.1$. Event *A* then has a probability proportional to $(0.01 + \alpha)$, or equal to 0.30. Event *B* has a probability proportional to $(0.06 + \alpha)$, or equal to 0.43. The unobserved event *C* would have a probability proportional to $(0 + \alpha)$, or equal to 0.27. These probabilities indicate both a substantial amount of memory decay

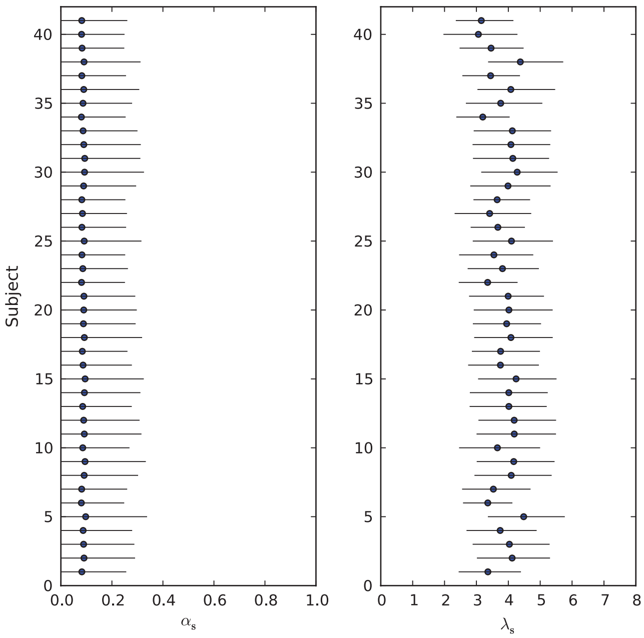

Figure 6 *This shows each subject's mean posterior estimate for the DM prior α_s (x-axis) and each subjects' mean posterior estimate for their memory decay parameter, λ_s. There is relatively little variation in individuals' estimated α_s and λ_s, but considerable uncertainty in the individual estimates. The numerical values of these variables results in a moderate range of prior beliefs and memory decay (see text).*

(the probability of A is not very different from C) and a moderate amount of smoothing (C has probability close to ⅓ even though it is unobserved). These are reasonable expectations and accord with our intuitive sense of which inferences in Figure 2 are most plausible: infants likely have difficulty remembering past events and so base inferences more strongly on recent events. Infants may smooth these distributions substantially because they observe multiple trials where objects appear from behind each of the boxes, and so may learn that as-of-yet unobserved events will eventually be seen. It will be important for future work to examine how these kinds of parameters may change over the course of development, and how this interacts with later (or earlier) stages of learning.

Importantly, these results provide strong evidence for infants' learning. There are several ways in which the data analysis model could have discovered that infants do not learn throughout the course of these sequences. For instance, if infants really did not form predictive expectations about the stimuli, the 'best' model parameters might have high α_s, making all events appear to have equal information content, or very high λ_s, making infants' inferences ignore essentially all data. Similarly, the 'best' linking functions for infants might have been flat (Figure 3(a)), indicating no relationship between information and look-aways. The fact that the model does *not* tend towards these parameter settings is indicative of regularities in infants' responses that would be expected by KPA's theory – statistical learning of sequences combined with a quadratic linking function between information and look-away. In Appendix A we simulate data and show that when given bimodal data (some infants prefer predictable events, some prefer unpredictable), the same data analysis model is able to recover these two kinds of infants and their corresponding increasing and decreasing curves. This suggests that the U-shapes recovered here are not artifacts of the analysis method, but represent actual learning and attentional effects in individual infants.

This finding broadly suggests that statistical learning and inference mechanisms themselves may play an active role in selecting which data should be subject to deeper processing. If correct, this paints a starkly different picture of statistical learning mechanisms than might be expected by application of the most common Bayesian models. In typical models, the data are simply provided 'as-is', usually through observation or an experimental protocol. Such models often well-explain a wide range of developmental phenomena (see Perfors *et al.*, 2011). However, real-world statistical learning may have a somewhat different nature, where learners actively select which data points should be ignored, and which should be studied in more depth (see Bardhan, Aslin & Tanenhaus, 2010; Gureckis & Markant, 2012, for recent examples of active sampling without instruction). This is a reasonable strategy for learners with constrained processing mechanisms (e.g. limited working memory, etc.) or limited memory. It is even plausible that the simple statistical learning paradigms studied in the lab may only scale to real-world sized data sets (e.g. learning an entire vocabulary) with inclusion of some kind of attentional selection mechanisms (Turk-Browne, Jungé & Scholl, 2005; Toro, Sinnett & Soto-Faraco, 2005; Creel, Newport & Aslin, 2004). As suggested by our findings, statistical learning mechanisms themselves may help decide which data points are most useful, leading to an interesting circularity. We believe that it is an important direction for future modeling work – especially related to very early development – to explore how such individual data selection might influence the trajectory and outcome of idealized statistical inference.

The method we have developed connects an idealized cognitive model with an idealized data analysis model, providing, in some sense, the best of both worlds. This technique allows us to specify a number of extraneous factors that we think might matter – like memory decay – and perform optimal statistical inferences about an aspect of infants' representational system given their behavior. The primary advantage of this is not just in avoiding possible pitfalls of null hypothesis significance testing (Edwards, Lindman & Savage, 1963; Cohen, 1994; Lee & Wagenmakers, 2005; Wagenmakers, 2007; Kruschke, 2010a, 2011), but that it provides a tractable way for formalized models to make contact with behavioral data, an important emerging trend that will be critical for working out precise scientific theories of learning and development. Importantly, our general approach is much more broadly applicable than just KPA's original experiment. Indeed, we expect that as modeling and experimental approaches converge, this type of method will be able to help resolve key issues relating to individual differences, such as distinctions between fast and slow habituators (McCall & Melson, 1969; McCall & Kagan, 1970; McCall, Hogarty, Hamilton & Vincent, 1973; DeLoache, 1976; Baillargeon, 1987).

Conclusion

These results demonstrate that the U-shaped relationship between information and look-away probability is likely *not* due to two separate kinds of infants, some who prefer low complexity and some who prefer high complexity.

Instead, individual infants appear to show the trend observed by KPA in group averages of a preference for medial information rates. This lends support to KPA's interpretation that a 'Goldilocks' preference allows individuals to learn in a complex environment, filtering out information sources which are either too simple or too complex.

The methods used in this paper illustrate that cognitive modeling can be usefully combined with rich data analysis. While it is compelling to build rational models, it is even more powerful to build rational models into analysis frameworks that allow for strong tests of the assumptions and limitations of the rational model, as well as its precise relation to behavior. From infant data, we infer intuitively plausible ranges of values for unobserved cognitive parameters, and perhaps more importantly, our key result of theoretical interest is confirmed while controlling for (marginalizing out) these unknown parameters. One important aspect of this work is our ability to do inference over the space of possible linking functions between hypothesized beliefs and observed behavior. This was possible by making plausible assumptions where absolutely necessary, and allowing the data to have the greatest influence on factors of theoretical interest. For instance, we 'build in' to the analysis that infants have similar linking functions and parameters, but allow substantial variation in the shape of the group-level and individual-level linking functions themselves. The assumption of similar parameters across individuals allows for partial pooling of information from each infant. This general approach of leaving the key parts of the model as unspecified as possible illustrates that strong quantitative analysis can be fruitfully applied to noisy infant data. We expect that even richer data analysis methods can be developed for infant experiments: for instance, one could apply functional mixed effect models (Guo, 2002) or richer nonparametric methods within individuals (e.g. N. Smith & Levy, 2008) in this type of framework, perhaps also with an explicit nonparametric Bayesian clustering model over individuals (Yurovsky *et al*., 2012).

We believe that the combination of probabilistic cognitive models with Bayesian data analysis is a powerful tool. Bayesian cognitive models encourage principled formalization of inferential theories, and Bayesian data analysis allows for principled formalization of statistical analysis. More generally, of course, one could combine rich data analysis with any parameterized cognitive model – not just Bayesian ones. However, our approach is especially well suited to Bayesian cognitive modeling: Bayesian cognitive models provide a superb 'first approximation' to any cognitive task because they embody the statistically optimal solution to a problem learners face.[13] Critically, Bayesian data analysis allows us to incorporate *limitations* into the rational model – in our case, a memory decay – without having to make a priori assumptions about how constraining or important those limitations are. We do not, for instance, assume a particular value of λ. Bayesian data analysis then gives the best of both worlds: a fully- formalized rational model, and an inferentially- optimal scheme to illuminate limitations to that rational model, given the behavioral data. As future work requires building and testing increasingly complex models, we believe that the general framework presented in this paper provides a powerful linking of theory, analysis, and experiment, using probabilities *all the way down*.

Acknowledgements

We'd like to thank Dan Yurovsky and Roger Levy for helpful discussions about this project, Holly Palmeri for running subjects, Johnny Wen for help programming the experiment, and two anonymous reviewers for great suggestions. We are grateful to the developers and maintainers of PyMC, free and open source software that substantially simplified modeling. In particular, we would like to thank Chris Fonnesbeck for bug fixes related to our model. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number F32HD070544. This work was also supported by NIH grant HD-037082 to R. Aslin and a National Science Foundation Graduate Research Fellowship Program (GRFP) award to C. Kidd. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Anderson, J., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, **2**(6), 396–408.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, **50**(1), 5–43.

[13] It is no accident that the best inferential tools available in science are also hypothesized to support humans' own development. Probability theory and Bayesian inference constitute what Jaynes (2003) called "the logic of science": they allow for optimal inferences to be made from data, and in many cases constitute the cleanest way to conceptualize induction-scientific or cognitive.

Aslin, R. (2007). What's in a look? *Developmental Science*, **10** (1), 48–53.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**(4), 390–412.

Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology*, **23**(5), 655–664.

Bardhan, N., Aslin, R., & Tanenhaus, M. (2010). Adults' self-directed learning of an artificial lexicon: the dynamic of neighborhood reorganization. In *Proceedings of the Thirty-second Annual Conference of the Cognitive Science Society.*

Berger, J., Bernardo, J., & Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, **37**(2), 905–938.

Bernardo, J. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, **41**, 113–147.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, **3**(2), 57 –65.

Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences*, **10**(7), 287–291.

Chen, S.F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, **13**(4), 359–393.

Civan, A., Teller, D., & Palmer, J. (2005). Relations among spontaneous preferences, familiarized preferences, and novelty effects: measurements with forced-choice techniques. *Infancy*, **7**(2), 111–142.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, **49**(12), 997–1003.

Cox, D., & Oakes, D. (1984). *Analysis of survival data*. London: Chapman & Hall.

Creel, S.C., Newport, E.L., & Aslin, R.N. (2004). Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **30**(5), 1119–1130.

DeLoache, J. (1976). Rate of habituation and visual memory in infants. *Child Development*, **47**(1), 145–154.

Dember, W., & Earl, R. (1957). Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychological Review*, **64**(2), 91–96.

Dewar, K., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge evidence from 9-month-old infants. *Psychological Science*, **21**(12), 1871–1877.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology.* Columbia, SC: Teachers College, Columbia University.

Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**(3), 193–242.

Fiser, J., & Aslin, R. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(24), 15822–15826.

Geisler, W. (2003). Ideal observer analysis. In L.M. Chalupa & J.S. Werner (Eds.), *The Visual Neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis.* Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gerken, L., Balcomb, F.K., & Minton, J.L. (2011). Infants avoid `labouring in vain' by attending more to learnable than unlearnable linguistic patterns. *Developmental Science*, **14**(5), 972–979.

Gilks, W., Thomas, A., & Spiegelhalter, D. (1994). A language and program for complex Bayesian modelling. *The Statistician*, **43**(1), 169–177.

Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of science*. Chicago, IL: University of Chicago Press.

Goodman, N., Mansinghka, V., & Tenenbaum, J. (2007). Learning grounded causal models. In *Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society* (pp. 305–310).

Griffiths, T., & Yuille, A. (2008). A primer on probabilistic inference. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 33–57). Oxford: Oxford University Press.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, **58**(1), 121–128.

Gureckis, T.M., & Markant, D.B. (2012). Self-directed learning: a cognitive and computational perspective. *Perspectives on Psychological Science*, **7**(5), 464–481.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models.* Boca Raton, FL: Chapman & Hall/CRC.

Hosmer, D., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data* (2nd edn.). Hoboken, NJ: John Wiley & Sons.

Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer.

Hunter, M., & Ames, E. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, **5**, 69–95.

Ibrahim, J., Chen, M., & Sinha, D. (2005). *Bayesian survival analysis.* Wiley Online Library.

James, W. (1890). *The principles of psychology* (Vol. 1). Cambridge, MA: Harvard University Press.

Jaynes, E. (2003). *Probability theory: The logic of science.* Cambridge: Cambridge University Press.

Jeffreys, S. (1961). *Theory of probability.* Oxford: Oxford University Press.

Kaldy, Z., Blaser, E.A., & Leslie, A.M. (2006). A new method for calibrating perceptual salience across dimensions in infants: the case of color vs. luminance. *Developmental Science*, **9**(5), 482–489.

Kidd, C., Piantadosi, S., & Aslin, R.N. (2012). The Goldilocks effect: human infants allocate attention to events that are neither too simple nor too complex. *PLoS ONE*, **7**(5), e36399.

Kinney, D., & Kagan, J. (1976). Infant attention to auditory discrepancy. *Child Development*, **47**(1), 155–164.

Klein, J. (1992). *Survival analysis: state of the art* (Vol. 211). New York: Springer.

Klein, J., & Moeschberger, M. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd edn.). New York: Springer-Verlag.

Kruschke, J. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, **1**(5), 658–676.

Kruschke, J. (2010b). Doing Bayesian data analysis: a tutorial with r and bugs. *Brain*, **1**(5), 658–676.

Kruschke, J. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, **6**(3), 299–312.

Lee, M., & Sarnecka, B. (2010a). A model of knower-level behavior in number concept development. *Cognitive Science*, **34**(1), 51–67.

Lee, M., & Sarnecka, B. (2010b). Number-knower levels in young children: insights from Bayesian modeling. *Cognition*, **120**(3), 391–402.

Lee, M., & Wagenmakers, E. (2005). Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychological Review*, **112**(3), 662–668.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**(4), 325–337.

McCall, R., Hogarty, P., Hamilton, J., & Vincent, J. (1973). Habituation rate and the infant's response to visual discrepancies. *Child Development*, 280–287.

McCall, R., & Kagan, J. (1970). Individual differences in the infant's distribution of attention to stimulus discrepancy. *Developmental Psychology*, **2**(1), 90–98.

McCall, R., & Melson, W. (1969). Attention in infants as a function of magnitude of discrepancy and habituation rate. *Psychonomic Science*, **17**(6), 317–319.

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

McMurray, B., & Aslin, R. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, **95** (2), B15–B26.

Patil, A., Huard, D., & Fonnesbeck, C. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, **35**(4), 1–81.

Perfors, A., Tenenbaum, J., Griffiths, T., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, **120**(3), 302–321.

Piantadosi, S. (2011). Learning and the language of thought. Unpublished doctoral dissertation, MIT.

Piantadosi, S., Tenenbaum, J., & Goodman, N. (2009). Beyond Boolean logic: exploring representation languages for learning complex concepts. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*

Roder, B., Bushnell, E., & Sasseville, A. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy*, **1**(4), 491–507.

Rose, S., Gottfried, A., Melloy-Carminar, P., & Bridger, W. (1982). Familiarity and novelty preferences in infant recognition memory: implications for information processing. *Developmental Psychology*, **18**(5), 704–713.

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

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

Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, **35**(4), 606–621.

Sarnecka, B., & Lee, M. (2009). Levels of number knowledge during early childhood. *Journal of Experimental Child Psychology*, **103**(3), 325–337.

Shannon, C. (1948). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Smith, L., & Yu, C. (2008). Infants rapidly learn word–referent mappings via cross-situational statistics. *Cognition*, **106**(3), 1558–1568.

Smith, N., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 595–600).

Sokolov, E. (1963). *Perception and the conditioned reflex*. Oxford England: Pergamon.

Sun, D., & Ye, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika*, **83**(1), 55–65.

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*, **27**(332), 1054–1059.

Toro, J.M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, **97**(2), B25–B34.

Turk-Browne, N.B., Jungé, J.A., & Scholl, B.J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, **134**(4), 552–564.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, **14**(5), 779–804.

Wagner, S., & Sakovits, L. (1986). A process analysis of infant visual and cross-modal recognition memory: implications for an amodal code. In L. Lipsitt & C. Rovee-Collier (Eds.), *Advances in infancy research*, (4), 195-217. Norwood, NJ: Ablex.

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, USA*, **105**(13), 5012–5015.

Yang, R., & Berger, J. (1996). *A catalog of noninformative priors.* Institute of Statistics and Decision Sciences, Duke University.

Yurovsky, D., Hidaka, S., & Wu, R. (2012). Quantitative linking hypotheses for infant eye movements. In *Proceedings of the Thirty-fourth Annual Conference of the Cognitive Science Society*

Received: 11 October 2012
Accepted: 29 November 2012

Appendix

Simulation results

To establish that the data analysis model presented here would correctly recover a mixture of two kinds of subjects – some who prefer high information content and some who prefer low – we created simulated data with this property and provided it to the data analysis model. For each of 41 simulated subjects, we 'ran' the experiment where each individual's probability of looking away at each point in the sequence was characterized by the four parameters, $\beta_0 \sim Normal(-2, 1)$, $\beta_1 \sim Normal(1.0, 0.1)$, $\beta_2 \sim Normal(\kappa \cdot 6, 1.0)$, $\beta_3 = 0$, where $\kappa = 1$ for even numbered subjects and $\kappa = -1$ for odd. Each of these parameters was sampled once for each individual, and used to determine their look-away probability throughout each sequence. The bimodal property of β_2 implemented through κ makes the even numbered subjects prefer to maintain attention to predictable (low information content) events, and the odd numbered subjects prefer to maintain attention to unpredictable (high information content) events. For the simulated data, we fix $\alpha = 0.1$ and $\lambda = 4.0$, approximately the values observed in KPA's data. Note that, as with the real data, all of these parameters must be estimated for each subject using the behavioral measure of which item in the sequence led to termination of attention.

The subject-wise curves recovered by the model are shown in Figure A1. This illustrates that the data analysis correctly recovers the increasing and decreasing curves for even and odd numbered subjects, respectively, with the exception of one subject (number 35). As in Figure 4, the red plots indicate curves where the left increase is more than 2 × the right, and the green indicate curves where the right is more than 2 × the left – 95% posterior probability curves are shown in dotted lines.

The success of the model on this simulated data indicates that the unimodal individual curves recovered by our model on real data reflects true properties of the data – the bimodality is not an artifact of the analysis approach. In particular, it is not due to the partial pooling of parameters implemented in our hierarchical Bayesian analysis. More generally, this indicates that our approach of recovering within-individual estimates of the relationship between predictors and behavior is amenable to discovering other population-level distributions of linking functions.

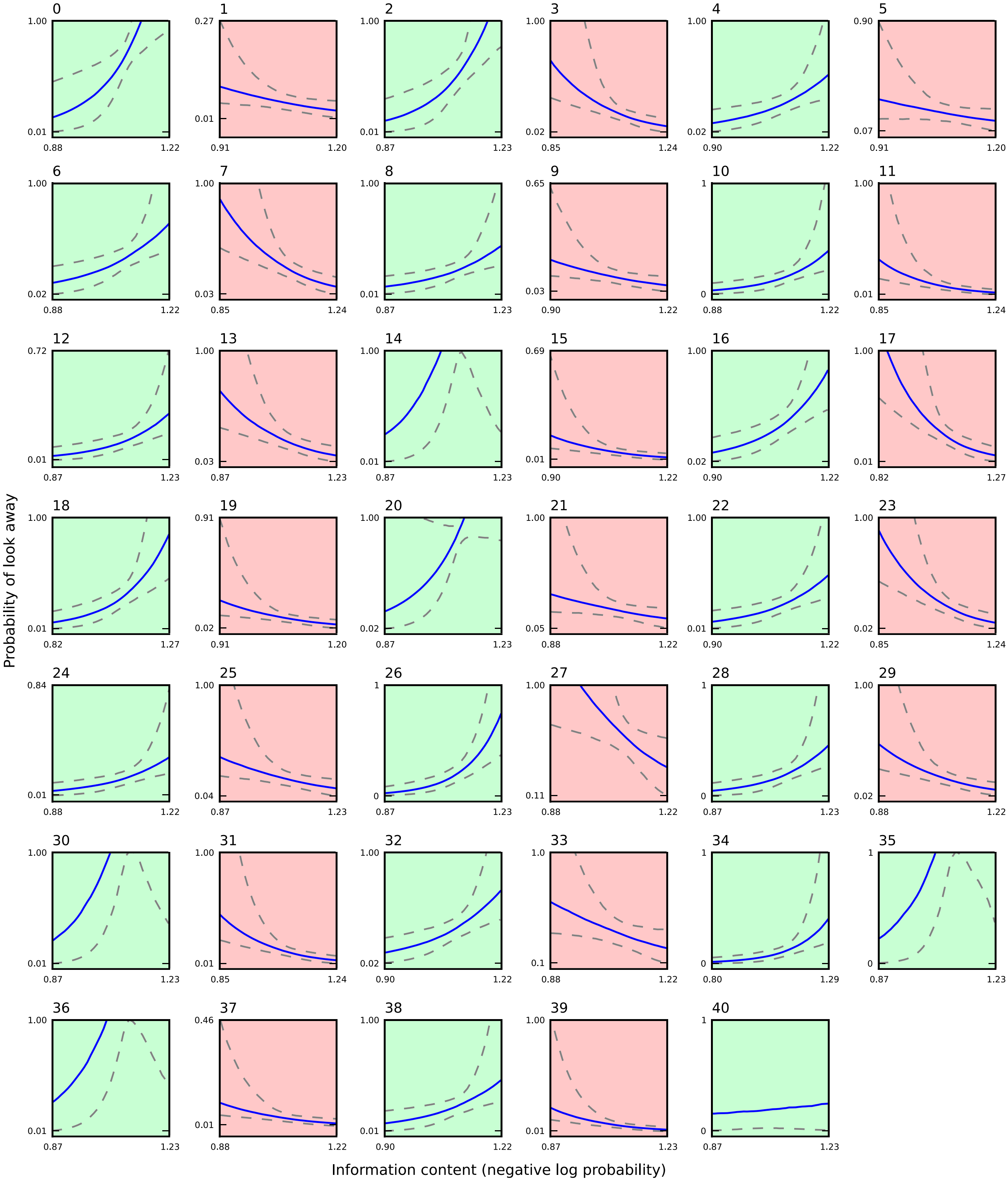

Figure A1 *Individual linking functions recovered by the data analysis model relating information content (x-axis) to probability of look-away (y-axis) for each infant (log scale), using* simulated *bimodal data. Green boxes indicate data biased for simplicity of information content (positive slope) and red boxes biased for complexity (negative slope). Simulated subject 35 in the only one misclassified in this analysis.*