Running head: ERROR DISCOUNTING

Supplementary material for

"Error Discounting in Probabilistic Category Learning"

by Craig, Lewandowsky, and Little

Stewart Craig and Stephan Lewandowsky

School of Psychology

The University of Western Australia

Daniel R. Little

The University of Western Australia, Indiana University, and The University of Melbourne

Keywords: categorization, learning, error, relevance shifts

School of Psychology

University of Western Australia

Crawley, W.A. 6009, AUSTRALIA

lewan@psy.uwa.edu.au

URL: http://www.cogsciwa.com/

Abstract

Craig, Lewandowsky, and Little (in press) used computational simulations to explore whether people discount errors in probabilistic category learning. This supplementary document provides full descriptions of the computational models and presents simulation results intended to complement those reported in Craig et al. (in press). This supplement is not intended as a stand-alone document. Refer to Craig et al. (in press) for further explanation.

Supplementary material for

"Error Discounting in Probabilistic Category Learning"

by Craig, Lewandowsky, and Little

Craig et al. (in press) explored the existence of error discounting in probabilistic categorization. Part of their analysis involved the use a series of computational models, the GCM (Nosofsky, 1986), MAC (Stewart, Brown, & Chater, 2002), RASHNL (Kruschke & Johansen, 1999), and ATRIUM (Erickson & Kruschke, 1998). Craig et al. (in press) found that the MAC, RASHNL, and ATRIUM fit data from two probabilistic categorization experiments significantly better when they included a mechanism to discount errors. The GCM did not include an error-discounting mechanism, instead poor performance of GCM ruled out an alternative sample-size explanation for error discounting. This supplement provides additional details of the models, modeling procedures, and results from the fits with each of the four models.

GCM: A Sample-Size Explanation

A possible explanation for the experimental results in Craig et al. (in press) was that the slow post-shift learning was not due to a discounting of error, but instead was an automatic by-product of the inevitable increase in memorized "sample size" during learning. Specifically, in the very early stages of training, when few items had been presented, on an exemplar view each further individual stimulus will have a relatively large impact. However, at later stages of learning, new items become increasingly insignificant in relation to the overall number of exemplars already encountered and memorized, thus limiting their impact. The slow adaptation to the shift in the experiments could therefore result from the small impact of items later in training relative to items early in training. We explored this alternative within the GCM (Nosofsky, 1986). The GCM contains no

associative learning mechanism but represents all encountered instances in memory, thus providing a quantitative instantiation of the sample-size hypothesis.

GCM Specification. The GCM assumes that on each trial the current item activates all previously encountered stimuli stored in memory according to:

$$s_{ij} = exp(-c \times d_{ij}), \tag{1}$$

where the similarity, s_{ij}, between items i and j is determined by the distance between them in psychological space, $d_{ij} = |x_i - x_j|$, which in the present case involves only a single dimension. (Note that for simplicity of exposition, all equations reported in this supplement are tailored to the fact that our stimuli were uni-dimensional.) The specificity parameter, c, determines the sharpness of the exponential function.

Similarities are converted to response probabilities by applying Luce's choice rule (Luce, 1963):

$$P(A|i) = \frac{(\sum_{j \in A} s_{ij})^{\gamma}}{(\sum_{j \in A} s_{ij})^{\gamma} + (\sum_{j \in B} s_{ij})^{\gamma})}, \tag{2}$$

where the response scaling parameter, γ, allows responding to vary between probability-matching when $\gamma \simeq 1$ and maximizing when $\gamma \gg 1$ (Ashby & Maddox, 1993; Nosofsky & Johansen, 2000). Thus, upon presentation of a test stimulus, it is compared against all stored exemplars in each of the categories separately, and a response is selected based on which category yields the greatest summed similarity.

GCM Simulations. For the square stimuli used in the present study, the perceived psychological distance between adjacent stimuli has been shown to be equivalent to the actual perceptual distance between the stimuli (Colreavy & Lewandowsky, 2008). Thus, for the simulations, the four stimuli were coded as the integers 1 to 4. The GCM was fit separately to each participant's mean response probabilities for all items in all blocks. The

GCM was presented with the training sequences shown to participants. Parameters (c and γ) were capped at 25. Both parameters were estimated using the SIMPLEX algorithm (Nelder & Mead, 1965) to minimize the negative binomial log-likelihood:

$$-\ln L = -\sum_i d_i \ln(p_i) + (n_i - d_i)\ln(1 - p_i), \tag{3}$$

where p_i is the model's predicted probability of category A for item i, d_i is the observed number of A responses made for item i, and n_i is the number of times item i was presented.

Table 1 shows the GCM's estimated parameters for both experiments. As shown in Figures 1 and 2, the GCM failed to capture the data in some crucial ways: In particular, it was unable to adjust its predictions in response to the probability shift. The GCM only managed to capture the early condition of Experiment 1, presumably because only 10 presentations of each item had occurred before the shift. However, in all other conditions, the GCM could not accumulate enough new evidence within the number of training trials to reverse its predicted probabilities. (Note, however, that there is a downward trend in the predicted slopes after the shifts, which indicates that given massively extended training, the GCM's performance might come to mirror the final outcome probabilities). In conclusion, the experimental results in Craig et al. (in press) cannot be accommodated by an explanation based solely on memorized sample size.

MAC model: Error discounting and Decisional Recency

The MAC model (Stewart et al., 2002) makes categorization decisions based only on perceptual and feedback information provided on the previous trial. With the exception of the immediately preceding item, the MAC model does not retain a representation of previously encountered exemplars.

Formal description of the MAC model. We present two versions of the MAC model. The first is the standard version of the model as outlined by Stewart et al. (2002). The second is a new version of the model, modified to incorporate error discounting via an annealing mechanism.

The MAC model makes categorization decisions by computing the psychological distance between the current stimulus and the previous one (this distance is isomorphic to the perceptual difference used in our decisional-recency analysis). The probability of repeating the same category response, P(same), is given by:

$$P(same\ category) = e^{-bd^2} \tag{4}$$

where d is the psychological distance between the current and previous stimulus, and b is a constant controlling the steepness of the Gaussian generalization gradient. Thus, b controls the rate at which psychological distance required to respond with a different category than the previous trial. Lower values of b will result in stimuli being perceived as more similar, thus increasing the probability that the response on the current trial will match the category feedback given on the previous trial.

The second, modified version of the MAC model includes an error-discounting mechanism, implemented by annealing the parameter b for trials following an error. Specifically, b was set to decrease across trials in this version, thus forcing the model to treat all items as if they were similar to the one just seen on the previous trial. Thus, the model effectively ignores the feedback from trials following errors.

To incorporate the error-discounting mechanism, we generated two forms of Equation 4. The first form of the equation was used on trials for which the response given on the previous trial was correct. The second version of the equation was used on trials following an error and instantiated annealing using the formalism provided in RASHNL

(Kruschke & Johansen, 1999). For this second form of the equation, the initial value of b was multiplied by an annealing factor, r, given by:

$$r(t) = 1/(1 + \rho \times t), \tag{5}$$

where ρ is a freely estimated, non-negative, scheduling parameter that controls the rate of annealing. Larger values of ρ lead to faster annealing, and when ρ is clamped at zero, the model exhibits no annealing at all (and saves a parameter in the process). This annealing mechanism has the effect of reducing the steepness of the Gaussian generalization gradient on the psychological distance between pairs of stimuli over trials. Thus, the mechanism gradually reduces the probability with which the model will shift its category response on a trial following an error. In summary, the standard MAC model has one parameter, the size parameter, b. The modified MAC model includes both the size parameter, b, and the annealing parameter, ρ.

MAC model simulations. The MAC model was fit in the same manner as the GCM. For the modified version, the ρ parameter was capped at 100. In order to closely model the participants' trial-by-trial behavior, we used each participant's individual responses, rather than the MAC model's own responses, to guide the feedback given to the model. That is, for each response, the MAC model computed its prediction based on the participant's response and feedback on the previous trial, rather than its own preceding response (use of the model's own response considerably worsened the model's fit). One of the participants for the late condition of Experiment 2 was not fit, as after removing trials with extreme reaction times, this participant did not have responses remaining for all items across all blocks. The best-fitting parameter values are presented in Table 2.

The MAC model captured the data qualitatively, but not quantitatively; the fits to Experiments 1 and 2 are shown in Figures 3 and 4, respectively. The slope increased

above chance early in training, and crossed through chance following the shift in reinforcement probabilities. However, the model's slopes fell short of participants' slopes and the model adapted to the shift much faster than did participants.

We do not consider the failure of the MAC model to quantitatively capture the data to be a core issue: The initial version of the MAC model was not designed to be a complete categorization model, but rather a means to explore perceptual recency effects in categorization Stewart et al. (2002). In that spirit, interest was on whether the model could better account for the data with the inclusion of an error-discounting mechanism.

RASHNL: Error Discounting via Annealing of Learning

Formal description of RASHNL. RASHNL (Kruschke & Johansen, 1999) is an exemplar-based connectionist model of probabilistic categorization that was developed as an extension of ALCOVE (Attention Learning COVEring map; Kruschke, 1992), which is itself an extension of the GCM (Nosofsky, 1986). Central to RASHNL is the concept of annealing of learning rates. This annealing mechanism captures error discounting by gradually decreasing the rate at which the model learns over time.

The annealed learning provides the model with various advantages over a fixed learning rate: In addition to the general notion that it allows for quick adaptation early in training, followed by slow fine-tuning of probabilistic responding (Amari, 1967; Heskes & Kappen, 1991; Murata, Kawanabe, Ziehe, Müller, & Amari, 2002), annealed learning may also help RASHNL avoid unduly high sensitivity to order among stimuli late in training, such as observed in ALCOVE (Lewandowsky, 1995). In confirmation, the limited tests of RASHNL available to date have consistently found that the annealing-based error-discounting mechanism improves the ability of the model to account for both probabilistic (Kruschke & Johansen, 1999) and non-probabilistic (Blair & Homa, 2005) categorization behavior.

RASHNL has a layer of input nodes that correspond to the dimensions of the stimulus. As RASHNL is an exemplar-based model, new items are categorized based on their similarity to previously encountered category members (Kruschke & Johansen, 1999; Nosofsky, 1986). Hence, the input nodes connect to a layer of hidden exemplar nodes, which correspond to the training stimuli. Activation of the jth exemplar node is given by:

$$h_j = \exp(-c \times |\psi_j - d|), \tag{6}$$

where c, the specificity, is a free parameter that determines the slope of the gradient of the receptive field of each exemplar; that is, the slope of the exponential decline in similarity with increasing distance between the current stimulus and the jth stored exemplar. (The present stimuli were uni-dimensional; accordingly, we removed RASHNL's gain activation, attention shifting, and attention-updating mechanisms, all of which only apply to multi-dimensional stimuli.)

Exemplar nodes connect to output nodes, which correspond to the available categories. Activation of the kth category node, a_k, is given by:

$$a_k = \sum_j w_{kj}\, h_j, \tag{7}$$

where w_{kj} is a weight associating an exemplar with a category. Category activations are mapped onto response probabilities using a version of Luce's (1963) choice rule, such that the probability of categorizing a stimulus into category K is determined by the exponentiated activation of category K over the sum of the exponentiated activation of all categories, given by:

$$P(K) = \exp(\varphi a_k) / \sum_i \exp(\varphi a_i), \tag{8}$$

where φ is a scaling parameter representing decisiveness. If φ is large, then a small activation advantage for category K will result in large preference for category K,

corresponding to maximizing behavior. Conversely, if φ is small, the response will be more uncertain and in proportion to the relative activations, thus corresponding to probability matching.

RASHNL is an error-driven learning model, such that each response is followed by feedback indicating the correct category in the form of teacher values for each category node. The error associated with a response is given by:

$$E = \frac{1}{2}\sum_{k}(t_k - a_k)^2, \tag{9}$$

where t is the teacher value, such that $t_k = 1$ if the stimulus is a member of category k, and $t_k = 0$ if the stimulus is not a member of category k.

Learning proceeds through the minimization of E via adjustment of association weights by gradient descent on error, given by:

$$\Delta w_{kj} = \lambda(t_k - a_k)h_j, \tag{10}$$

where λ represents the learning rate. Crucially for the current experiments, this learning rate is annealed, such that as training progresses the rate of learning is slowed. Although several annealing mechanisms have been explored within the neural network literature (Amari, 1967; Bös & Amari, 1998; Heskes & Kappen, 1991; Müller, Ziehe, Murata, & Amari, 1998; Murata et al., 2002), RASHNL uses a "search and converge" mechanism to decrease learning rates (see e.g., Darken & Moody, 1991). On each trial, t, the initial learning rate is multiplied by an annealing factor, r, as per the earlier Equation 5. The annealing function allows the model to make large shifts in learning early in training, whereas from around trial $1/\rho$ onward, the learning rates rapidly reduce and converge to zero. In addition to the annealing rate, ρ, this version of RASHNL had three other free parameters: specificity, c, the probability-mapping parameter, φ, and the weight-learning rate, λ.

RASHNL simulations. The model was applied to the data in the same manner as the GCM. RASHNL was fit using the models' own corrective feedback to guide model behavior. In both fits, we compared two versions of RASHNL: One in which the annealing parameter, ρ , was freely estimated and another one in which it was set to zero.

Figures 5 and 6 show the mean predictions of RASHNL when fit to the data of individual participants with annealing turned on for Experiments 1 and 2 respectively. Quite in contrast to the GCM and the MAC model, RASHNL captured the fast initial learning and slower post-shift learning that was displayed by participants. The corresponding mean (and median) best-fitting parameter values (aggregating across the fits to individual subjects) are shown in Table 3.

rATRIUM: Annealing Without Exemplars

The majority of rule-based models, including the GRT, do not include associative learning mechanisms. As an associative learning mechanism is particularly suited for investigating error discounting, we selected the rule module of ATRIUM (Erickson & Kruschke, 1998) as an alternative candidate model for the present data. ATRIUM's rule module learns to associate rules with particular categories via a standard network learning algorithm, permitting implementation of annealing in the same manner as in RASHNL.

ATRIUM is a hybrid model that relies on both exemplars and rules; here, we eliminated the exemplar module because we were exclusively interested in the generality of annealing and its applicability within a rule-based architecture (hence we use the label *r*ATRIUM for this variant of the model from here on). *r*ATRIUM divides the category space by a rule boundary set perpendicular to the relevant stimulus dimension. The stimulus dimension is represented by two rule nodes, r_{small} and r_{large}, whose activations are given by:

$$r_{small} = 1 - \frac{1}{1 + \exp[-\mu(d + \beta)]}, \tag{11}$$

and by:

$$r_{large} = \frac{1}{1 + \exp[-\mu(d + \beta)]}, \tag{12}$$

where d represents the value of a given item on the stimulus dimension. Each of these rule nodes forms a sigmoid threshold function, centered on the rule boundary, such that larger dimensional inputs will result in higher activation of the large rule node, while smaller dimensional inputs will result in a higher activation of the small rule node. The parameter μ represents the gain of the sigmoid (i.e., its steepness), and thus controls the level of perceptual noise (or its equivalent) as dimensional values approach the rule boundary. Large values of μ result in stimuli close to the rule boundary being more confusable. The parameter β controls the position of the rule boundary.

The rule nodes are connected to output nodes that correspond to the possible category selections. The activation of output nodes, a_k, for each category, k, is calculated as the sum of the activations of the small and large rule nodes, given by:

$$a_k = w_{k_{large}} r_{large} + w_{k_{small}} r_{small}, \tag{13}$$

where the activation is moderated by the learned association weights, w_k, between the rule and output nodes. As in RASHNL, the association weights are updated by minimizing mean square error during learning:

$$\Delta w_{kj} = \lambda(t_k - a_k) r_j, \tag{14}$$

where λ is a freely estimated parameter which controls the rate of learning. Finally, output activations are converted into probabilities as in Equation 8 in RASHNL.

For present purposes, the annealing mechanism from RASHNL, given in Equation 5, was imported into *r*ATRIUM: Thus, an annealing rate, ρ, controlled the rate at which the weight learning rate, λ, was adjusted on successive trials.

In summary, the model has four free parameters: A gain constant, μ, which sets the standard deviation of the perceptual noise; a scaling constant, φ, which maps output probabilities to participant responses; the annealing rate, ρ; and learning rate, λ.

The model was fit in the same manor as the fits with RASHNL. Figures 7 and 8 show the fits of *r*ATRIUM to Experiment 1 and Experiment 2 respectively. Like RASHNL, *r*ATRIUM provided a good fit to the data. As shown in the figures, the model closely tracked the behavior of the participants throughout training.

Fit Statistics

In Craig et al. (in press), a series of fit statistics were used to compare the fits of the four models, viz. AIC (Akaike Information Criterion; Akaike, 1974) and w_i(AIC) (AIC weights; Wagenmakers & Farrell, 2004), and BIC (Bayesian Information Criterion) and w_i(BIC) (BIC weights; Wagenmakers & Farrell, 2004) to compare the models. AIC and BIC adjust for model complexity and flexibility. When corrected for small sample sizes, the AIC (AIC_c) is given by:

$$AIC_c = -2 \ln L + 2V + \frac{2V(V+1)}{n - V - 1}, \tag{15}$$

where L is the maximum likelihood for the given model with V free parameters taken over n observations. The corrected AIC is recommended for use of samples where the ratio of data points to parameters is less than 40. The AIC thus combines two sources of information: Lack of fit (represented by the log likelihood) and a penalty term for model complexity (represented by the second and third terms in the above equation). The BIC is given by:

$$BIC = -2 \ln L + V \ln n. \tag{16}$$

Unlike AIC, whose penalty relies on the number of parameters only, the BIC additionally penalizes models based on the number of data points being fitted.

AIC_c and BIC values were converted into AIC and BIC weights (Wagenmakers & Farrell, 2004). The w_i(AIC), and w_i(BIC), represent the conditional probabilities that the model M_i is the best of the set of models being compared.

In Craig et al. (in press), a likelihood-ratio test (Lamberts, 1997) was used to determine whether the loss of fit associated with removal of error discounting, by setting ρ to zero, was statistically significant. The likelihood-ratio test is given by:

$$\chi^2 = -2[\ln L(restricted) - \ln L(general)], \tag{17}$$

where $\ln L(general)$ is the log-likelihood of the version of a model that includes annealing (i.e., $\rho > 0$), whereas $\ln L(restricted)$ is the log-likelihood of the restricted version of a model, with annealing set to zero.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Amari, S. (1967). Theory of adaptive pattern classifiers. *IEEE Transcriptions in Electronic Computers*, *16*, 299–307.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Blair, M., & Homa, D. L. (2005). Integrating novel dimensions to eliminate category exceptions: When more is less. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 258–271.

Bös, S., & Amari, S. (1998). On-line learning in switching and drifting environments with application to blind sources separation. In D. Saad (Ed.), *On-line learning in neural networks* (pp. 209–230). Cambridge: Cambridge University Press.

Colreavy, E., & Lewandowsky, S. (2008). Strategy development and learning differences in supervised and unsupervised categorization. *Memory & Cognition*, *36*, 762–775.

Craig, S., Lewandowsky, S., & Little, D. R. (in press). Error discounting in probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Darken, C., & Moody, J. E. (1991). Note on learning rate schedules for stochastic optimization. In R. P. Lippman, J. E. Moody, & D. S. Tourestzky (Eds.), *Advances in neural information processing systems* (pp. 832–838). San Mateo, CA: Kaufmann.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Heskes, T. M., & Kappen, B. (1991). Learning process in neural networks. *Physical Review A*, *44*, 2718–2726.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083–1119.

Lamberts, K. (1997). Process models of categorization. In K. Lamberts & D. Shanks. (Eds.), *Knowledge, concepts, and categories* (pp. 371–403). Hove, East Sussex: Psychology Press.

Lewandowsky, S. (1995). Base-rate neglect in ALCOVE: A critical reevaluation. *Psychological Review*, *102*, 185–191.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Müller, K.-R., Ziehe, A., Murata, N., & Amari, S. (1998). On-line learning in switching and drifting environments with application to blind source separation. In D. Saad (Ed.), *On-line learning in neural networks* (pp. 93–110). Cambridge: Cambridge University Press.

Murata, N., Kawanabe, M., Ziehe, A., Müller, K.-R., & Amari, S. (2002). On-line learning in changing environments with applications to supervised and unsupervised learning. *Neural Networks*, *15*, 743–760.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308–313.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*, 375–402.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 3–11.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

Author Note

Preparation of this paper was facilitated by a Discovery Grant from the Australian Research Council and an Australian Professorial Fellowship to the second author. The first and third authors were supported by Jean Rogerson Scholarships. Additionally, the third author was supported by an Australian Postgraduate Scholarship and an NIH-NIMH training grant #:T32 MH019879-14. We wish to thanks Charles Hanich for assisting with data collection and Gordon Brown for his comments on an earlier version of this manuscript. Address correspondence to the second author at the School of Psychology, University of Western Australia, Crawley, W.A. 6009, Australia. Electronic mail may be sent to lewan@psy.uwa.edu.au. Web page: http://www.cogsciwa.com.

Table 1

Median (Mdn), mean (M), and standard deviation (SD) of estimated parameter values across participants and negative log-likelihood values (-lnL) for the fits with GCM for each condition of both experiments.

Exp.	Cond.	Parameters						
		γ			c			$-\ln L$
		Mdn	*M*	*SD*	*Mdn*	*M*	*SD*	
1	Early	2.55	2.37	1.86	1.42	13.20	20.76	5939.48
	Mid	1.07	1.11	0.71	2.90	9.87	10.96	6163.92
	Late	1.19	1.74	1.34	1.43	5.46	8.81	5924.28
2	Early	1.55	1.57	1.04	2.09	9.23	11.30	5321.91
	Late	1.50	1.67	0.87	2.10	6.25	9.92	3946.33

Table 2

Median (Mdn), mean (M), and standard deviation (SD) of estimated parameter values across participants and negative log-likelihood values (-lnL) for the fits of the MAC model with annealing on and off for each condition of both experiments.

Exp.	Cond.	Parameters						
		ρ			b			$-\ln L$
		Mdn	*M*	*SD*	*Mdn*	*M*	*SD*	
1	Early	0.194	7.795	25.704	0.43	0.41	0.22	6425.07
	Mid	0.414	36.089	49.447	0.45	0.47	0.18	5963.39
	Late	0.497	14.751	36.122	0.39	0.54	0.51	6182.40
	Early	0.	0.	0.	0.35	0.35	0.23	6663.88
	Mid	0.	0.	0.	0.44	0.50	0.40	6171.74
	Late	0.	0.	0.	0.34	0.41	0.22	6420.13
2	Early	0.812	41.810	51.368	0.47	0.43	0.19	5337.18
	Late	1.004	13.287	32.690	0.46	0.51	0.13	3693.88
	Early	0.	0.	0.	0.38	0.44	0.35	5647.71
	Late	0.	0.	0.	0.40	0.44	0.15	3972.02

Table 3

Median (Mdn), mean (M), and standard deviation (SD) of estimated parameter values across participants and negative log-likelihood values (-lnL) for the fits with RASHNL with annealing free or set to zero for each condition of both experiments.

Exp.	Cond.	Parameters												
		ρ			φ			λ			c			$-\ln L$
		Mdn	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	
1	Early	0.019	0.457	1.586	3.89	4.88	3.88	0.26	1.46	3.21	1.17	5.39	15.43	5688.31
	Mid	0.006	0.024	0.030	4.13	3.95	2.14	0.23	0.55	0.72	0.98	2.97	7.52	5303.15
	Late	0.032	0.054	0.088	2.37	3.03	1.79	0.78	1.66	2.55	1.57	8.48	13.19	5696.02
	Early	0.	0.	0.	4.05	10.91	24.97	0.08	0.35	0.58	1.42	1.42	0.91	5729.54
	Mid	0.	0.	0.	4.09	4.33	3.04	0.17	0.39	0.50	0.82	3.00	8.30	5442.77
	Late	0.	0.	0.	2.73	3.36	2.16	0.09	0.44	0.61	1.38	5.62	11.18	5764.01
2	Early	0.035	0.285	0.748	3.14	3.50	2.61	0.39	2.71	7.08	0.79	18.43	55.05	4937.05
	Mid	0.025	0.038	0.040	4.44	7.10	8.53	0.30	0.76	0.90	0.76	1.05	0.82	3632.29
	Early	0.	0.	0.	3.70	6.03	7.29	0.07	0.47	0.63	1.17	6.38	12.44	5115.43
	Late	0.	0.	0.	5.21	8.68	9.32	0.03	0.08	0.11	0.91	4.16	10.10	3722.47

Table 4

Median (Mdn), mean (M), and standard deviation (SD) of estimated parameter values across participants and negative log-likelihood values (-lnL) for the fits with r*ATRIUM with annealing free or set to zero for each condition of both experiments.*

Exp.	Cond.	Parameters												
		ρ			φ			λ			μ			$-\ln L$
		Mdn	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	
1	Early	0.041	0.092	0.142	1.75	1.98	1.83	3.41	3.69	3.18	0.93	1.02	0.43	5769.22
	Mid	0.002	0.375	1.181	1.66	1.96	1.26	0.88	3.74	6.72	0.94	1.10	0.70	5375.43
	Late	0.011	0.049	0.118	1.27	8.65	26.31	1.20	2.06	2.25	0.74	0.95	0.65	5754.83
	Early	0.	0.	0.	1.76	2.08	1.80	0.41	0.64	0.83	1.01	1.07	0.66	5892.26
	Mid	0.	0.	0.	2.55	2.29	1.41	0.26	0.48	0.46	0.92	0.97	0.34	5484.35
	Late	0.	0.	0.	1.28	1.66	1.18	0.27	0.38	0.49	0.80	0.94	0.53	5822.08
2	Early	0.039	0.089	0.101	1.20	1.53	1.70	1.28	3.68	3.69	0.88	0.85	0.26	5010.39
	Late	0.019	0.126	0.246	1.87	2.65	2.06	1.30	4.69	8.09	0.83	0.82	0.29	3727.13
	Early	0.	0.	0.	1.71	2.25	2.22	0.46	0.52	0.51	0.90	0.96	0.56	5197.12
	Late	0.	0.	0.	2.86	3.27	2.44	0.22	0.47	0.54	0.79	0.79	0.31	3897.96

Figure Captions

Figure 1. Observed slopes through response probabilities across all 4 training items in Experiment 1 (solid lines and error bars), slopes predicted by GCM (solid lines and open circles), and objective slopes (dotted lines). Error bars indicate 95% confidence intervals. The three panels show the early (top), mid (middle), and late (bottom) condition, respectively.

Figure 2. Observed slopes through response probabilities across all 4 training items in Experiment 2 (solid lines and error bars), slopes predicted by GCM (solid lines and open circles), and objective slopes (dotted lines). Error bars indicate 95% confidence intervals. The two panels show the early (top) and late (bottom) condition, respectively.

Figure 3. Observed slopes through response probabilities across all 4 training items in Experiment 1 (solid lines and error bars), slopes predicted by the MAC model (solid lines and open circles), and objective slopes (dotted lines). Error bars indicate 95% confidence intervals. The three panels show the early (top), mid (middle) and late (bottom) condition, respectively.

Figure 4. Observed slopes through response probabilities across all 4 training items in Experiment 2 (solid lines and error bars), slopes predicted by the MAC model (solid lines and open circles), and objective slopes (dotted lines). Error bars indicate 95% confidence intervals. The two panels show the early (top) and late (bottom) condition, respectively.

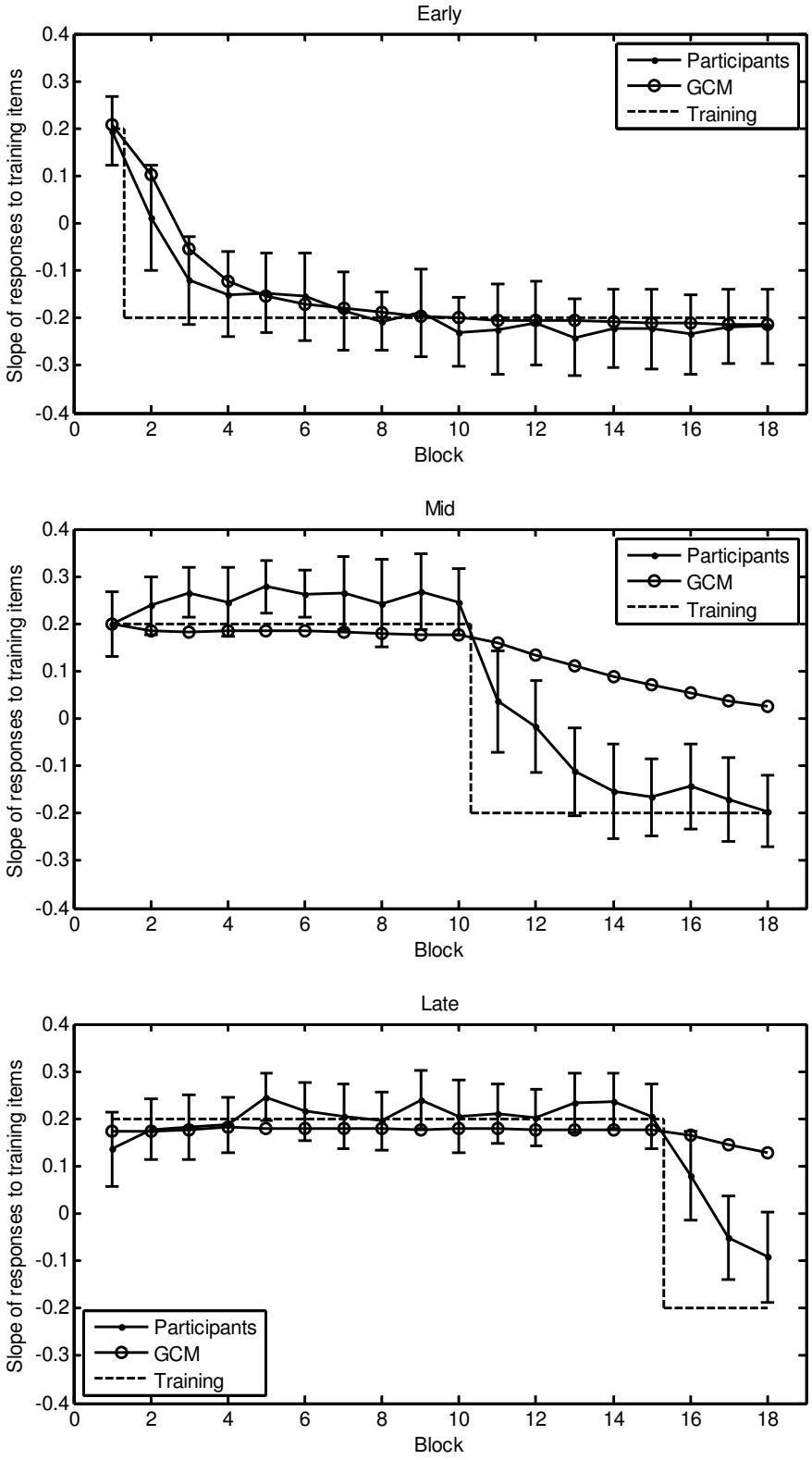
Figure 5. Observed slopes through response probabilities across all 4 training items in Experiment 1 (solid lines and error bars), slopes predicted by RASHNL (solid lines and open circles), and objective slopes (dotted lines). RASHNL's predictions were obtained with the annealing parameter, ρ, being freely estimated. Error bars indicate 95%

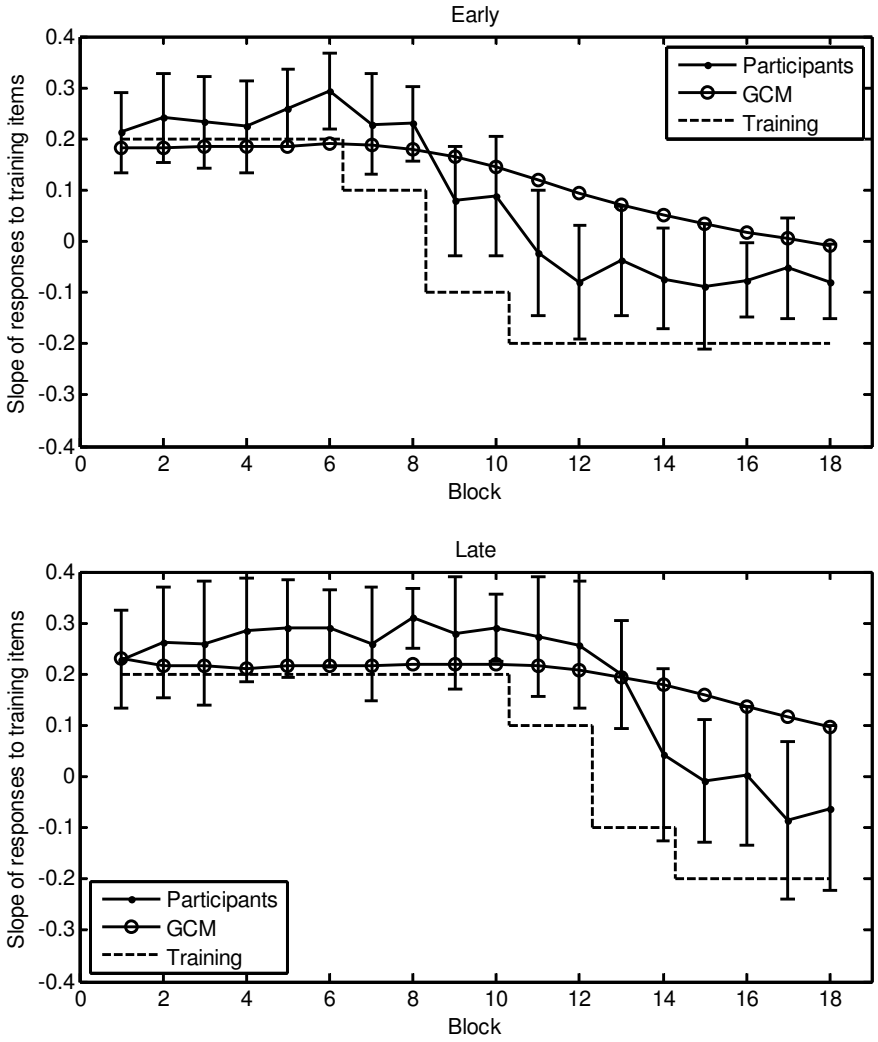
confidence intervals. The three panels show the early (top), mid (middle), and late (bottom) condition, respectively.

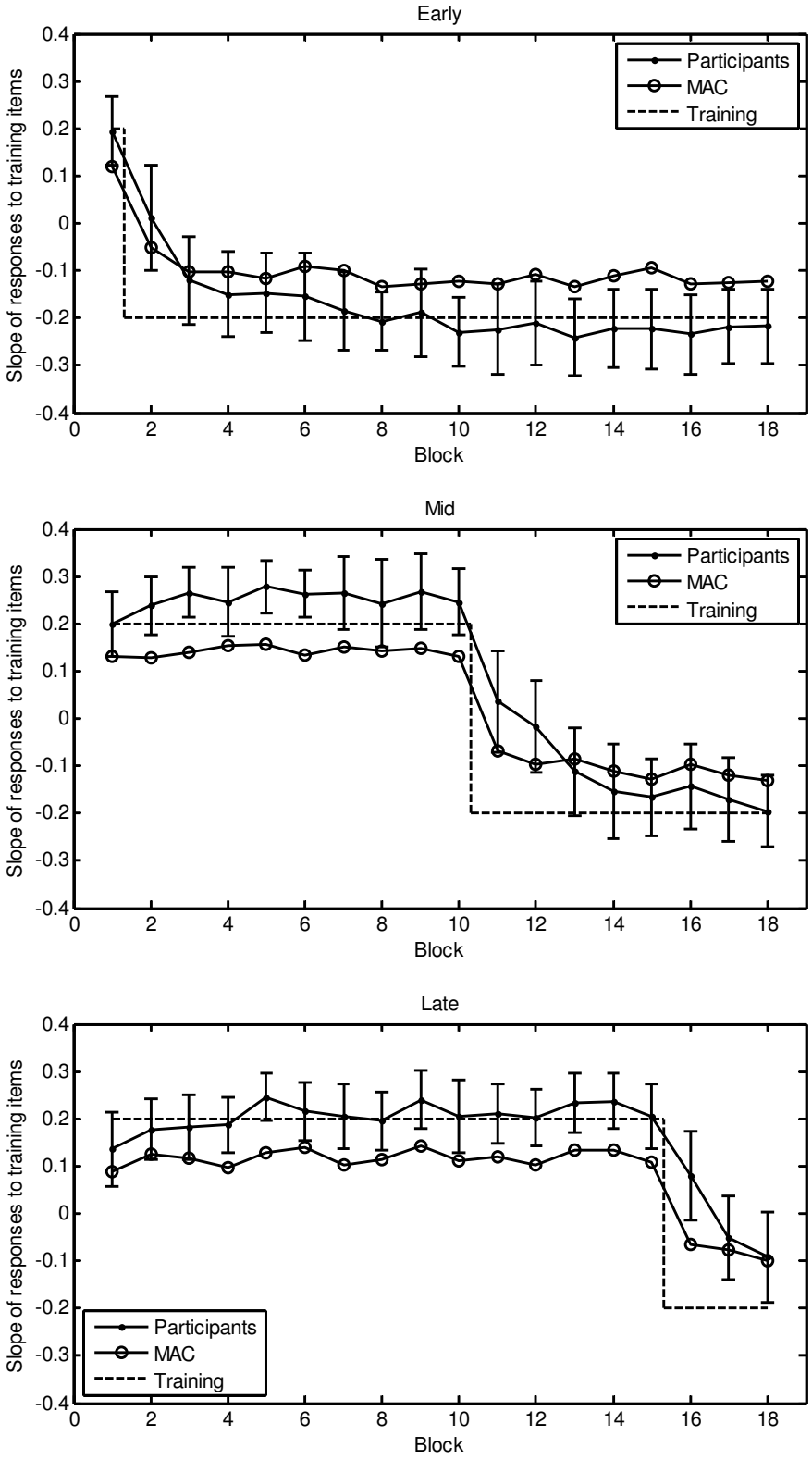
Figure 6. Observed slopes through response probabilities across all 4 training items in Experiment 2 (solid lines and error bars), slopes predicted by RASHNL (solid lines and open circles), and objective slopes (dotted lines). RASHNL's predictions were obtained with the annealing parameter, ρ, being freely estimated. Error bars indicate 95% confidence intervals. The two panels show the early (top) and late (bottom) condition, respectively.

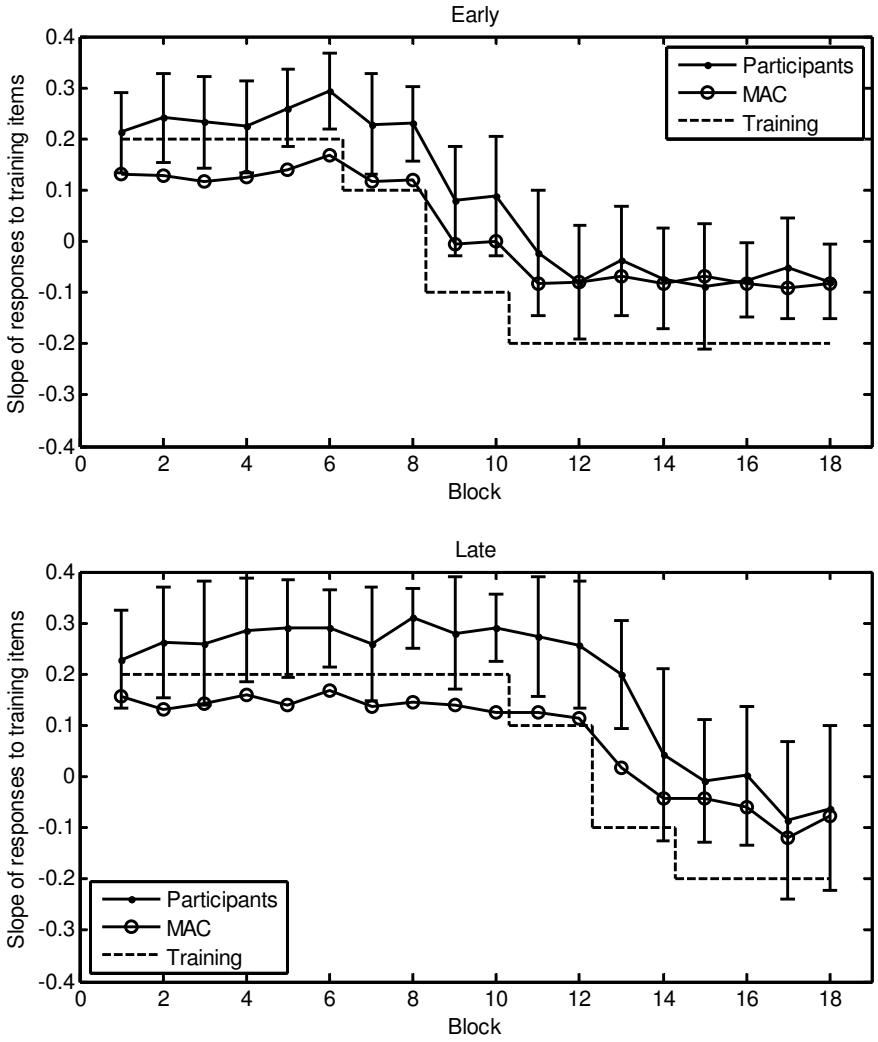
Figure 7. Observed slopes through response probabilities across all 4 training items in Experiment 1 (solid lines and error bars), slopes predicted by rATRIUM (solid lines and open circles), and objective slopes (dotted lines). rATRIUM's predictions were obtained with the annealing parameter, ρ, being freely estimated. Error bars indicate 95% confidence intervals. The three panels show the early (top), mid (middle), and late (bottom) condition, respectively.

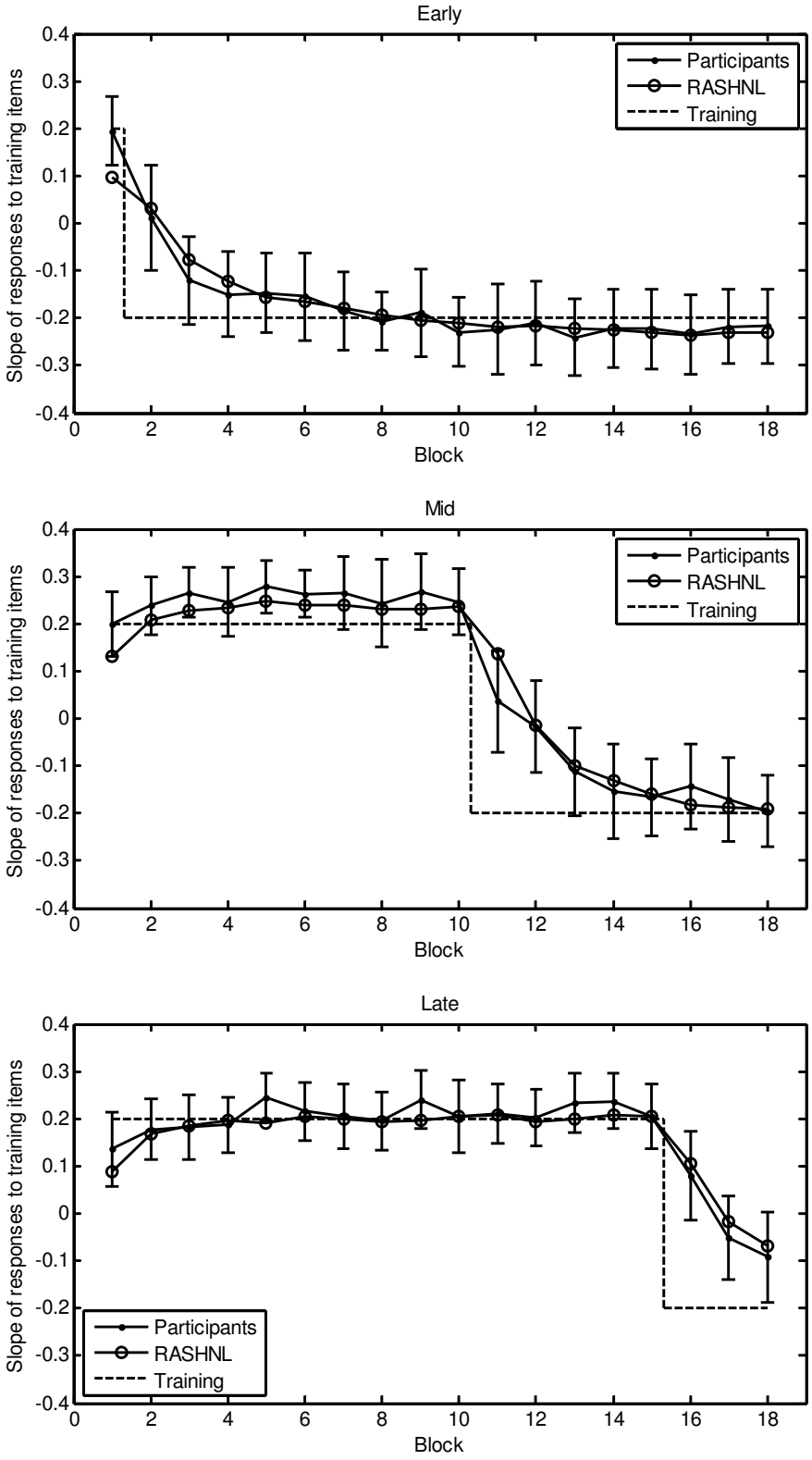
Figure 8. Observed slopes through response probabilities across all 4 training items in Experiment 2 (solid lines and error bars), slopes predicted by rATRIUM (solid lines and open circles), and objective slopes (dotted lines). rATRIUM's predictions were obtained with the annealing parameter, ρ, being freely estimated. Error bars indicate 95% confidence intervals. The two panels show the early (top) and late (bottom) condition, respectively.

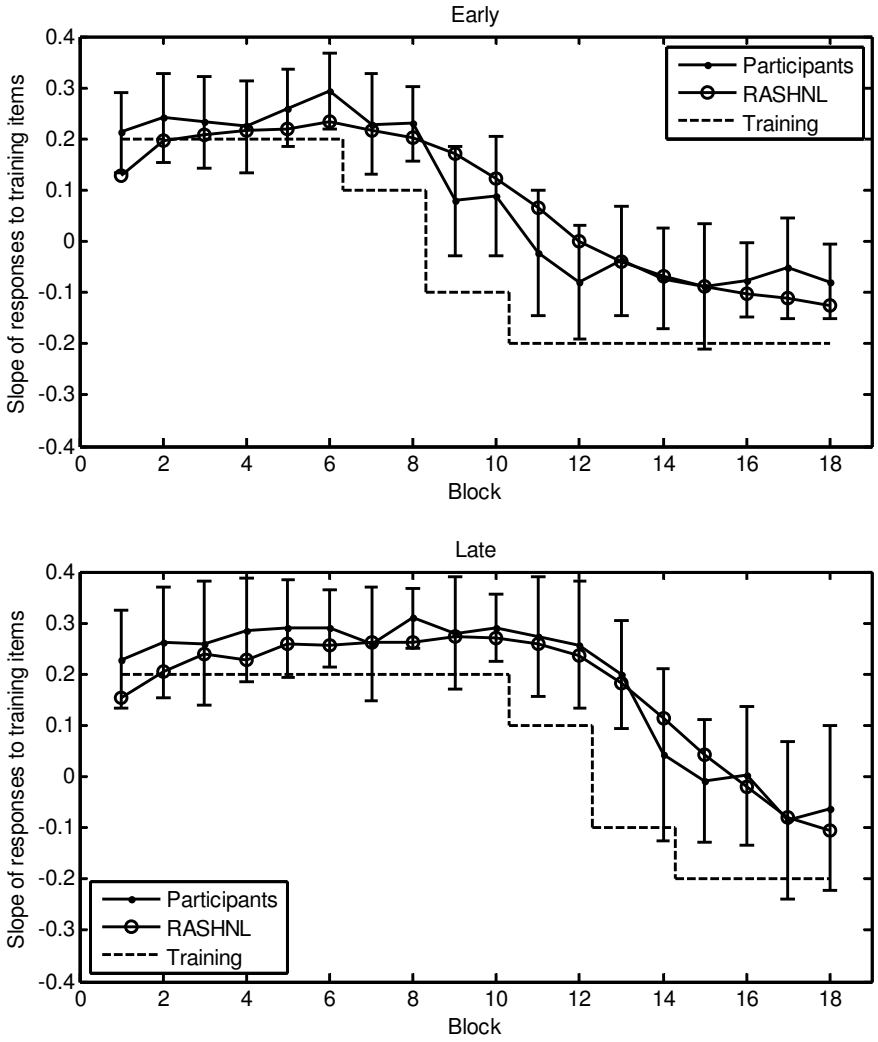

Early
Mid
Late
Participants
GCM
Training
Slope of responses to training items
Block
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
2
4
6
8
10
12
14
16
18

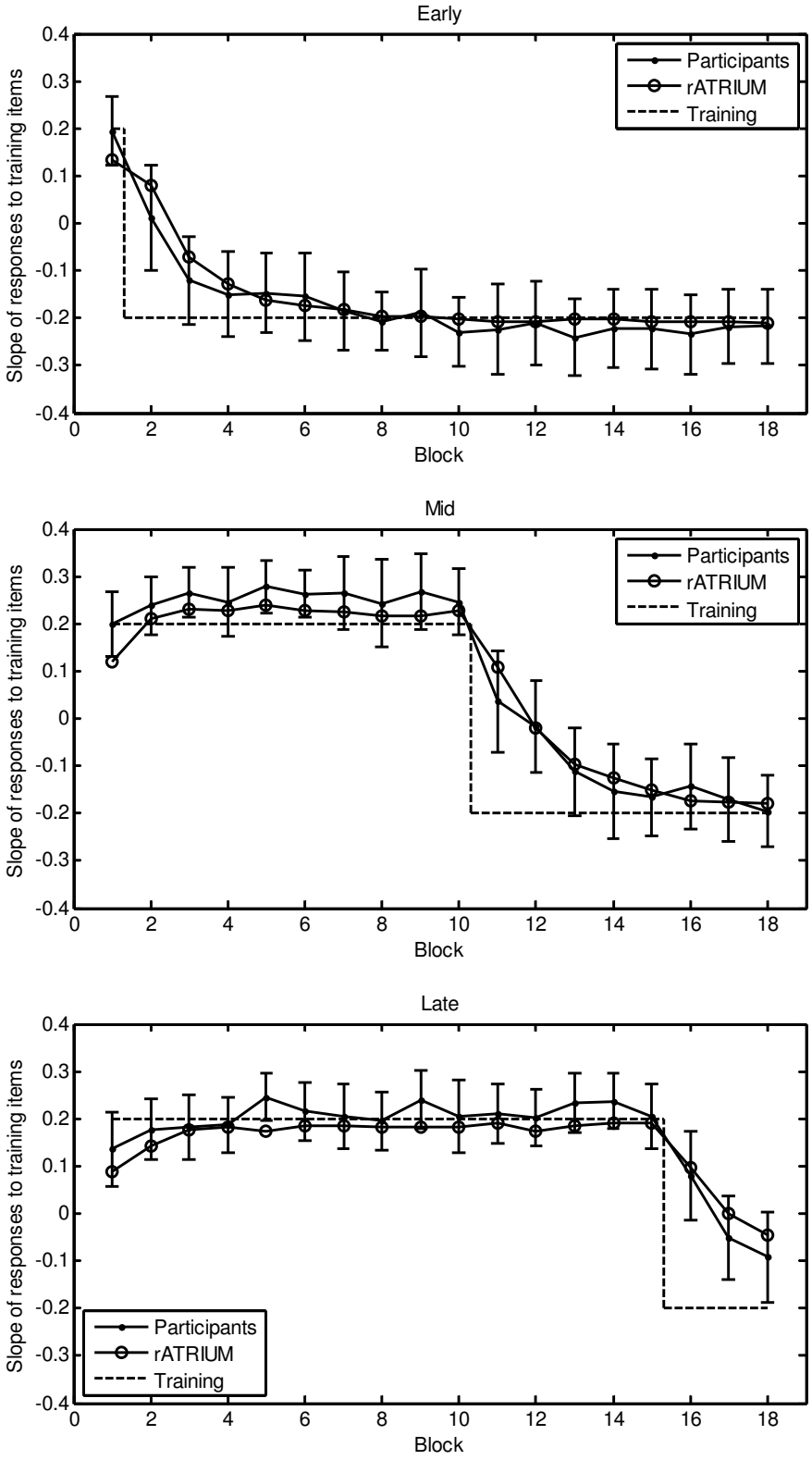

Early
Slope of responses to training items
Block
Participants
GCM
Training
Late
Slope of responses to training items
Block
Participants
GCM
Training

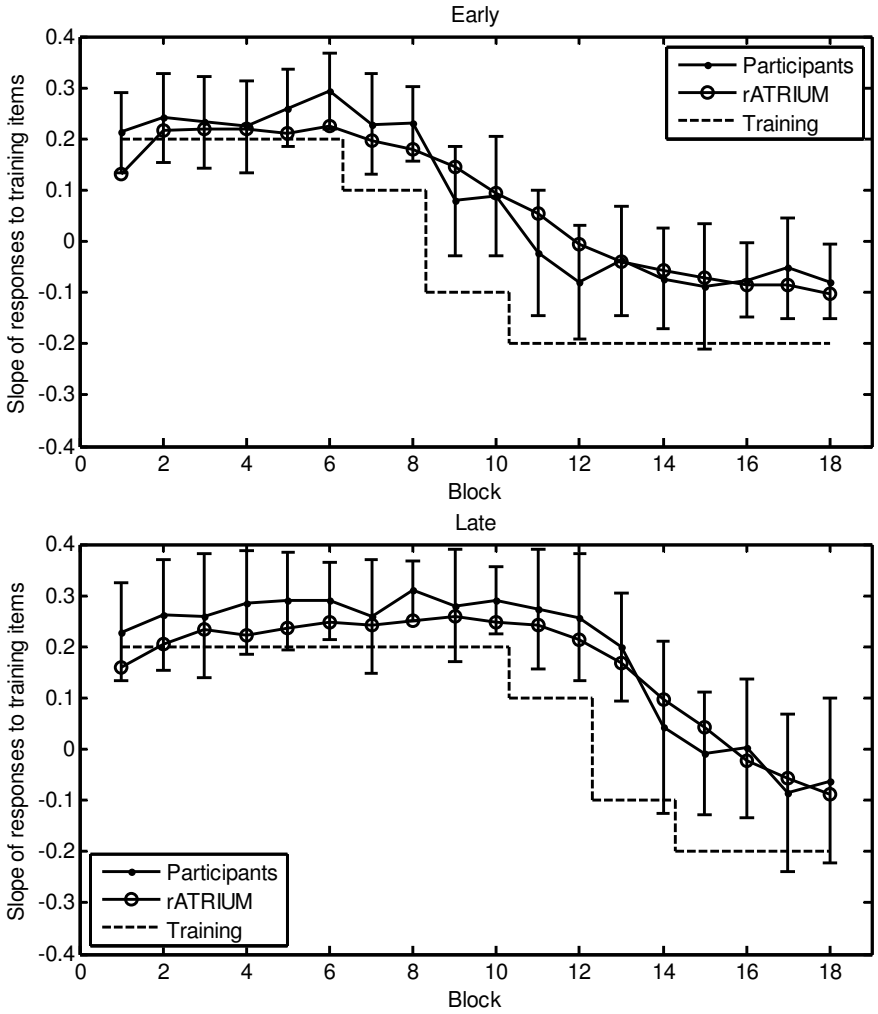

Early
Participants
MAC
Training
Slope of responses to training items
Block
Mid
Participants
MAC
Training
Slope of responses to training items
Block
Late
Participants
MAC
Training
Slope of responses to training items
Block


Early
Slope of responses to training items
Block
Participants
MAC
Training
Late
Slope of responses to training items
Block
Participants
MAC
Training


Early
Mid
Late
Participants
RASHNL
Training
Slope of responses to training items
Block
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
2
4
6
8
10
12
14
16
18


Early
Slope of responses to training items
Block
Participants
RASHNL
Training
Late
Slope of responses to training items
Block
Participants
RASHNL
Training


Early
Mid
Late
Participants
rATRIUM
Training
Slope of responses to training items
Block
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
2
4
6
8
10
12
14
16
18


Early
Slope of responses to training items
Block
Participants
rATRIUM
Training
Late
Slope of responses to training items
Block
Participants
rATRIUM
Training