Graduate School of Decision Sciences

Universität Konstanz

GSDS
Working Paper
No. 2017-06

Testing learning mechanisms of rule-based judgment

Janina A. Hoffmann
Bettina von Helversen
Jörg Rieskamp

February 2017

Graduate School of Decision Sciences

All processes within our society are based on decisions – whether they are individual or collective decisions. Understanding how these decisions are made will provide the tools with which we can address the root causes of social science issues.

The GSDS offers an open and communicative academic environment for doctoral researchers who deal with issues of decision making and their application to important social science problems. It combines the perspectives of the various social science disciplines for a comprehensive understanding of human decision behavior and its economic and political consequences.

The GSDS primarily focuses on economics, political science and psychology, but also encompasses the complementary disciplines computer science, sociology and statistics. The GSDS is structured around four interdisciplinary research areas: (A) Behavioural Decision Making, (B) Intertemporal Choice and Markets, (C) Political Decisions and Institutions and (D) Information Processing and Statistical Analysis.

GSDS – Graduate School of Decision Sciences
University of Konstanz
Box 146
78457 Konstanz

Phone: +49 (0)7531 88 3633
Fax: +49 (0)7531 88 5193
E-mail: gsds.office@uni-konstanz.de

-gsds.uni-konstanz.de

ISSN: 2365-4120
February 2017

Testing learning mechanisms of rule-based judgment

Janina A. Hoffmann

University of Konstanz

Bettina von Helversen

University of Zürich

Jörg Rieskamp

University of Basel

Author Note

This research was supported by Swiss National Science Foundation Grant 100014 _146169/1.

Correspondence concerning this article should be addressed to Janina A. Hoffmann, Department of Psychology, University of Konstanz, Universitätsstrasse 10, 78 468 Konstanz, Germany. E-mail: janina.hoffmann@uni-konstanz.de

Abstract

Weighing the importance of different pieces of information is a key determinant of making accurate judgments. In social judgment theory, these weighting processes have been successfully modeled with linear models. How people learn to make judgments has received less attention. Although the hitherto proposed least mean squares or delta learning rule can perfectly learn to solve linear problems, we found in a first study that it does not adequately describe human learning. To provide a more accurate description of learning processes we amended the delta learning rule with three learning mechanisms —a decay, an attentional learning mechanism, and a capacity limitation —and tested in a further study how well those learning mechanisms can describe and predict learning in linear judgment tasks.

In the study, participants first learned to predict a continuous criterion based on four cues. To test the three learning mechanisms rigorously against each other, we changed the importance of the cues after 200 trials so that the mechanisms make different predictions with regard to how fast people adapt to the new environment. On average, judgment accuracy improved from trial 1 to 200, dropped when the task structure changed, but improved again until the end of the task. The capacity-restricted learning model best described and predicted the learning curve of the majority of participants. Taken together, these results suggest that human learning when making inferences is governed by cognitive capacity limitations.

Keywords: Multiple-cue Judgment; Rule-based Processes; Learning

Testing learning mechanisms of rule-based judgment

When making judgments, such as predicting a job candidate's future performance or assessing the value of a used car, people usually rely on information about the object of interest, such as the job candidate's skills or the car's mileage and accident records. An important predictor for judgment accuracy is the ability to correctly weigh the available aspects according to their importance. For instance, a car's mileage may accurately predict for how long the car will still run, whereas the time since its last cleaning may be less important. Social judgment theory has proposed that the weight people assign to different pieces of information (or cues) when making a judgment can be estimated by linear regression models —following the assumption that judgments are formed by weighting and then combining the cue values linear additively (e.g., Brehmer, 1994; Cooksey, 1996). In the following decades, social judgment theory has been successfully employed to understand which aspects people consider in judgment and decision problems in a range of applied areas, such as personality judgments (Hirschmüller, Egloff, Nestler, & Back, 2013), sentencing decisions (von Helversen & Rieskamp, 2009), personal selection (Graves & Karren, 1992), or medical diagnoses (Wigton, 1996). Furthermore, the notion that people preferably weigh and add information has inspired theories of information processing across a variety of domains ranging from probability judgments (Nilsson, Winman, Juslin, & Hansson, 2009) to impression formation (Anderson, 1971; Fishbein & Ajzen, 1975).

Despite the success of linear, additive models in describing how people combine different pieces of information (i.e. cues) when making judgments, our knowledge about how people learn to infer each cues' importance is still limited. Previous research has proposed that the additive integration of weighted information emerges from a serial, capacity-constrained hypothesis-testing process restricting people to consider only linear, additive rules (Juslin, Karlsson, & Olsson, 2008). Yet, the psychological mechanisms that may limit this rule-based learning process have rarely been spelled out (but see Kelley & Busemeyer, 2008; Kelley & Friedman, 2002; Rolison, Evans, Dennis, & Walsh, 2012;

Speekenbrink & Shanks, 2010) and learning models incorporating these constraints have seldom been tested against each other. The goal of the current research was to fill this gap and to investigate how the learning of cue weights in linear judgment problems can be described. To this goal we examined how a simple and widely used learning rule (the least mean squares or delta-learning rule) can be extended with different psychological mechanisms to explain how people learn the importance of cues in multiple-cue judgment tasks.

In the following we give an overview on how people weight information in multiple-cue judgments and review the least mean squares rule as a model describing the learning process as well as how it deviates from human learning. Next, we extend this learning rule by different psychological mechanisms to capture human performance and test these psychological mechanisms against each other in two studies.

Rule-based models of human judgment

Social judgment theory (SJT) has proposed that people approach judgment problems such as assessing the selling price of a car by considering the different aspects that could affect the car's worth, weighting them by their importance, and summing up the weighted cue values. This idea has been formalized by portraying a persons' judgments $\hat{j}$ as a linear, additive function of the cue values x_i weighted by their importance, the cue weights w_i, which can be mathematically modeled by a linear regression.

$$\hat{j} = \sum_i w_i \cdot x_i \tag{1}$$

with $x_i = \begin{bmatrix} x_1 & \dots & x_n & 1 \end{bmatrix}$ where n denotes the number of cues and 1 denotes the constant intercept. Accordingly, these *rule-based models* assume that people abstract the importance of each cue and prescribe how the abstracted knowledge should be combined.

Linear additive rules capture human judgments very well in applied settings but also in experimental studies in which people learn to solve new judgment tasks (for a review see

Karelaia & Hogarth, 2008). In particular, linear rules describe judgments well in tasks in which the criterion is a linear, additive function of the cues (Hoffmann, von Helversen, & Rieskamp, 2016; Juslin et al., 2008; Scheibehenne, von Helversen, & Rieskamp, 2015). In addition, the cue weights implied by linear rules have been found to successfully predict participants' judgments for unknown objects (Hoffmann, von Helversen, & Rieskamp, 2014) and correspond well with people's explicit judgment rules (Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Lagnado, Newell, Kahan, & Shanks, 2006; Speekenbrink & Shanks, 2010). Furthermore, when considering a task in which participants learn the correct weights of cues over repeated trials with feedback, it has been shown that the cue weights estimated from a rolling regression —a series of linear regressions fitted to a fixed set (or window) of training trials and repeatedly moved one trial ahead —match people's stated importance of each cue across the learning phase (Lagnado et al., 2006). However, although the rolling regression provides insights into the question of how the importance people assign to different cues changes over time, this descriptive model is mute about the cognitive learning processes underlying changes in cue importance. Attempts to model these learning processes mathematically have predominantly relied on the least mean squares rule to adjust the cue weights over trials (Gluck & Bower, 1988; Kelley & Busemeyer, 2008; Kelley & Friedman, 2002; Rolison et al., 2012).

The Least Mean Squares (LMS) rule

Learning the importance of each cue requires repeatedly updating the cue weights based on feedback about the correct criterion. It has been suggested that people update these cue weights by comparing two successively presented objects and relating the difference in judgment criteria to the difference in cue values (Juslin et al., 2008; Pachur & Olsson, 2012). This trial-by-trial updating process is mathematically captured by the delta-learning or "LMS rule" (called "LMS rule" because it converges to the least mean squares, LMS, solution Gluck & Bower, 1988; Sutton & Barto, 1981). In each trial, the

judgment is made based on a linear regression model (Equation 1). After each trial, the cue weights are updated for the next trial depending on how much the judgment $\hat{j}$ deviated from feedback y. The more the judgment deviated from the correct judgment and the higher the learning rate λ is, the more strongly the cue weights should change in the next trial Δw_i. Changes in the cue weights are attributed to those cues with higher cue values.

$$\Delta w_i = \lambda \cdot x_i \cdot (y - \hat{j}) \tag{2}$$

At the end of each trial the cue weights are updated with their associated changes.

$$w_i = w_i + \Delta w_i \tag{3}$$

The LMS rule is identical to the Rescorla-Wagner learning model (Rescorla & Wagner, 1972; Sutton & Barto, 1981) and has been applied to describe conditioning (Siegel & Allan, 1996), category learning (Gluck & Bower, 1988; Shanks, 1991), learning in multiple-cue judgment (Kelley & Busemeyer, 2008), and reward-related learning in neuroscience (O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Schultz & Dickinson, 2000; Tobler, O'doherty, Dolan, & Schultz, 2006). In a first step, we aimed to evaluate how well the LMS rule can describe participants' judgments over the course of learning and then investigated whether extending it with psychologically informed mechanisms can improve the prediction of human learning.

Reanalysis: Comparing the LMS rule to a rolling regression

To investigate this question, we compared the performance of the LMS rule to a rolling regression model. The rolling regression can be used as a measurement model to detect which cues people apply and how the cue weights change over time (Kelley & Friedman, 2002; Lagnado et al., 2006). In a rolling regression, a linear regression model is repeatedly estimated for a fixed number of judgments starting from the first to the n^{th} learning trial (where n indicates the window size) and this window is then repeatedly

moved by one trial ahead until it includes the last learning trial. For instance, using a window size of 50 trials the rolling regression is estimated in the first step using trial 1 to 50, next using trial 2 to 51, and so forth. With sufficiently small window sizes the rolling regression can reflect any kind of changes in cue weights that occur during the learning phase and thus its goodness of fit provides an upper limit for the fit of any learning model of the cue weights. We also estimated a baseline model as a lower limit any learning model has to beat, which simply learns participant's mean judgment over the learning trials. To evaluate the performance of the LMS rule against the rolling regression and the baseline model, we reanalyzed the linear judgment task from Hoffmann et al. (2014). In this study, a linear regression model described participant's judgment well at the end of training and also best predicted judgments for new objects for the majority of participants compared to an exemplar model.

Judgment task. In the judgment task, the criterion value ranging from 0 to 50 was perfectly predicted by four quantitative cues that could take values from 0 to 5. The criterion value was a linear, additive function of these cues, $y = 4x_1 + 3x_2 + 2x_3 + x_4$. Participants learned to predict the judgment values of 25 objects over 10 blocks with items presented in random order in each block, resulting in 250 training trials. In each trial, participants were asked to make a judgment and afterwards received feedback about the correct outcome. After 250 trials, participants moved to a test phase in which they judged 15 unknown objects four times.

Comparison to a rolling regression. We used a rolling regression with a fixed window of 50 trials and calculated the RMSD between its prediction for the last trial in a window (hence trial 50, 51...) and participants' judgment for this trial from trial 50 to 250 in the training phase. The last window of the rolling regression is akin to a linear regression fitted to the last 50 training trials. For the LMS rule we assumed that at the beginning of the task all cues have starting weights of zero, but that participants have a starting bias corresponding to the intercept in a linear regression. This starting bias was

set to the participants' judgment in the first trial. The models' learning rate and standard deviation were estimated by minimizing the maximum likelihood between participants' judgments and model predictions over all trials in the training phase (for details on model estimation and comparison see Appendix A, for model parameters see Appendix B). Based on the cue weights in each trial we then calculated the RMSD between model predictions and participants' judgments from trial 50 to 250 as well as the RMSD for the last block of training (from trial 226 to 250). We only considered trials from trial 50 onwards to enable a better comparison with the rolling regression.

Table 1 summarizes the model fits, that is to what degree model predictions deviate from participants' judgments. Considering all training trials, the LMS rule outperformed the baseline model, V = 12775 (paired Wilcoxon test), $p < 0.001$, 95% CI [-1.2,-0.6], but did not meet the performance of the rolling regression model, V = 41328, $p < 0.001$, 95% CI [3.1,3.7]. More importantly, the average RMSD of the LMS rule was almost twice as high as the average RMSD of the rolling regression and close to the average RMSD for the baseline model. To more closely track the learning path in the training phase, we compared the cue weights estimated from trial 51 to trial 250 for the LMS rule and the rolling regression (Figure 1). For the most important cue, cue 1, the rolling regression and the LMS rule propose a similar learning path, but the LMS rule systematically underestimates the importance participants gave to all other cues. Furthermore, the LMS rule suggests a slow, but steady learning of cue 2 and cue 3, whereas the rolling regression weights suggest that people only update the importance of cue 2 in early learning trials and do not update the importance of cue 3. Hence, the estimated cue weights from the rolling regression and the LMS rule show systematic deviations during the learning phase.

Although we think that the rolling regression represents a good measurement model to identify the importance people give to different cues (without identifying the learning process), it faces the danger of overfitting when being estimated using only a small window size (Pitt & Myung, 2002). The LMS rule, in contrast, was estimated based on all training

trials, thus restricting parameter estimates more strongly. A more conservative test of model performance requires predicting new data based on the cue weights. Accordingly, we used the resulting cue weights at the end of training (or the cue weights obtained from the last 50 trials for the rolling regression) to predict participants' judgments for unknown items in the test phase.[1] Similar to the training results, the LMS rule captured judgments for unknown items better than the baseline model, V = 5073, $p < 0.001$, 95% CI [-5.2,-3.4], but still did not outperform the predictive performance of the rolling regression, V = 40759, $p < 0.001$, 95% CI [3.1,3.9]. Taken together, these results suggest that the LMS rule cannot appropriately reproduce the learning path in rule-based learning, nor accurately predict judgments for new objects after training.

Psychological constraints in rule-based learning

Why may the LMS rule fail to account for rule-based learning? The LMS rule implies that the learning rate is stable across all learning trials and all cue weights are updated with the same learning rate. Past evidence has accumulated that human rule-based learning diverges in important ways from such an idealized learning process. First, studies in which the cues' importance changes over time indicate that people adjust to this change more slowly than they acquired the solution to the initial judgment problem (Dudycha, Dumoff, & Dudycha, 1973; Peterson, Hammond, & Summers, 1965; Summers, 1969; but see Speekenbrink & Shanks, 2010). Second, increasing the validity of one cue has been shown to attenuate learning about the predictive validity of a second cue (cue competition effects, Birnbaum, 1976; Busemeyer, Myung, & McDaniel, 1993a, 1993b) indicating that learning rates for one cue depend on the existence of another cue.

These phenomena have been been traced back to different psychological mechanism altering the learning process. First, it has been assumed that people adapt to a task more

[1] In Hoffmann et al. (2014), the RMSD between model predictions and judgments in the test phase was calculated using participants' average judgment for each test item, not the individual responses on each test trial. For this reason, the RMSD reported here deviates from the one reported in the article.

slowly, the more experience they gain with the task. Accordingly, this explanation proposes that learning speed decays across learning trials (Kelley & Busemeyer, 2008; Rolison et al., 2012). Second, it has been argued that learning rules in multiple-cue judgment tasks is restricted by a limit in working memory capacity (Hoffmann, von Helversen, & Rieskamp, 2013, 2014; Juslin et al., 2008). A capacity limitation would constrain how much people update the set of hypotheses on a single trial and, in turn, would cause cue competition effects (Busemeyer et al., 1993b). Finally, it has been proposed that it is not a capacity restriction per se that limits learning, but limited attentional resources and psychological mechanisms guiding the distribution of attention during learning (Kruschke, 1996; for a review see Le Pelley, Mitchell, Beesley, George, & Wills, 2016). Accordingly, attention may limit which cues people focus on during learning and how strongly they update different cues. In the remainder of this article, we will first specify these psychological learning mechanisms mathematically and then test these mechanisms against each other and the LMS rule in two studies.

LMS rule with decaying learning speed (Decay)

Previous research supports the idea that the more experience people gain with a judgment task the more slowly they adapt to a change in the underlying task structure (Dudycha et al., 1973; Peterson et al., 1965; Summers, 1969) suggesting that people may not learn with a constant learning rate, but the learning rate may decrease over time. A decay in learning speed has been mostly instantiated in rule-based learning models by decreasing the updating of cue weights based on the number of previous trials (Kelley & Busemeyer, 2008; for a similar version see Rolison et al., 2012)

$$\Delta w_i = \frac{\lambda \cdot x_i \cdot (y - j)}{t^{\delta}} \tag{4}$$

Parameter δ controls the decay rate with $\delta > 0$. A higher decay rate implies that the learning rate more strongly declines with a higher number of learning trials. Indeed,

including a decay parameter has been shown in some tasks to provide a better description of the learning process than the LMS rule (Kelley & Busemeyer, 2008; Rolison et al., 2012).

LMS rule with a capacity restriction (Capacity)

Theories of rule-based judgment put forth the idea that cognitive capacity restrictions may affect rule-based learning (Hoffmann et al., 2014; Juslin et al., 2008). Specifically, the comparison processes involved in learning from feedback require storing and manipulating the judgment objects and thus may pose high demands on working memory (Juslin et al., 2008). Supporting this idea, higher working memory capacity has been related to a more accurate solution of rule-based judgment tasks (Hoffmann et al., 2014). Reducing working memory demands by facilitating a direct comparison of cue values in contrast speeds up learning in linear tasks (Juslin et al., 2008). Finally, cue competition effects as well point towards the idea that learning is restricted by a cognitive capacity limitation (Busemeyer et al., 1993a, 1993b). Specifically, Busemeyer et al. (1993b) found that a moderately valid cue is perceived as less valid when paired with a highly valid cue than when paired with a moderately valid cue. Based on these results, the authors argued more generally that previously proposed learning models, for instance the LMS rule, are not able to account for this effect because they gradually converge to the optimal weights (Busemeyer et al., 1993a). Instead models predicting cue competition effects need to impose a capacity constraint on the weights.

To our knowledge, past research has not yet specified, nor tested a rule-based learning model for human judgment specifying this cognitive capacity restriction. We implemented capacity restricted learning in our model by restricting the cue weights to sum up to a capacity restriction r, $\sum_i |w_i| \leq r$. If the capacity restriction is reached, each of the updated cue weights is reduced by the difference between summed cue weights and the capacity restriction, divided by the number of weights.

$$w_i = w_i - \text{sgn}(w_i) \cdot \frac{\sum_i |w_i| - r}{i} \tag{5}$$

Thus, increasing one cue weight above the capacity limit reduces all cue weights by the same magnitude and, in effect, decreases all other cue weights. Accordingly, the capacity restriction slows down updating of the cue weights and gives rise to cue competition effects. These cue competition effects are most pronounced if the capacity restriction falls below the optimal sum of weights because people will not learn to weight the cues optimally and hence will not reach optimal performance. In case, the capacity limit matches or somewhat exceeds the optimal sum of weights, the capacity model will converge over the long run to the optimal weights, as do all other models. Compared to the LMS rule, however, to what degree the cue weights are updated still hinges upon the capacity limit preventing an overly strong adaptation of the cue weights and accordingly more stable learning even for high learning rates. If the capacity limit strongly exceeds the optimal sum of weights, however, learning proceeds similarly as in the LMS rule.

LMS rule with attention weights (Attention)

Learning research has emphasized the role of attentional processes in associative learning (Denton & Kruschke, 2006; Kruschke, Kappenman, & Hetrick, 2005; Le Pelley et al., 2016), category learning (Kalish & Kruschke, 2000; Kruschke, 1996), or causal learning (Lachnit, Schultheis, König, Üngör, & Melchers, 2008). Recent research has identified the predictiveness of the cues, the salience of the cues, and the value of the outcome as major determinants of attentional biases in associative learning (Le Pelley et al., 2016). Similarly, categorization research has argued that people may shift attention between different cues depending on their importance, but also in response to the salience of single cues (Kalish & Kruschke, 2000; Kruschke, 1996). Following this previous research, we adapted an *attentional shift mechanism* specified for categorization problems to account for rule-based learning in judgment (Kruschke, 1996). The model assumes that the

judgment is in each trial a linear additive function of the cues, the cue weights, and the attention weights.

$$\hat{j} = \sum_i \alpha_i \cdot w_i \cdot x_i \tag{6}$$

where $\alpha_i = \frac{1}{n+1}$ are the attention weights. Accordingly, attention is equally distributed across all cues when people make a judgment and attention previously paid to a cue does not carry over to the next trial. Attention, however, plays an important role in updating the cue weights. Specifically, before any cue weight is updated, the model adjusts the attention weights.

$$\Delta\alpha_i = \lambda_\alpha \cdot x_i \cdot (y - \hat{j}) \cdot w_i \tag{7}$$

where λ_α is the learning rate for the attention weights. Thus, the model focuses attention more strongly on previously important cues and salient cues, that is cues with high cue values. The attention weights are then updated, $\alpha_i = \alpha_i + |\Delta\alpha_i|$. Considering only absolute changes in attention weights implies that higher cue weights draw a higher attention towards them, independent of the direction of the predicted relationship. Similarly, independent of the direction of judgment error, high judgment errors pronounce the effect of focusing attention on the most important and salient cues, compared to trials with only small errors. The updated attention weights are then normalized to reflect limits in attentional capacity.

$$\alpha_i = \frac{e^{\theta \cdot \alpha_i}}{\sum_i e^{\theta \cdot \alpha_i}} \tag{8}$$

The θ parameter determines the sensitivity to differences in attentional strength. In a next step, the cue weights are adjusted based on the cue value, feedback, and the attention attached to each cue.

$$\Delta w_i = \lambda_w \cdot (y - \hat{j}) \cdot x_i \cdot \alpha_i \tag{9}$$

Accordingly, the more attention a cue receives, the more strongly the cue weights are updated. In sum, the attention model postulates that more attention flows towards errors that have been caused by cues that are salient and have been previously predictive. This means that people should adapt faster to changes in importance of previously important cues, but adapt more slowly to changes in the importance of previously unimportant cues.

Reanalysis: Comparing psychological learning models to the LMS rule

The proposed psychological learning models aim to incorporate key processes that alter and limit people's learning abilities in rule-based judgment. Compared to the LMS rule, can those psychological mechanisms better capture how people learn to solve rule-based judgment tasks? To understand which learning mechanism best describes and predicts participant's judgments in the experiment, we compared the models on two model comparison criteria: the Bayesian Information Criterion (BIC, Schwarz, 1978) and a generalization test (Busemeyer & Wang, 2000, see Appendix A for a more detailed description). Whereas the BIC penalizes more complex models by accounting for the number of free parameters,[2] the generalization test measures to what degree the models can also predict independent, unseen data. To calculate the BIC, we estimated each model's parameters based on all training trials. Based on the BIC, we then derived the corresponding Bayesian Information weights, BIC_w, that yield the posterior probability of each model given the data (Wagenmakers & Farrell, 2004). For the generalization test we used the cue weights from the last learning trial to predict participants' judgments for new objects. Next, we computed the deviances between model predictions and participants' judgments for all test trials, D, and similarly derived deviance weights, D_w.

Descriptively, the average BIC is lower for the capacity and the attention model than

[2]Of the three extensions of the LMS rule that we tested the decay model and the capacity model include each one additional free parameter (the decay parameter δ, and the capacity restriction parameter r respectively) and the attention model includes two additional parameters, the attentional learning parameter λ_α and the sensitivity parameter θ.

for the decay model and the LMS model with the capacity model reaching the highest BIC_w (see Table 1). The decay model overall does not outperform the LMS rule. Using the BIC_w to classify participants to each model further suggests that the majority of participants is best described by the capacity model and only a minority is classified to the decay or the attention model. Also, the LMS rule and the baseline model do not describe a substantial number of participants.

Reflecting the results from training, the generalization test suggests an overall lower D for the capacity and the attention model. Classifying participants to each model based on the D_w again indicates that the capacity model best predicted judgments of the majority of participants, whereas the decay model only described a minority of participants best. The attention model best predicted a slightly larger number of participants compared to the results based on BIC —mostly at the cost of the capacity model.

To gain more insight into the learning path, we compared the cue weights predicted by each model to the weights of the rolling regression (Figure 1). These graphs suggest that the decay model underestimates the importance of most cues for making a judgment. In comparison, the capacity model in general catches the change in weights but slightly overestimates the importance of the most predictive cue and underestimates the importance of the least important cue. Finally, the attention model most precisely estimates the cue weight of the most important cue and manages to match the rolling regressions' cue weights for the second and third most important cues at the end of the learning phase. However, the cue weights deviate from the rolling regression in the first two thirds of the learning phase and the model underestimates the weight for the least important cue.

In sum, the reanalysis suggested that a capacity-restricted learning model best described rule-based learning, whereas a decay in learning speed or an attentional mechanism fared less well. Compared with a stable rule-based judgment model at the end of training, the capacity-restricted learning also predicted judgments for new objects fairly well.

Testing mechanisms of rule-based learning in a relearning experiment

The results from the reanalysis suggested that learning models incorporating psychological mechanisms may better capture the learning path and also improve predictions for unseen objects. In this reanalysis, participants only had to find out once how important the cues are for making a judgment and the importance assigned to different cues did not change over trials. The vast benefit of learning models is, however, that they are able to predict how people learn to adapt their behavior to a new task. Specifically, the decay model predicts that people should adapt more slowly to the new task. In contrast, the capacity model predicts that people will not reach optimal performance if the capacity limit has been exceeded. Finally, the attention model suggests that attention focuses on cues that were previously relevant or are highly salient when relearning a new task. The learning models hence allow fine-grained predictions about how people should change their judgment policy, if the underlying task changes. In consequence, to further evaluate the learning models and to test them rigorously against each other it is necessary to contrast the models' predictions in an experiment in which the importance of the cues changes over trials and people have to adapt their judgment policy.

Therefore, we designed a relearning experiment in which 51 participants solved a multiple-cue judgment task that changed over the course of learning. In the first half of the learning phase, participants learned to predict the judgment based on four predictive cues. After 200 trials, the least important cues gained importance for predicting the criterion, whereas the two most important cues lost their importance.

Method

Participants. 51 participants (39 females, $M_{\text{Age}} = 22.1$, $SD_{\text{Age}} = 3.6$) were recruited from the participant pool of the University of Basel. Participants received course credit for their participation in the experiment. In addition, they could earn a performance-dependent bonus ($M = 6.2$ CHF, $SD = 2.9$ CHF).

Design and material. The cover story in the multiple-cue judgment task asked participants to judge how many small animals different comic figures, the Sonics, caught on a scale from 0 to 50. Participants were presented with pictures of these Sonics that varied on four different quantitative cues. The Sonics had different sizes of their ears and their nose, and a different number of hairs and stripes on their shirt. These pictorial cues could be used to predict the criterion (the success of the Sonic).

To test how well the different learning models predict the learning path of the participants, we changed during the judgment task how these cues had to be combined to form the judgment criterion. Specifically, in the first 200 trials of the judgment task the criterion was a linear, additive function of the cues, $y = 4x_1 + 3x_2 + 2x_3 + x_4$. After 200 trials, the task structure suddenly changed so that the two most important cues lost their predictive power, whereas the two least important cues gained importance, $y = 4x_3 + 6x_4$. To select this particular task structure, we used in a first step the median (as well as the most likely) parameters from the reanalysis to predict learning performance as well as the adaptation to a new task structure across a range of task combinations. For instance, we also generated predictions for an experiment in which only one cue was predictive in the first half of the experiment and after the task structure changed all cues were equally predictive. We selected the task combination that best allowed us to make distinct predictions for the learning models. Within this task combination, we next constructed a presentation sequence from all possible items that maximized the possibility to discriminate between the learning models. To generate this sequence, we randomly drew 1000 learning sequences, each consisting of 400 items, and estimated the models' parameters for this sequence. We finally selected the presentation sequence that maximized differences in model predictions across the whole learning phase as well as for the first 50 trials of the relearning phase.

For each participant, the cues x_1 to x_4 were randomly assigned to the pictorial cues (e.g., ears or nose). Higher cue values, however, were always associated with more salient

pictorial features. For instance, a cue value of zero corresponded to a Sonic without stripes on the belly and a cue value of five to a Sonic with five stripes on the belly. Likewise, a cue value of zero on the cue 'nose' corresponded to a Sonic without a (visible) nose, whereas a Sonic with a cue value of five had a big nose.

Procedure. The experiment consisted of 400 learning trials, divided into 16 learning blocks with 25 trials each. In each trial, participants first estimated the criterion on a scale from 0 to 50 and afterwards received feedback about their own answer, the correct outcome, and the points they earned. After 200 learning trials (i.e., after the eight block), the task structure changed and participants had to relearn the importance of the cues. Participants were informed in the beginning of the experiment about a potential change in the task, yet were not informed when the change would happen but had to infer that a change occurred from the feedback they received.

To motivate participants to achieve a high judgment accuracy, they could earn points in each trial depending on how much their judgment j deviated from the correct criterion y:

$$\text{Points} = 20 - \frac{(j-y)^2}{7.625} \tag{10}$$

This function was truncated so that participants could win at most 20 points in each trial and could not lose any points. In addition, participants could earn a bonus of 3 CHF in the last learning block of each task (learning block 8 and 16) if they reached 80% of the points.

Results

Learning performance. The learning performance suggested that participants on average adapted to the change in the task structure (Figure 2). Descriptively, judgment error, measured as the root mean square deviation (RMSD) between participants' judgments and the criterion, dropped from the first learning block (M = 9.1, SD = 1.9, with each block including 25 items) to the eighth block, M = 6.2, SD = 2.0, $t(50)$ = -9.0, d

$= -1.45$, $p < 0.001$ (d calculated using an effect size based on the change score for repeated measures Morris & DeShon, 2002). When the task changed after the eighth block, judgment error suddenly increased, $M = 10.6$, $SD = 2.3$, $t(50) = 12.0$, $d = 1.92$, $p < 0.001$, but dropped again until the end of the experiment, $M = 7.4$, $SD = 3.6$, $t(50) = -7.1$, $d = -0.88$, $p < 0.001$. Inspecting individual learning paths indicated that participants varied strongly in the degree to which they successfully adapted to the change in task structure. Compared to the first eight blocks of the experiments, judgment performance varied more strongly between participants after the task structure changed. Whereas some participants quickly achieved a high judgment accuracy, other participants did not show any improvement in judgment accuracy. This qualitative pattern indicates that how people learn to adapt to a change in task structure may vary between participants and may suggest different underlying learning mechanisms.

Average performance of the learning models. To understand which learning mechanism best describes and predicts participant's judgments over the experiment, we compared each model's performance based on the BIC and based on the generalization test (see Appendix A for a more detailed description). For the BIC, we estimated each model's parameters based on all trials in the experiment and calculated the BIC weights (see Appendix B for model parameters). To consider as well how accurately all models predict new data, we further performed a generalization test (Busemeyer & Wang, 2000). Specifically, we estimated each models' parameters based on participants' judgments in the first 200 trials, used the obtained parameters (as well as the final weights) to predict participants' judgments in the second half of the experiment, and calculated the deviance D based on these predictions. Table 2 summarizes the model fits, the relative performance of all models within the set of considered models (BIC_w and D_w) as well as the absolute fit between model predictions and participants' judgments (RMSD).

As in the reanalysis, the capacity and the attention model possess a lower BIC than the decay model or the LMS rule with the capacity model outperforming all other models.

The decay model describes judgments slightly better than the LMS rule. But can the capacity model also predict how well participants adapt to the change in task structure? Matching the results based on BIC, the capacity model best predicts participants' judgments in the second half of the experiment. The LMS rule and the decay model also outperform a baseline model in predicting how participants adapt to the change in task structure, but the decay model fares worse than the LMS rule. Yet, the relative advantage of the attention model vanishes in generalization. Specifically, the D of the attention model is similar to the D of the baseline model indicating that the attention model has problems to predict how participants relearn the task. In fact, the model generated a higher D than the baseline model for 21 participants and the RMSD between model predictions and participants' judgments suggests a stronger increase for the attention model than for the other models compared to fitting.

To more closely investigate to what extent the learning path of the learning models agrees with the average learning path of all participants, we generated the predicted RMSD in each learning block based on each model's predictions and the models' implied standard deviation. Figure 3 depicts for each learning block the average judgment error of all participants (black lines) as well as the average judgment error predicted by each model (gray lines, in columns), separately for model estimation and generalization (in rows). White diamonds illustrate the absolute difference between the model's implied learning path and participants' learning path, averaged across participants. Light gray lines show the model predicted judgment error for each single participant. Early in training, the LMS rule, the decay, and the attention model on average underestimate how well the average participant adapts to the judgment task, but this difference mostly vanishes in later learning blocks (Figure 3, upper row). In contrast, the capacity model captures quite well the average learning path in the first half of the experiment, but generates overly optimistic predictions about how successful participants adapt to the change in task structure and, hence, the absolute difference between the learning paths increases in later learning blocks.

Focusing more on the variation in individual model predictions, the graphs illustrate that predicted judgment error is more variable for the LMS rule and the decay model than for the capacity or the attention model across all learning blocks. Specifically, the capacity model suggests for most participants a steady learning path in the first half of the experiment as well as an improvement after the task structure changed with (mostly) faster learning in the first learning blocks. The LMS rule and the decay model instead allow for the possibility that after the task structure changed, judgment error does not sufficiently decrease even after several learning blocks. Finally, the attention model describes a slower learning path compared to the capacity model and participants in the first learning phase, and allows a strong increase in judgment error even after several learning blocks in the relearning phase.

In generalization, the overall predictive performance of all models drops and differences in model implied learning paths are even more pronounced (Figure 3, lower row): Whereas the capacity model mostly optimistically predicts a steady improvement for the second half, the LMS rule and —to an even stronger degree —the decay model are more likely to predict a high amount of judgment errors. Furthermore, the variability in learning paths is much lower for the capacity model than for all other models. Finally, even though the average predictions of the attention model appear to match the average learning curve of participants well, the absolute difference between learning paths suggests a rather strong increase in mismatch. Particularly, the attention model emphasizes that judgment error may increase for some participants and predicts for a sizeable number of participants large judgment errors late in learning (block 14-16). This might have contributed to its inability to predict the learning path of individual participants.

Taken together, all learning models incorporating an additional psychological learning mechanism outperformed the LMS rule in terms of BIC, but only the capacity model keeps this advantage in generalization. Average model fits suggest that the capacity model overall describes the learning path best. Yet, the high variability in model predictions as well as in

learning performance of individual participants make it likely that different learning mechanisms account better for different subgroups of individuals. Accordingly, in a next step we assessed which learning mechanism best describes the majority of participants and how individuals classified to different learning mechanisms differ in their overall learning path.

Learning path for individual participants. To identify whether the learning mechanisms best described different subgroups of individuals, we classified participants to the different learning models based upon the relative performance of those models (that is, the BIC_w or the D_w in the generalization test, respectively). Reflecting the results from the reanalysis, the classification based on the BIC_w indicated that the capacity model described the majority of participants best (66.7 %) and only a minority of participants were better described by the decay model (23.5 %), the attention model (7.8 %), or the LMS rule (2.0 %). A classification based on the D_w suggested similarly that the majority of participants (52.9 %) was classified as best predicted by the capacity model. Further, some participants were classified as best predicted by the decay model (17.6 %), whereas the LMS rule or the attention model only predicted judgments of a few participants best (7.8 % and 5.9 %, respectively). Yet, a substantial number of participants were classified to the baseline model (15.7 %) indicating that the learning models are prone to overfitting the data.

Figure 4 displays the learning path for participants best described by each model (black lines, in columns) as well as the average judgment error predicted by each model (gray lines, in columns), separately for model fitting and generalization (in rows). Light gray lines show the model predicted judgment error for each single participant. Considering first the learning path for model classifications based on BIC weights (Figure 4, upper row), the learning models seem to capture different learning patterns best. Specifically, the decay model proposes predominantly that performance steadily improves in the first half of the experiment for participants best fit by the model, but judgment error only slowly decreases in the second half indicating that those participants may only slowly

adapt to the changing task structure. The capacity model similarly proposes a steady improvement during the first half, but in comparison predicts a faster decline in judgment error after the change indicating a more successful relearning of the judgment task. In contrast, the attention model captures the learning path of participants best for whom it predicts only slow improvements in the first half of learning as well as adaptation problems after the task structure changed, as indicated by even a slight increase in judgment error from learning block 9 to 12. Finally, the participant classified to the LMS rule displays a learning path that systematically deviates from the learning path implied by the model. The qualitative differences in model predictions between the decay and the capacity model are even more pronounced for participants classified based on D_w obtained from the generalization test (Figure 4, lower row). Whereas the decay model predicts for most participants a high judgment error after the task structure changed, the capacity model predicts a more successful learning path for most participants.

Taken together, the capacity model best described and predicted how the majority of participants learned to adapt their judgments over trials suggesting on average a steady adaptation to the change in task structure. The decay model fared best in describing and predicting those participants who more slowly detect this change and in turn show a delayed improvement in judgment accuracy.

Robustness check. In the preceding analysis, all learning models included the strong assumption that participants make the first judgment without considering any cues or cue values, that is, the starting weights in the first trial were set to [0 0 0 0 j] with j reflecting a starting bias corresponding to an intercept in a regression model. It is possible that this assumption may have biased our analysis and another learning model may yield a better performance if we relax this assumption. To control for this possibility, we varied the starting weights systematically from assuming that all cues equally contribute to the judgment in the first trial, but not the bias (0 % bias) to no contribution of the cues to the judgment (100 % bias corresponding to our previous analysis) in steps of 12.5 % bias. The

weights in the first trial were thus calculated as

$$w_n = \frac{j * (100\% - b)}{\sum x_n} \tag{11}$$

with b varying the percentage of bias. If the starting weights biased our analysis towards the capacity model and another model, for instance the attention model, performs better considering a different set of starting weights, we would expect that this competitor shows consistently a higher BIC_w (or D_w, respectively) than the capacity model for different sets of starting weights. A mere reduction in BIC_w for the capacity model, however, could also result because we maximized the possibility to discriminate between models using the starting weights with a 100 % bias. Thus, the ability to discriminate between the models may be lower for different starting values and the reduction in BIC_w is not a sufficient indicator for a worse model performance.

Figure 5 displays how the average BIC_w (left panel) and D_w (right panel) for each model (separate lines: baseline, rule, decay, capacity, and attention) vary as a function of the percentage of bias. For both BIC_w and D_w, the pattern suggests that the weights for the capacity model increase with a higher percentage of bias. In contrast, with a lower bias the BIC_w for the LMS rule and the decay model increase. In generalization, the D_w for the LMS rule and the baseline model increase similarly with a lower bias. Still, the capacity model possesses a higher BIC_w and a higher D_w across all starting weights we used. In sum, although the discrimination between the models varies with the bias, the advantage of the capacity model appears to be robust against variations in starting weights.

General Discussion

Weighing information according to its importance has been deemed one of the core competences in human judgment. However, how people form these weights has received less attention. The predominant model to describe the learning process, the LMS rule, assumes that people will be able to learn the optimal cue weights when receiving appropriate

feedback —an assumption that contradicts prior evidence showing that human learning depends on the relative weight of the cues (Busemeyer et al., 1993b) and on average slows down when people have to adapt to a new task structure (Betsch, Brinkmann, Fiedler, & Breining, 1999, 2001; Bröder & Schiffer, 2006; Rieskamp, 2006). Still, little research has tried to capture how people learn the importance of cue weights within a formal modeling approach (for exceptions see Kelley & Busemeyer, 2008; Speekenbrink & Shanks, 2010). In our study, we aimed to fill this gap by a) systematically comparing the LMS rule to human learning in judgment and b) by implementing three psychological mechanisms into the LMS rule that have explained deviations from optimal learning in related research areas: a decay of the learning rate, a capacity restriction, and attentional learning.

Overall, we found that the psychological learning mechanisms better described the learning process than the simple LMS rule. In the reanalysis, all psychological learning models predicted judgments more accurately than the LMS rule. To tear the psychological learning mechanisms apart, we designed an experiment assessing how well people relearn a task after the judgment environment changed making it necessary that participants adjusted the learned cue weights to accurately predict the criterion. In this experiment, however, only the capacity model outperformed the LMS model and the baseline model in both fitting the whole learning phase and predicting learning in a generalization test. These results show that considering psychological constraints is essential to understand how people learn to solve judgment problems and suggest that capacity restrictions are the most likely mechanism explaining the systematic differences between human performance and an optimal learning algorithm.

The advantage of capacity restrictions in learning

Why is there a benefit for the capacity model? The capacity model has been motivated by research on cue competition and capacity limits in judgment and decision making (Birnbaum, 1976; Busemeyer et al., 1993a, 1993b; Juslin et al., 2008). It assumes

that a capacity limit restricts how much importance people assign to the cues, which slows down updating of the cue weights and allows for the possibility that people do not learn the optimal weights even after a long time. Further, increases in one weight imply decreases in other weights. The version we proposed here assumes that if the capacity limit is exceeded all cues weights are reduced by the same amount. Thus, compared to previous evidence suggesting that the less valid cue suffers more from cue competition than the highly valid cue (an asymmetric effect Busemeyer et al., 1993a), our model proposes that all cues are affected by cue competition to the same degree (a symmetric effect). Although it would be possible to adapt the updating rule (Equation 5) to allow for asymmetric effects, previous work also suggests that cue competition effects may be less pronounced in judgment than they are in related domains. Specifically, Speekenbrink and Shanks (2010) only observed cue competition effects in a minority of participants and a model incorporating competitive effects only described a few participants best. Here, further research should specify the conditions under which cue competition effects are likely to occur in judgment and investigate which updating rule is called for.

Interestingly, the model proposes that this capacity restriction may possess an adaptive value. Specifically, in the LMS rule high learning rates are not always beneficial because they result in an over-adaptation of the cue weights and consequently the model does not learn the task. A capacity limit close to the optimal sum of weights limits the possibility that people update the cue weights too strongly and thus may enable higher learning rates. Another important implication of a capacity limitation is that the degree to which updating is affected increases with the number of cues. In line with this idea, research suggests that varying one cue between trials can lead to faster learning (Juslin et al., 2008) because in this case the respective cue weight can be learnt without reaching the capacity limit.

Slower learning with more experience

A common finding in the learning literature is that people are able to relearn a task —albeit more slowly than in the original task (Betsch et al., 1999, 2001; Bröder & Schiffer, 2006; Dudycha et al., 1973; Peterson et al., 1965). Overall, in the second study people were slower to learn the cue weights after the change than in the first learning phase. However, there were large individual differences in the overall pattern of how people adapt to a change in task structure. Whereas some people rather quickly adapted to the task structure, others had problems with relearning the task. These results resonate well with the findings by Speekenbrink and Shanks (2010) who also found large individual difference in the ability to adapt to changes in cue validity. However, in both studies only a minority of participants was best described and predicted by the decay model suggesting that a pure slowing of learning over time is not enough to capture how people learn each cues' importance.

One reason for why decay only played a minor in our experiment is potentially that we informed participants about a potential change in the task structure and introduced a rather salient shift in the cues' importance. This shift in task structure resulted in a strong reduction in judgment accuracy in the ninth block which may have clearly signaled to participants that they should change their judgment policy. In this vein, previous research suggests that learning rates depend on whether people expect a change or not. For instance, Behrens, Woolrich, Walton, and Rushworth (2007) found on average higher learning rates in variable environments including a lot of changes than in stable environments in which no change occured. Accordingly, including a mechanisms allowing for decay in the learning rate may gain importance when changes occur gradually and are unexpected.

Attentional learning

Attentional learning has been considered an important mechanism in learning and evidence for the idea that attention influences learning processes is widespread (Le Pelley et al., 2016). Effects of attention on learning have also been successfully demonstrated in fields closely related to judgment research such as category learning. For instance, measures of attention such as eye movements have been shown to reflect the importance of cues in categorization decisions (Beesley, Nguyen, Pearson, & Le Pelley, 2015; Hoffman & Rehder, 2010; Rehder & Hoffman, 2005). Further, attentional shifts can explain learning phenomena such as blocking (Kruschke et al., 2005), overshadowing (Denton & Kruschke, 2006) and may also explain cue competition (Kruschke, 2001). Thus, attentional learning seemed a promising candidate to explain how people deviate from optimal learning. Here, we adapted a mechanism from category learning (Kruschke, 1996) to account for multiple-cue judgments assuming that high error on that trial, previously important cues, and high cue values increase attention to specific cues.

On average, the attention model performed quite well when fitting participants' data (second runner up). However, on the individual level only a small number of participants were classified to the model. The model suffered even more in the generalization test indicating strong overfitting. These misfits are potentially caused by an overly strong focus on errors that can strongly change cue weights even at the end of the learning phase and consequently a large variability in learning performance. Possibly, reducing the strength of the attentional effect, limiting the factors that guide attention, or reducing the sensitivity over the course of the learning phase (see also Le Pelley et al., 2016) could make the model more resistant to overfitting and better suited to explain learning in judgment problems.

Further frameworks for learning multiple cue judgments

In the present work we focused on evaluating psychological constraints within the framework of rule-based models learning from prediction error. Assuming that people rely

on a linear additive judgment rule, the proposed rule-based learning models update the cue weights based on the difference between the judgment and external feedback. However it has been argued recently that human learning processes may be better described by Bayesian learning mechanisms. In this vein, Speekenbrink and Shanks (2010) proposed a Bayesian model of how people learn the cue weights in a linear judgment rule, which described participants' judgments better than the LMS rule.

In addition, it has been doubted that people solve judgment tasks by relying on explicit linear rules, but may rather learn the associations between specific patterns of cue and criterion values. In this vein, the associative learning model (ALM, Busemeyer, Byun, Delosh, & McDaniel, 1997) has been shown to describe people's performance in a variety of judgment tasks well and to outperform a simple LMS model (Kelley & Busemeyer, 2008; Speekenbrink & Shanks, 2010). In category learning, the predominant model is the exemplar-based neural network ALCOVE (Kruschke, 1992) that could be adapted to describe learning in judgment tasks. Similar to the LMS rule, ALM and ALCOVE are, however, unable to predict cue competition effects (Busemeyer et al., 1993b) suggesting that both models would need to include psychological constraints such as a capacity restriction to describe human learning. Future research may try to disentangle the question of which framework describes learning processes best from the question of which psychological constraints are necessary to account for learning processes by implementing these constraints in a similar fashion across different frameworks.

Conclusion

In sum, we aimed to investigate the psychological mechanisms constraining how people learn to weigh different pieces of information in multiple cue judgment tasks. All three mechanisms improved how well the LMS rule described the learning process, but including a capacity restriction matched human performance most closely. These results suggest that limited cognitive resources that confine knowledge updating may cause

deviations from optimal learning and highlight that considering psychological constraints is crucial to understand human behavior.

References

Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review*, *78*(3), 171–206. doi:10.1037/h0030834

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *The Quarterly Journal of Experimental Psychology*, *68*(11), 2175–2199. doi:10.1080/17470218.2015.1009919

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. doi:10.1038/nn1954

Betsch, T., Brinkmann, B. J., Fiedler, K., & Breining, K. (1999). When prior knowledge overrules new evidence: Adaptive use of decision strategies and the role of behavioral routines. *Swiss Journal of Psychology*, *58*(3), 151–160. doi:10.1024//1421-0185.58.3.151

Betsch, T., Haberstroh, S., Glöckner, A., Haar, T., & Fiedler, K. (2001). The Effects of Routine Strength on Adaptation and Information Search in Recurrent Decision Making. *Organizational Behavior and Human Decision Processes*, *84*(1), 23–53. doi:10.1006/obhd.2000.2916

Birnbaum, M. H. (1976). Intuitive Numerical Prediction. *The American Journal of Psychology*, *89*(3), 417. doi:10.2307/1421615

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, *87*(2-3), 137–154. doi:10.1016/0001-6918(94)90048-5

Bröder, A. & Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 361–380. doi:10.1037/0278-7393.32.4.904

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. M. (1997). Learning functional relations based on experience with input-output pairs by humans and

artificial neural networks. In Lamberts K. & D. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 408–437). Cambridge, MA, US: MIT Press.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993a). Cue competition effects: Theoretical Implications for Adaptive Network Learning Models. *Psychological Science*, *4*(3), 196–202. doi:10.1111/j.1467-9280.1993.tb00487.x

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. M. (1993b). Cue competition effects: Empirical tests of adaptive network learning models. *Psychological Science*, *4*(3), 190–195. doi:10.1111/j.1467-9280.1993.tb00486.x

Busemeyer, J. R. & Wang, Y.-M. (2000). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189. doi:10.1006/jmps.1999.1282

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods and applications*. San Diego, CA: Academic Press.

Denton, S. E. & Kruschke, J. K. (2006). Attention and salience in associative blocking. *Learning & behavior*, *34*(3), 285–304. doi:10.3758/BF03192884

Dudycha, A. L., Dumoff, I. G., & Dudycha, L. W. (1973). Choice Behavior in Dynamic Environments. *Organizational Behavior and Human Performance*, *9*, 328–338.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*(5), 465–485. doi:10.1037//0033-295X.86.5.465

Fishbein, M. & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Reading, MA: Addison-Wesley.

Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–47. doi:10.1037/0096-3445.117.3.227

Graves, L. M. & Karren, R. J. (1992). Interviewer Decision Processes and Effectiveness: An experimental Policy-capturing Investigation. *Personnel Psychology*, *45*, 313–340. doi:10.1111/j.1744-6570.1992.tb00852.x

Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *104*(2), 335–353. doi:10.1037/a0030383

Hoffman, A. B. & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–40. doi:10.1037/a0019042

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, *24*(6), 869–879. doi:10.1177/0956797612463581

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*(6), 2242–2261. doi:10.1037/a0037989

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1193–1217. doi:10.1037/xlm0000241

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298. doi:10.1016/j.cognition.2007.02.003

Kalish, M. L. & Kruschke, J. K. (2000). The role of attention shifts in the categorization of continuous dimensioned stimuli. *Psychological Research*, *64*(2), 105–116. doi:10.1007/s004260000028

Karelaia, N. & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. doi:10.1037/0033-2909.134.3.404

Kelley, H. & Busemeyer, J. R. (2008). A comparison of models for learning how to dynamically integrate multiple cues in order to forecast continuous criteria. *Journal of Mathematical Psychology*, *52*(4), 218–240. doi:10.1016/j.jmp.2008.01.009

Kelley, H. & Friedman, D. (2002). Learning To Forecast Price. *Economic Inquiry*, *40*(4), 556–573. doi:10.1093/ei/40.4.556

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. doi:10.1037/0033-295X.99.1.22

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 3–26. doi:10.1037/0278-7393.22.1.3

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812–863. doi:10.1006/jmps.2000.1354

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 830–845. doi:10.1037/0278-7393.31.5.830

Lachnit, H., Schultheis, H., König, S., Üngör, M., & Melchers, K. (2008). Comparing elemental and configural associative theories in human causal learning: A case for attention. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*(2), 303–313. doi:10.1037/0097-7403.34.2.303

Lagnado, D. a., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162–183. doi:10.1037/0096-3445.135.2.162

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*(10), 1111–1140. doi:10.1037/bul0000064

Morris, S. B. & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. doi:10.1037/1082-989X.7.1.105

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, *138*(4), 517–534. doi:10.1037/a0017351

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, *38*(2), 329–337. doi:10.1016/S0896-6273(03)00169-7

Pachur, T. & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. doi:10.1016/j.cogpsych.2012.03.003

Peterson, C. R., Hammond, K. R., & Summers, D. A. (1965). Multiple Probability-Learning with Shifting Weights of Cues. *The American Journal of Psychology*, *78*(4), 660. doi:10.2307/1420932

Pitt, M. A. & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425. doi:10.1016/S1364-6613(02)01964-2

Rehder, B. & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. doi:10.1016/j.cogpsych.2004.11.001

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*(6), 64–99. doi:10.1101/gr.110528.110

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1355–1370. doi:10.1037/0278-7393.32.6.1355

Rolison, J. J., Evans, J. S. B. T., Dennis, I., & Walsh, C. R. (2012). Dual-processes in learning and judgment: Evidence from the multiple cue probability learning paradigm. *Organizational Behavior and Human Decision Processes*, *118*(2), 189–202. doi:10.1016/j.obhdp.2012.03.003

Scheibehenne, B., von Helversen, B., & Rieskamp, J. (2015). Different strategies for evaluating consumer products: Attribute- and exemplar-based approaches compared. *Journal of Economic Psychology*, *46*, 39–50. doi:10.1016/j.joep.2014.11.006

Schultz, W. & Dickinson, A. (2000). Neuronal Coding of Prediction Errors. *Annual Review of Neuroscience*, *23*(1), 473–500. doi:10.1146/annurev.neuro.23.1.473

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 433–443. doi:10.1037//0278-7393.17.3.433

Siegel, S. & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321. doi:10.3758/BF03210755

Speekenbrink, M. & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*(2), 266–98. doi:10.1037/a0018620

Summers, D. A. (1969). Adaptation to Change in Multiple Probability Tasks. *The American Journal of Psychology*, *82*(2), 235–240.

Sutton, R. S. & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*(2), 135–170. doi:10.1037/0033-295X.88.2.135

Tobler, P. N., O'doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*, *95*(1), 301–10. doi:10.1152/jn.00762.2005

von Helversen, B. & Rieskamp, J. (2009). Predicting sentencing for low-level crimes: Comparing models of human judgment. *Journal of Experimental Psychology: Applied*, *15*(4), 375–395. doi:10.1037/a0018024

Wagenmakers, E.-J. & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. doi:10.3758/BF03206482

Wigton, R. S. (1996). Social Judgement Theory and Medical Judgement. *Thinking & Reasoning*, *2*(2-3), 175–190. doi:10.1080/135467896394492

Table 1

Model Fits in the Reanalysis. SD in Parantheses

Model	Training					Test			
	BIC	BIC_w	n	RMSD	$RMSD_{LB}$	D	D_w	n	RMSD
Baseline	2848 (156)	0.01 (0.12)	4	9.2 (1.2)	9.2 (1.6)	879 (272)	0.04 (0.2)	13	14.2 (6.2)
LMS rule	2740 (279)	0.04 (0.17)	13	8.4 (2.4)	7.7 (2.4)	672 (82)	0.04 (0.12)	10	9.0 (3)
Decay	2740 (288)	0.14 (0.34)	41	8.3 (2.5)	7.9 (2.6)	663 (82)	0.15 (0.33)	45	8.7 (2.9)
Capacity	2493 (222)	0.68 (0.46)	196	5.9 (1.6)	5.5 (1.9)	593 (78)	0.56 (0.46)	163	6.3 (2)
Attention	2580 (192)	0.12 (0.32)	33	6.6 (1.3)	5.9 (1.6)	625 (107)	0.21 (0.36)	56	7.0 (2.7)
Regression	—	—	—	4.7 (1.3)	4.3 (1.5)	—	—	—	5.4 (1.7)

Note. BIC = Bayesian Information Criterion; BIC = Bayesian Information Criterion weight; RMSD = Root Mean Square Deviation; $RMSD_{LB}$ = Root Mean Square Deviation in the last training block; D = Deviance; D_w = Deviance weight. RMSD in the training phase was calculated only for trial 50-250 where all models yield predictions.

Table 2

Model Fits in the Relearning Experiment. SD in Parentheses

Model	BIC				Generalization			
	BIC	BIC_w	n	RMSD	D	D_w	n	RMSD
Baseline	4674 (153)	0 (0)	0	10.0 (0.9)	2382 (98)	0.15 (0.35)	8	10.5 (1.2)
LMS rule	4396 (349)	0.01 (0.10)	1	8.7 (1.9)	2280 (218)	0.09 (0.22)	4	9.2 (2.0)
Decay	4369 (360)	0.24 (0.43)	12	8.5 (1.9)	2291 (185)	0.18 (0.36)	9	9.3 (2.0)
Capacity	4127 (353)	0.67 (0.47)	34	7.3 (1.5)	2212 (352)	0.52 (0.49)	27	8.0 (2.4)
Attention	4256 (283)	0.08 (0.27)	4	7.8 (1.3)	2375 (534)	0.06 (0.18)	3	9.2 (2.1)

Note. BIC = Bayesian Information Criterion; BIC_w = Bayesian Information Criterion weight; RMSD = Root Mean Square Deviation; D = Deviance; D_w = Deviance weight.

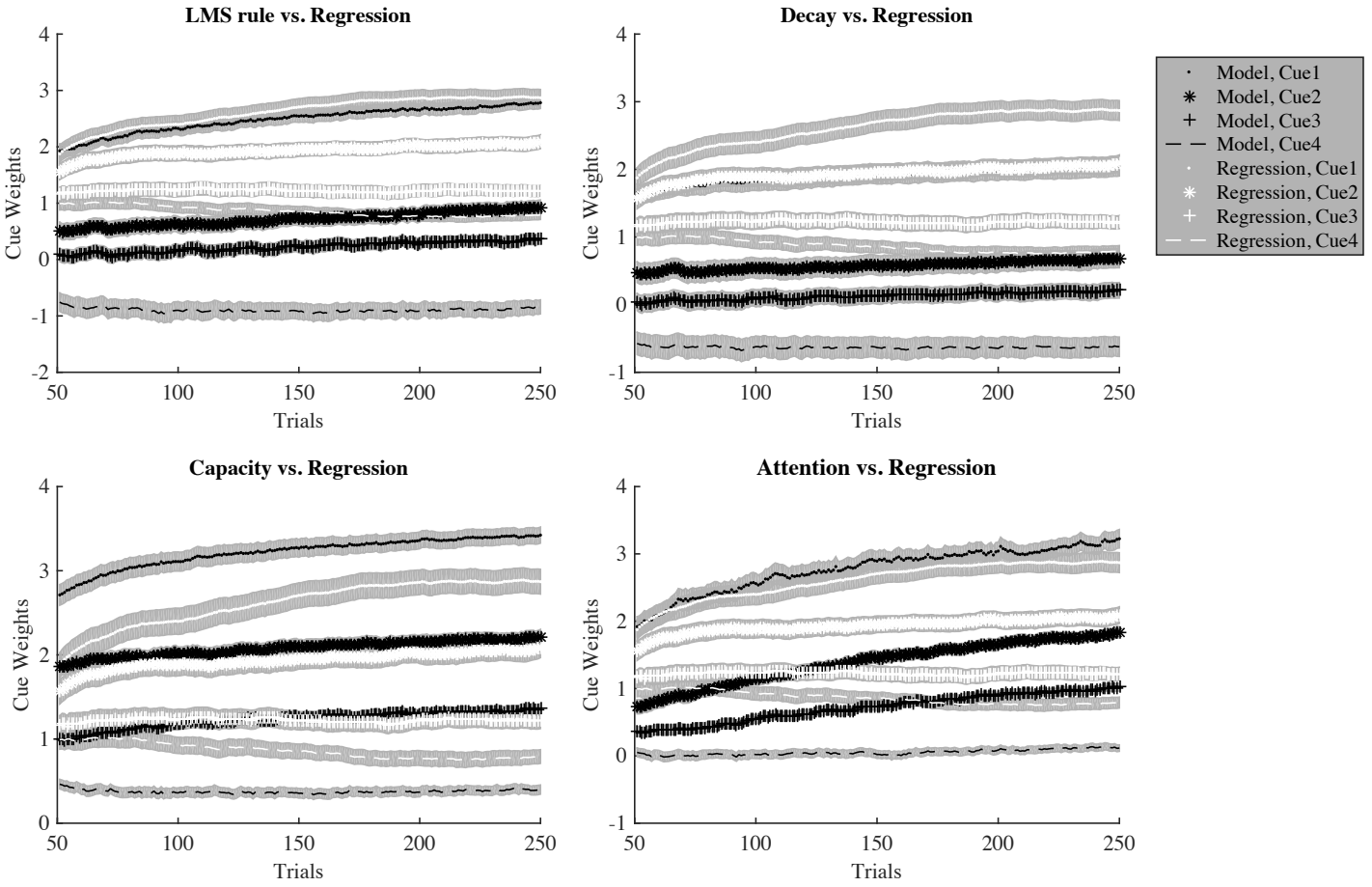

Figure 1. Cue weights predicted by each model in the reanalysis compared to cue weights from a rolling regression. Grey shaded areas show confidence intervals.

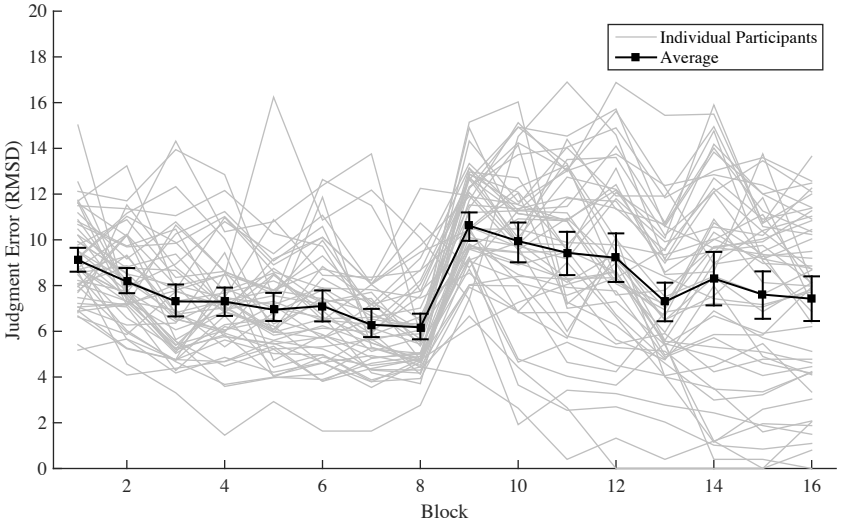

Figure 2. Judgment error (Root Mean Square Deviation, RMSD) in the Relearning Experiment. The black line shows the average judgment error, gray lines show judgment error of individual participants in each learning block. Error bars plot bootstrapped confidence intervals.

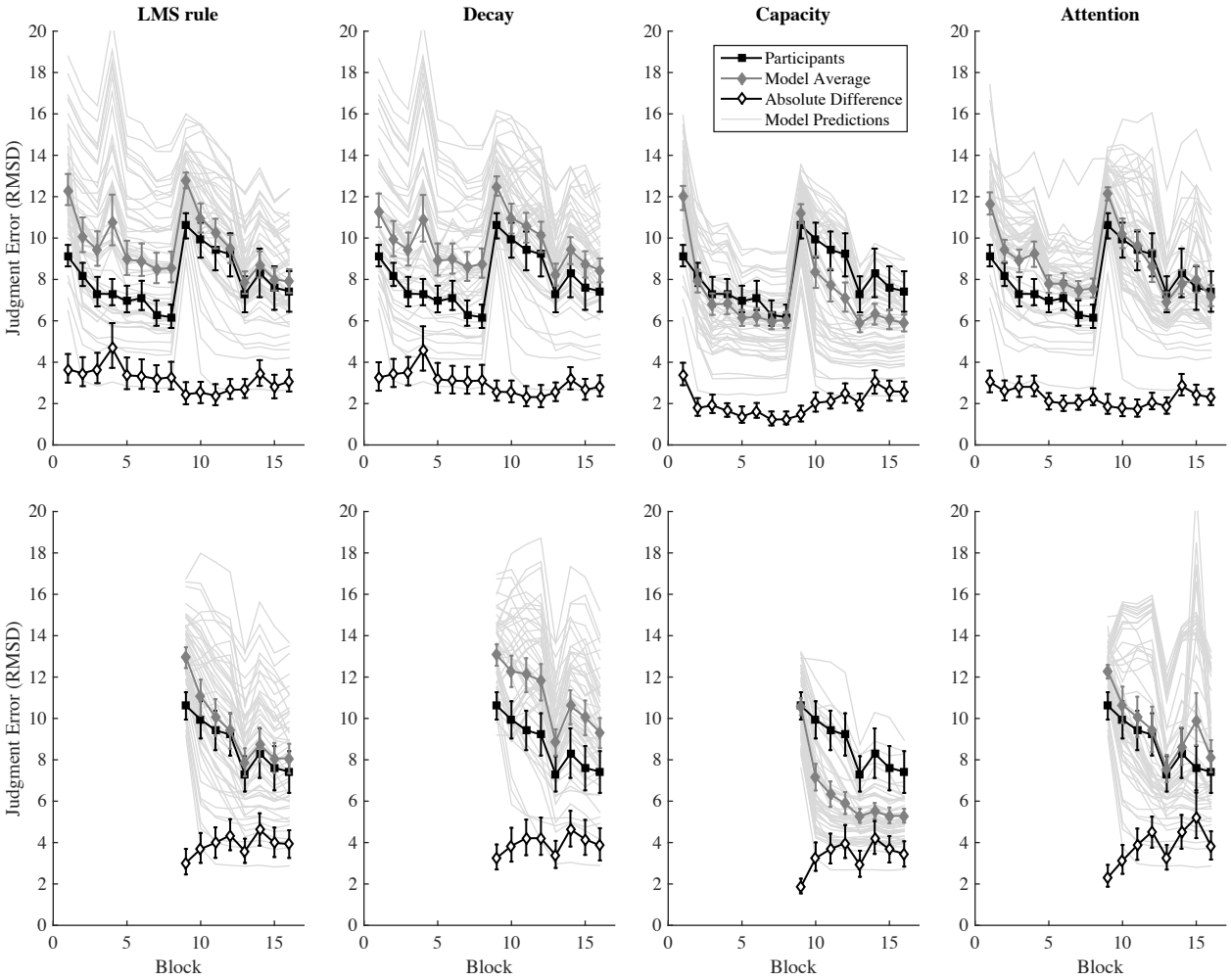

Figure 3. Judgment error (Root Mean Square Deviations, RMSD) averaged across all participants (black lines) and judgment error predicted on average by each model (dark grey lines) in each learning block. Filled white diamonds illustrate the absolute difference between both learning paths; light grey lines depict model predictions for each single participant. Columns show judgment error separately for each model (LMS rule, Decay, Capacity, Attention). The upper row shows predicted judgment error when model parameters were estimated using all learning trials; the lower row shows predicted judgment errors when model parameters were estimated based the first 200 learning trials and used to predict the learning path in the second half of the experiment. Error bars indicate bootstrapped confidence intervals.

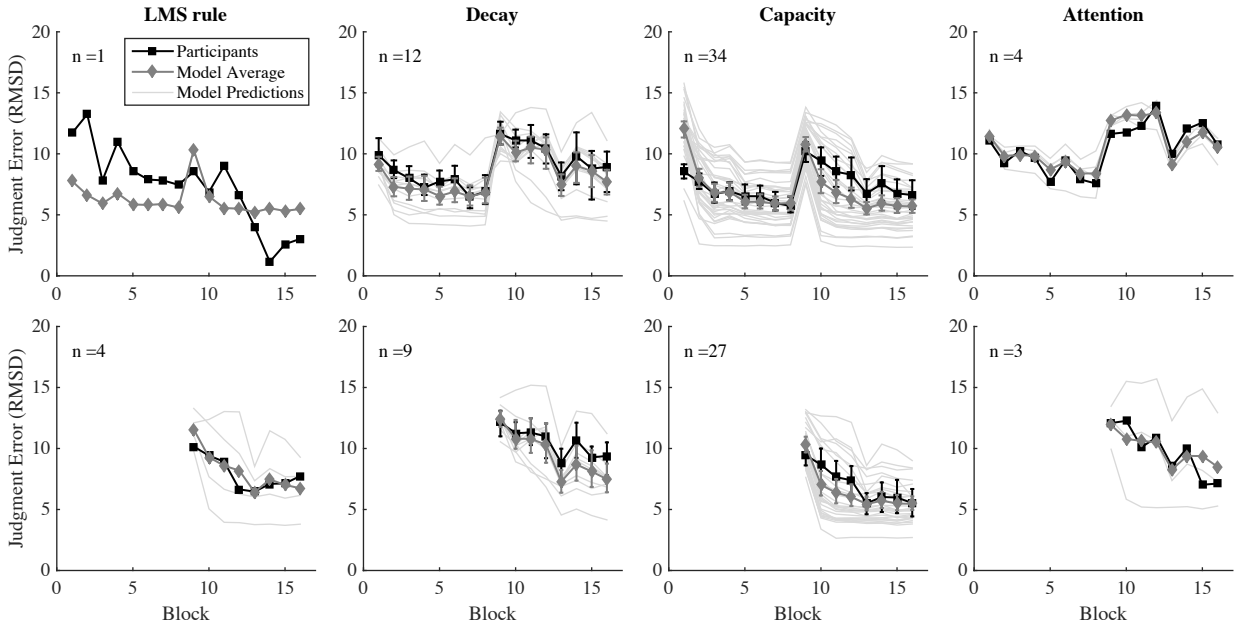

Figure 4. Judgment error in Root Mean Square Deviations (RMSD) for participants classified to each model (black lines) and judgment error predicted by the model on average for those participants (dark gray lines) in each learning block. Light gray lines depict model predictions for each single participant. Columns show judgment error separately for each model (LMS rule, Decay, Capacity, Attention), rows show predicted judgment error separately for each fit indicator (Upper row: BIC, Lower row: Generalization). Error bars indicate bootstrapped confidence intervals.

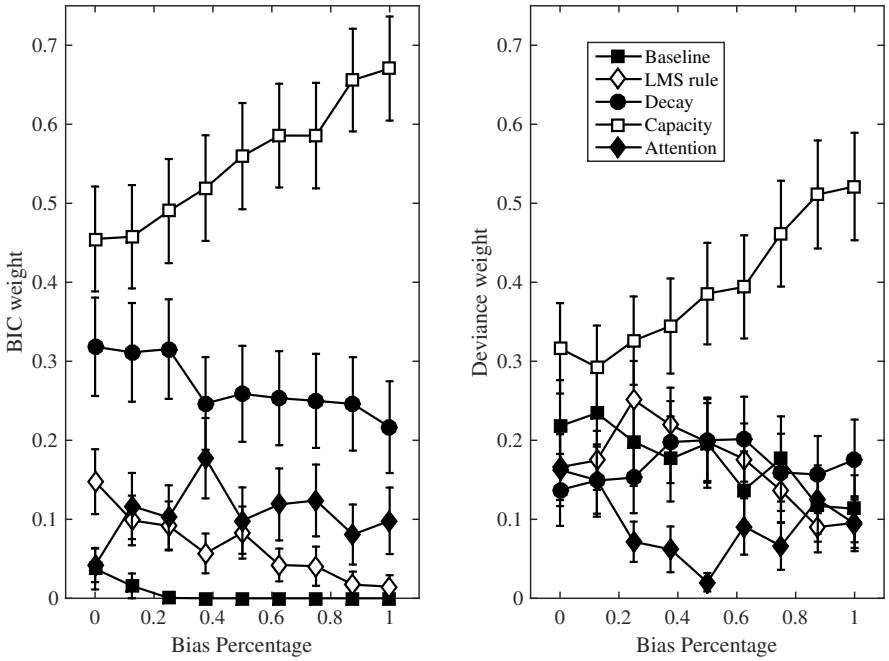

Figure 5. BIC and deviance weights for each model (separate lines) as a function of the starting value for the cue weights (expressed in percentage of bias). The left panel shows BIC weights from estimating the models based on all learning trials. The right panel shows deviance weights from generalization for the second half of the experiment. Error bars show ± 2 standard errors.

Appendix A

Model estimation and model comparison

To evaluate the models' relative performance, we employed two different model fit indicators that vary in the degree to which they consider model generalizability and model complexity: the Bayesian Information Criterion (BIC; Schwarz, 1978) as well as a generalization test (Busemeyer & Wang, 2000). Both techniques can be used to compare non-nested models, but consider different sources of model flexibility. The BIC penalizes more complex models by accounting for the number of free parameters, but does not account for model complexity in terms of the functional form. In contrast, the generalization test is an indicator of projective fit and assesses to what degree the models' performance also generalizes to a range of new items or a new condition. In doing so, it implicitly accounts for both model complexity in terms of the number of parameters as well as functional form.

All models were fitted to participants' responses by minimizing the deviance -2LL, the negative summed log-likelihood L of the model given the data.

$$-2LL = -2 \cdot \sum \ln(L) \tag{12}$$

We calculated the likelihood as the probability density of participants' judgments j assuming a truncated normal distribution, with the models' predicted responses $\hat{c_p}$ as the mean of the normal distribution and a fitted standard deviation σ. This truncated normal distribution was chosen because it matched the response scale from 0 to 50.

$$L = \frac{1}{\sigma} \frac{\phi(j|\hat{c}_p, \sigma)}{\Phi(50|\hat{c}_p, \sigma) - \Phi(0|\hat{c}_p, \sigma)} \tag{13}$$

Bayesian Information Criterion

To calculate the BIC for each model, we estimated parameters of all learning models based on all training trials for the reanalysis. In the relearning study, we estimated each models' parameters using all trials in the experiment. The BIC was then calculated from

each models' deviance penalized with the number of free model parameters k:

$$\text{BIC} = -2LL + k \ln n, \tag{14}$$

where n denotes the number of observations. Smaller BIC values indicate a better model fit. BICs were converted into BIC weights $\text{BIC}_{w,M}$ that give the posterior probability of each model given the data (Wagenmakers & Farrell, 2004).

$$\text{BIC}w_M = \frac{e^{-.5\Delta\text{BIC}_M}}{\sum_i e^{-}.5\Delta\text{BIC}_i} \tag{15}$$

with Δ BIC_M as the difference between model M and the best model in the set and Δ BIC_i as the difference between a specific model i the best model.

Model fit measured in RMSD was calculated as the RMSD between model predictions and participants' judgments for the complete learning sequence. Model predictions were constrained to the range of the scale from 0 to 50. To derive model predictions for each learning block, we included for each participant a truncated normally distributed random error matching the standard deviation from fitting and generated the model predictions 100 times. We then calculated the RMSD for each learning block and simulation and averaged across the simulations, separately for each learning block.

Generalization Test

To account for model flexibility introduced by the functional form and to test generalizability to new items and conditions, we also conducted a generalization test (Busemeyer & Wang, 2000). Specifically, in the reanalysis we used the regression weights obtained from model fitting at the end of the training phase to generate model predictions for validation items. In the relearning study, we estimated each models' parameters on the first half of the learning blocks (before changing the task structure) and predicted participants' learning performance in the second half of the learning blocks (after the task structure changed). In accordance with the BIC weights, we computed deviance weights to classify participants to each model. The reported overall RMSD was calculated as the

RMSD between model predictions and participants' judgments for the second half of the experiment. Model predictions were truncated to match the range of the scale. To derive thze predicted RMSD for each learning block in the validation trials, we generated model predictions for all validation trials a 100 times using the estimated standard deviation from each model. The predicted RMSD for each learning block was calculated separately for each learning block and simulation and averaged across the simulations.

Appendix B

Model parameters for the reanalysis and the relearning experiment

Appendix B lists the estimated mean parameter values for the reanalysis (Table B1) and the relearning experiment (Table B2) with standard deviations in parentheses. Parameter estimates for the reanalysis were estimated based on all training trials. In the reanalysis, parameter estimates for the BIC were estimated based on all trials in the experiment (cf. Appendix A). Parameter estimates for the generalization test were estimated based on the first 200 trials experiment.

Table B1

Model Parameter in the Reanalysis. SD in Parentheses

Model	Parameter					
	λ	δ	r	λ_α	θ	*SD*
Baseline	0.082 (0.077)	–	–	–	–	6.6 (0.9)
LMS rule	0.008 (0.006)	–	–	–	–	6.2 (1.9)
Decay	0.043 (0.091)	11.7 (25.4)	–	–	–	6.2 (1.9)
Capacity	0.019 (0.014)	–	17.2 (5.1)	–	–	4.7 (1.1)
Attention	0.446 (0.214)	–	–	0.042 (0.112)	7.6 (23.0)	5.1 (1.0)

Note. Note. λ = Learning rate; δ = Decay rate; r = Capacity restriction; λ_α = Learning rate for attention weights; θ = Sensitivity to attentional strength; *SD* = Standard Deviation.

Table B2

Model Parameter in the Relearning Experiment. SD in Parentheses

Criterion	Model	Parameter					
		λ	δ	r	λ_α	θ	*SD*
BIC	Baseline	0.089 (0.095)	–	–	–	–	7.1 (0.7)
	LMS rule	0.01 (0.008)	–	–	–	–	6.2 (1.5)
	Decay	0.034 (0.036)	0.4 (0.5)	–	–	–	6.1 (1.5)
	Capacity	0.012 (0.011)	–	16.6 (5.2)	–	–	5.2 (1.1)
	Attention	0.339 (0.23)	–	–	0.034 (0.143)	11 (29.2)	5.6 (0.9)
Generalization	Baseline	0.119 (0.109)	–	–	–	–	6.8 (0.7)
	LMS rule	0.009 (0.007)	–	–	–	–	6.2 (1.7)
	Decay	0.036 (0.049)	2.4 (13)	–	–	–	6.1 (1.7)
	Capacity	0.018 (0.013)	–	18.2 (7.7)	–	–	4.9 (1.0)
	Attention	0.419 (0.287)	–	–	0.120 (0.254)	11 (26.5)	5.6 (1.1)

Note. BIC = Bayesian Information Criterion; λ = Learning rate; δ = Decay rate; r = Capacity restriction; λ_α = Learning rate for attention weights; θ = Sensitivity to attentional strength; *SD* = Standard Deviation.