Testing learning mechanisms of rule-based judgment

Janina A. Hoffmann
University of Bath
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 Bath, 10W Claverton Campus, Bath, BA27AY, United Kingdom. E-mail: jah253@bath.ac.uk

We thank Maarten Speekenbrink and two anonymous reviewers for their comments on earlier versions of this manuscript.

Data and material associated with the manuscript are made available via https://osf.io/w7f4u/.

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 described with linear models. How people learn to make judgments has received less attention. Although the hitherto proposed delta learning rule can perfectly learn to solve linear problems, reanalyzing a previous experiment shows 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. An additional study tested the different learning mechanisms in predicting learning in linear judgment tasks. In this 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 environment changed, but improved again until the end of the task. The capacity-restricted learning model, restricting how much people update the cue weights on a single trial, best described and predicted the learning curve of the majority of participants. Taken together, these results suggest that considering cognitive constraints within learning models may help to understand how humans learn when making inferences.

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 how long the car will still run, whereas the time since its last cleaning may be less prognostic. 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). Consequently, the notion that people preferably weigh and add information has inspired theories of information processing across a broad range of research areas 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 (Juslin, Karlsson, & Olsson, 2008). Within a trial-by-trial updating process these changes in each cues' importance can be described within a simple and widely used learning model, the least mean squares (LMS) or delta-learning rule (Gluck & Bower, 1988; Sutton & Barto, 1981). Yet, although the LMS rule has been successfully applied to a variety of learning problems (Gluck & Bower, 1988; Schultz & Dickinson, 2000; Siegel & Allan, 1996), it neither describes learning of judgment problems adequately, nor does it successfully predict how people weigh the importance of different cues, as we demonstrate below. In addition, psychological mechanisms that may limit this rule-based learning process have rarely been spelled out in detail (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 people may learn the cue weights in linear judgment problems within the class of LMS models. To this goal, we examined how the LMS 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 LMS 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 $\hat{j}_t$ on each trial t as a linear, additive function of the cue values $x_{t,i}$ weighted by their importance, the cue weights $w_{t,i}$, which can be mathematically modeled by a linear regression.

$$\hat{j}_t = \sum_i w_{t,i} \cdot x_{t,i} \tag{1}$$

with $x_{t,*} = \begin{bmatrix} x_{t,1} & \dots & x_{t,n} & 1 \end{bmatrix}$ 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.

In principle, people can learn different kind of cue-criterion relations. In the current manuscript we focus on how people learn the validities of multiple uncorrelated linear predictors. We take this focus because research shows that people can abstract linear rules 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). The cue weights implied by linear rules also 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, although people are able to learn more complex cue-criterion relationships such as quadratic or even cyclic functions (Bott & Heit, 2004) or configural cue-criterion patterns (Mellers, 1980), people show a preference for learning positive linear functions and have a hard time picking up interactions between cues (Brehmer, 1994; Lucas, Griffiths, Williams, & Kalish, 2015). Thus, before successful learning in more complex tasks is tackled it seems important to understand how cue weights are learnt in a linear task. We will consider potential extensions to nonlinear tasks in the general discussion.

In linear additive judgments tasks, 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 t, the judgment is made based on a linear regression model (Equation 1). After each trial t, the cue weights are updated for the next trial $t+1$ depending on how much the judgment $\hat{j}_t$ deviates from feedback y_t. The more the judgment deviates from the correct judgment and the higher the learning rate λ is, the more strongly the cue weights should change in the next trial. Changes in the cue weights $\Delta w_{t,i}$ are more strongly attributed to those cues with higher cue values, that is here akin to more salient features.[1]

[1] In the present tasks all cue values covered the same range from 0 to 5 with higher cue values indicating a higher salience based on cue intensity. Care was taken in stimulus design that the single cues are approximately equally salient and that the cue values could be discriminated from each other. Therefore, we assume here that all cues are measured on the same scale. One way to accommodate different scales in learning models would be to standardize all cues, for instance to a range between between 0 and 1. However, different scales imply as well different psychophysical properties of the stimulus, such as differences in salience, and may impact on learning. How different scales interact with learning in judgment tasks shall be addressed in future work.

$$\Delta w_{t,i} = \lambda \cdot x_{t,i} \cdot (y_t - \hat{j}_t) \tag{2}$$

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

$$w_{t+1,i} = w_{t,i} + \Delta w_{t,i} \tag{3}$$

The LMS rule is formally similar 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-based learning (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 how well the LMS rule can describe learning, 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), without providing a description of the underlying learning mechanism. In a rolling regression, a linear regression model is repeatedly estimated for a fixed number of judgments starting from the first to the m^{th} learning trial (where m 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 participants' 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 model's learning rate and standard deviation were estimated by minimizing the negative log-likelihood between participants' judgments and model predictions over all trials in the training phase (for details on model estimation and comparison see Appendix B, for model parameters see Appendix C). 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 considered trials from trial 50 onwards because the rolling regression estimates only one set of cue weights for trials 1-50 so that one cannot compare changes in cue weights and model fit for early trials.

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 = 12900 (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 [3.1,3.8]. 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 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.[2] Similar to the training results, the LMS rule captured judgments for unknown items better than the baseline model, V = 5362, $p < 0.001$, 95% CI [-5.0,-3.2], but still did not outperform the predictive performance of the rolling regression, V = 40772, $p < 0.001$, 95% CI [3.2,4.0]. 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; Busemeyer, Myung, & McDaniel, 1993b) indicating that learning rates for one cue depend on the existence of another cue.

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

[2] 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.

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 environment (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)

$$\Delta w_{t,i} = \frac{\lambda \cdot x_{t,i} \cdot (y_t - \hat{j}_t)}{t^{\delta}} \tag{4}$$

Parameter δ controls the decay rate with $\delta > 0$ and thus the decay model possesses the parameters λ and δ. A higher decay rate implies that the learning rate more strongly declines with a higher number of learning trials. Indeed, it has been shown in some tasks that including a decay parameter provides a better description of the learning process than the pure 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, findings that the cues' importance is often adjusted relative to the importance of all other cues as well point towards the idea that learning is restricted by a cognitive capacity limitation (Birnbaum, 1976; Busemeyer et al., 1993a; Busemeyer et al., 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 (a cue competition effect). 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 these cue competition effect because they gradually converge to the optimal weights (Busemeyer et al., 1993a). Instead, according to Busemeyer et al. (1993a), models predicting cue competition effects need to impose a capacity constraint on the weights (but see Speekenbrink & Shanks, 2010, for a more recently developed model that predicts cue competition without a capacity constraint). Importantly, this capacity restriction implies that if the capacity restriction is reached, the cue weights are adjusted relative to the importance of all other cues (Birnbaum, 1976; Busemeyer et al., 1993a).

To our knowledge, past research has not yet specified, nor tested a rule-based learning model for human judgment spelling out how this cognitive capacity restriction affects knowledge updating. We implemented capacity restricted learning in our model by restricting the updated cue weights to sum up to a capacity restriction r, $\sum_i |w_{t+1,i}| \leq r$ (Busemeyer et al., 1993a).[3] In each trial, the model updates the cue weights first using the LMS rule, resulting in the unconstrained updated cue weights, $v_{t,i}$. In case, the capacity restriction is not reached, the unconstrained cue weights are used in the next trial, $w_{t+1,i} = v_{t,i}$. In case, the capacity limit is reached, the cues are adjusted by the difference between the summed cue weights still considered important after the update (i.e., they are larger than 0) and the capacity restriction, divided by the number of non-zero weights p (see Duchi, Shalev-Shwartz, Singer, & Chandra, 2008, and Appendix A for details).

$$\beta_{t,i} = \left[|v_{t,i}| - \frac{1}{p} * \left(\sum_p |v_{t,p}| - r\right)\right]^+ \tag{5}$$

where $[x]^+$ is x for $x > 0$, else 0. Thus, the capacity model estimates a parameter for the learning rate λ and the capacity constraint r. In a last step, the restricted weights $\beta_{t,i}$ are multiplied by the sign function so that each non-zero weight keeps its initial direction.

$$w_{t+1,i} = \text{sgn}(v_{t,i})\beta_{t,i} \tag{6}$$

Consider, for instance, a case in which capacity is restricted to $r = 1$ and the updated cue weights are $v_{t,*} = [0.5\ 0.5\ 0.25\ 0.25]$. In this example, each cue weight will be reduced by $\frac{1}{4}$ (1.5 - 1) = 0.125 and $w_{t+1,*}$ =[0.375 0.375 0.125 0.125]. 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. Consider another case, in which one cue strongly overshoots the capacity limit, for instance $v_{t,*} = [2\ 0.25\ 0.25\ 0.25]$. Here, the smaller cue weights are set to zero to adhere to the capacity restriction and only the highest cue weight determines the reduction, (2 - 1) = 1, resulting in $w_{t+1,*}$=[1 0 0 0]. Thus, setting a restriction on the absolute sum of weights can lead to ignoring the least important cues.

[3] Restricting the absolute sum of cue weights transforms the linear regression model into a LASSO regression.

Accordingly, the capacity limit reduces how much cue weights are updated, which can give 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 in response to high errors. For instance, consider a trial in which a participant's judgment (e.g. 10) was very far off from the correct response (e.g. 45). The LMS rule (similar to all other models proposed here) assumes that the participant should strongly adjust all cue weights in the direction of the error. This adjustment may in some trials lead astray and the adjusted cue weights in the next trial are further away from the correct weights than the original ones causing a decline in judgment accuracy. In contrast, the capacity restriction limits this adjustment and therefore learning performance is more stable even for high learning rates. However, learning proceeds similarly as in the LMS rule, if the capacity limit strongly exceeds the optimal sum of weights.

LMS rule with attention learning (Attention)

Attentional mechanisms allow us to selectively process information and to prioritize certain pieces of information, while ignoring less relevant cues (Awh, Vogel, & Oh, 2006; Le Pelley et al., 2016). 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). For instance, evidence from the related field of category learning suggests that measures of attention such as eye movements reflect the importance of cues in categorization decisions (Beesley, Nguyen, Pearson, & Le Pelley, 2015; Hoffman & Rehder, 2010; Rehder & Hoffman, 2005). 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).

Here, we implemented an attentional learning model that adapts the learning rate for each cue weight over time by adjusting the cue weights proportionally to the correlation between the current change in cue weights and recent changes in cue weights —a gain adaptation learning mechanism (Sutton, 1992). As in the LMS rule, the judgment is in every trial a linear additive function of the cue values and the cue weights, but instead of assuming a global learning rate for all cues, the model considers a separate learning rate $\alpha_{t,i}$ for every cue i which can be learned over trials:

$$w_{t+1,i} = w_{t,i} + \alpha_{t,i} \cdot x_{t,i} \cdot (y_t - \hat{j}_t) \tag{7}$$

with $\alpha_{t,i} = e^{b_{t,i}}$ to keep the learning rates strictly in a positive range. Accordingly, each cue weight is learned with a different learning speed. Specifically, these learning rates are adjusted before any cue weight is updated by considering the global learning rate λ, the salience of the cues, and a decaying memory trace, $h_{t,i}$, that stores previous weight updates:

$$b_{t,i} = b_{t-1,i} + \lambda \cdot x_{t,i} \cdot (y_t - \hat{j}_t) \cdot h_{t,i} \tag{8}$$

with an initial learning rate for the single cues b_0 and the global learning rate λ estimated as free parameters. Thus, the model updates knowledge faster about salient cues, that is cues with high cue values, and cues that recently changed their weights, reflecting that those cues may have been important to consider in recent trials. This stored memory trace decays from the trial t to the next trial $t + 1$ depending on the current change in learning rate (Sutton, 1992). The degenerated trace $h_{t,i}$ is then updated with the current change in cue weights, $\alpha_{t,i} \cdot x_{t,i} \cdot (y_t - \hat{j}_t)$, and, hence, the memory trace accumulates recent changes in cue weights over time.

$$h_{t+1,i} = h_{t,i}\left[1 - \alpha_{t,i} \cdot x_{t,i}^2\right]^+ + \alpha_{t,i} \cdot x_{t,i} \cdot (y_t - \hat{j}_t) \tag{9}$$

Accordingly, how strongly the learning rate changes for a single cue in one trial depends on the visual salience of the cue, judgment error on this trial, and the accumulated knowledge about cues recently changing in importance.[4]

In sum, the attention model postulates that attention, as modelled by differences in learning speed for different trials, modulates how fast people update their knowledge about the cue weights. A higher learning rate reflects a higher attention towards this cue, implying a stronger update of the cue weights. Attention flows towards errors caused by salient and recently changing cues. As a result, if attention plays a major role during relearning, people should adapt faster to changes in previously important cues and for cues strongly changing their predictive weight.

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 participants' judgments in the experiment, we compared the models with two model comparison criteria: the Bayesian Information Criterion (BIC, Schwarz, 1978) and a generalization test (Busemeyer & Wang, 2000, see Appendix B for a more detailed description). Whereas the BIC penalizes more complex models by accounting for the

[4] Note that the product $\lambda x_{t,i} \cdot (y_t - \hat{j}_t)$ reflects the update for each cue without assuming differences in learning speed. As a result, the change in learning rate is proportional to the correlation between the current change in cue weights and the accumulated recent changes in cue weights (Sutton, 1992).

number of free parameters[5], the generalization test measures to what degree the models can also predict independent, unseen data, thereby taking the functional complexity of the models into account. To calculate the BIC, we estimated each model's parameter values based on all training trials using maximum likelihood estimation. Based on the BIC, we then derived the corresponding Bayesian Information weights, BIC_w, that yield the probability that a given model is the best model within the model set given the data, that is, the one that minimizes the Kullback-Leibler discrepancy (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 for the decay model and the LMS model with the capacity model reaching the lowest BIC and 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 the model with the best BIC 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 best.

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

[5] Using the Akaike's information criterion (AIC) instead of the BIC does not yield different results. Therefore, we report only the BIC here.

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 model 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 additional 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, learning models predict how people should adapt their judgment policy in a dynamic environment in which the cues' importance for predicting a criterion changes. The decay model predicts that people should adapt more slowly to a 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 are highly salient, were previously relevant or implied a recent change in cue weights. The learning models hence allow fine-grained predictions about how people should change their judgment policy, if the underlying environment 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 correct criteria 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. Each cue could take on six different ordered cue values ranging from 0 (e.g. a non-visible nose) to 5 (a very large nose). 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 environment 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 combination, we used in a first step the median parameters from the reanalysis to predict learning performance as well as the adaptation to a new task environment across a range of task combinations. 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 eighth block), the task environment changed and participants had to relearn the importance of the cues. Participants were explicitly informed in the beginning of the experiment about a potential change in the task. Yet they were not informed when the change would happen but had to infer that the change occurred from the feedback they received. Specifically, the instructions highlighted that "how you could best solve the judgment task may change during the experiment." After introducing the comic figures and their features, participants were again told that "how you could use the features [of the comic figure] to judge the comic figure may change over time".

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, if participants reached more than 80% of the points in the last two learning blocks of each task (learning block 8 and 16), they received an additional bonus of 3 CHF. Participants were informed that we selected two learning blocks in advance for paying out the bonus, but we did not tell them which blocks we chose.

Results

Learning performance. The learning performance suggested that participants on average adapted to the change in the task environment (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 environment 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 environment. Compared to the first eight blocks of the experiments, judgment performance varied more strongly between participants after the task environment 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 of the task environment 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 participants' judgments over the experiment, we compared each model's performance based on the BIC and based on the generalization test (see Appendix B 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 C 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 first estimated each models' parameters based on participants' judgments in the first 200 trials. In a next step, we used the obtained parameters to predict participants' judgments for all learning trials, that is the learning models continued to learn in the second half of the experiment. As a measure of model fit, we determined the deviance D based on the predictions for the second half of the experiment. 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).

Similar to the results of 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 environment? 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 environment, but now the decay model fares worse than the LMS rule. More importantly, the relative advantage of the attention model vanishes in generalization. Specifically, the D of the attention model is similar to the D of the LMS rule and the decay 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 23 of the 51 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.[6]

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 (blue lines) as well as the average judgment error predicted by each model (black lines, in columns), separately for model estimation and generalization (in rows). Red 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 strongly decreases in later learning blocks (Figure 3, upper row). In contrast, the capacity model captures quite well the average learning path in most learning blocks, but makes slightly more errors on average in the first and last learning blocks.

Focusing more on the variation in individual model predictions, the graphs based on estimation illustrate that predicted judgment error is more variable for the LMS rule and

[6] For 12 participants the attention model generated extremely high D that we restricted to D = 10'000. In addition, model-implied cue weights in generalization were bound to the scale between -10 and 10.

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 environment 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 environment changed, judgment error does not much decrease even after several learning blocks. In addition, both models propose that judgment error strongly increases in a few learning blocks (e.g. in learning block 4) —a pattern that is neither as pronounced in participants (compare Figure 2) nor picked up by the capacity model. Finally, the attention model describes a slower learning path in the first learning phase compared to the capacity model as well as participants' learning path. Furthermore, similar to the decay model it allows a strong increase in judgment error even after several learning blocks in the relearning phase (learning block 14 and 15). Although judgment error also increases for a few participants in learning block 14, the increase in block 15 is less pronounced in participants.

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 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 large amount of judgment errors. Finally, the attention model predicts larger judgment errors in the last 200 trials and the absolute difference between learning paths suggests a rather strong increase in mismatch. Particularly, the attention model emphasizes that judgment error may strongly increase for some participants and predicts for a substantial number of participants larger judgment errors late in learning (particularly in block 15). 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 reported above, 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 (25.5 %), the attention model (5.9 %), or the LMS rule (2.0 %). A classification based on the D_w suggested similarly that most participants (45.1 %) were classified as best predicted by the capacity model. Further, some participants were classified as best predicted by the decay model (15.7 %), the LMS rule (9.8 %) or the attention model (13.7 %), 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 (blue lines, in columns) as well as the average judgment error predicted by each model (black 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 environment. 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 problems in adapting to the change in task environment, 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 environment changed, the capacity model predicts a more successful learning path for most participants.

Figure 5 illustrates how the model-implied cue weights change over the course of the experiment compared to cue weights estimated from a rolling regression. Inferred cue weights were predicted based on the learning parameters estimated on the first 200 learning trials. The upper row shows implied average cue weights for all participants, the lower row shows implied cue weights only for those participants best predicted by each model. This graph suggests that on average the decay model predicts for every cue a rather slow change in cue weights and particularly underestimates how fast participants relearn to assign a higher importance to cue 3 and cue 4. In comparison, the capacity model overestimates how fast participants react to a change in the environment, predicting an almost immediate detection of the most important cue in the second phase of the experiment and a quick suppression of the no longer predictive cues. Finally, the attention model relearns cue 4 slightly more slowly than the capacity model, but similarly overestimates its importance.

In addition, it inhibits more strongly the previously important cues than the capacity model does.

Taken together, the capacity model best described and predicted how most participants learned to adapt their judgments over trials suggesting on average a steady adaptation to the change in task environment. 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.

Does a combination of psychological mechanisms outperform the single process models? It is possible that the psychological mechanisms we proposed do not act in isolation, but a learning model incorporating, for instance, attentional learning as well as a decay mechanism may outperform the isolated learning mechanisms. To account for this possibility, we formulated learning models combining two out of the three proposed psychological mechanisms. We also specified one "full model" relying on decay, a capacity constraint, as well as attentional learning (for the specification of the full model see Appendix D, for BICs and D see Appendix E). Next, we calculated how likely each model is the best model within this candidate set in terms of BIC_w or D_w. Overall, this comparison suggested that a model combining a capacity restriction and a decay possesses a higher $\text{BIC}_w = 0.62$ ($SD = 0.46$) than a model assuming attentional learning and decay ($\text{BIC}_w = 0.06$, $SD = 0.22$), a model assuming attentional learning and a capacity restriction ($\text{BIC}_w = 0.22$, $SD = 0.38$) or the full model ($\text{BIC}_w = 0.1$, $SD = 0.28$). Similarly, D_w were slightly higher for the model combining a capacity restriction and decay ($D_w = 0.37$, $SD = 0.48$) than for a model assuming attentional learning and decay ($D_w = 0.29$, $SD = 0.46$), a model assuming attentional learning and a capacity restriction ($D_w = 0.1$, $SD = 0.29$), or the full model ($D_w = 0.24$, $SD = 0.42$). Overall, this result suggests that attentional learning only plays a minor role in learning compared to a capacity restriction and a decay in learning speed.

Does a combination of capacity-restricted learning and a decay in learning speed describe and predict participant's judgments better than all of the single mechanisms alone? Thus, in a second step, we estimated to what degree the combined model is preferred compared to the isolated mechanisms using BIC_w or D_w. Overall, this analyses yielded mixed results. In describing participants' judgments across all experimental trials, the combined model ($\text{BIC}_w = 0.6$, $SD = 0.44$) outperformed the decay model ($\text{BIC}_w = 0.2$, $SD = 0.36$), the capacity model, ($\text{BIC}_w = 0.15$, $SD = 0.32$), and the attention model ($\text{BIC}_w = 0.05$, $SD = 0.2$). In generalization, however, the capacity model ($D_w = 0.48$, $SD = 0.5$) better predicted judgments compared to the decay model ($D_w = 0.27$, $SD = 0.39$), the attention model ($D_w = 0.13$, $SD = 0.33$), or the combined model ($D_w = 0.13$, $SD = 0.27$). One potential reason why the complex model does not fare too well in predicting new judgments is that those models may adjust to flexibly to participants' learning in the first phase and are, hence, less robust to variations in the learning path in the second phase. Taken together, these results suggests that assuming only capacity-restricted learning is more parsimonious than considering a combination of different psychological mechanisms and might be the more robust model in predicting new independent behavior.

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 the 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.[7] 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 6 displays how the average BIC_w (left panel) and D_w (right panel) for each model (separate lines: Baseline, LMS rule, Decay, Capacity, and Attention model) 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 attention 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.

Contrasting capacity constrained learning with a Bayesian Learning model. The current work mainly focussed on evaluating how well different psychological mechanisms supplementing the LMS rule fare within an updating process that hinges on gradient descent learning. Our work shows that within the family of LMS rule models the capacity-restricted LMS was best in describing the observed learning process.

Alternatively, it has been proposed that updating of the cues could follow a Bayesian

---

[7] We conducted as well a robustness check varying the intercept in the probe, $x_{t,*}=\begin{bmatrix} x_{t,1} & \dots & x_{t,n} & k \end{bmatrix}$ with k =[1, 2, 5]. Results do not suggest that another LMS model systematically outperforms the capacity model if a different value for the intercept is assumed.

learning process (Speekenbrink & Shanks, 2010) or that people may learn the associations between specific patterns of cue and criterion values (Busemeyer, Byun, Delosh, & McDaniel, 1997; Kruschke, 1992). However, it is beyond the focus of the present work to test the capacity-restricted model against a variety of alternative frameworks and learning models. To evaluate the predictive accuracy of the capacity-restricted LMS rule we thought it would be beneficial to compare it to the Bayesian approach. Thus, is a capacity-restricted LMS rule a viable competitor for Bayesian learning models? To compare capacity-restricted learning with Bayesian models, we implemented the Bayesian Linear Filter (Speekenbrink & Shanks, 2010) using a version with 3 parameters (σ_0 for the diagonal elements of the covariance matrix of the prior distribution, σ for the diagonal elements of the covariance matrix of validity changes, and σ_y as the standard deviation around predicted judgments). Comparing the performance of the Bayesian Linear Filter with the capacity model suggests that the Bayesian Linear Filter on average describes the learning process better than the capacity model (BIC = 2693, *SD* = 194 vs. capacity model: BIC = 2715, *SD* = 182), but does not predict as successfully as the Capacity model how people adapt to a change in task environment (D = 1458, *SD* = 165 vs. capacity model: D = 1439, *SD* = 180). Likewise, considering the relative evidence for the Bayesian Linear Filter, as expressed by the BIC_w and D_w, suggests that the Bayesian Linear Filter performs better at describing the learning path (BIC_w = 0.73, *SD* = 0.44), but the evidence does not favor one over the other model in generalization (D_w = 0.53, *SD* = 0.49). More fine-grained analyses of the cue weights indicated that the models predict similar adjustments of the cue weights as a function of learning. For instance, both models predict a rapid change in importance for the previously least important cue (Cue 4), but do not capture the slower learning of the second least important cue (Cue 3). Classifying participants suggested that those participants are better predicted by the capacity model who quickly learn to focus on the most important cue after the task changed and efficiently suppress the cue weights for the least important cues. Taken together, these results suggest that the capacity constrained LMS rule might be a suitable competitor for Bayesian Learning models.

General Discussion

Weighing information according to its importance has been deemed one of the core competences in human judgment. However, how people learn 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 environment (Betsch, Brinkmann, Fiedler, & Breining, 1999; Betsch, Haberstroh, Glöckner, Haar, & Fiedler, 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 additional 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 adding those psychological learning mechanisms better described the learning process than the pure LMS rule. In the reanalysis, all extended LMS learning models predicted judgments more accurately than the pure LMS rule. To tease 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 suggest that considering psychological constraints within optimal learning algorithms may help to better understand how people learn to solve judgment problems and that —within the class of LMS model —a capacity constraint likely best explains the systematic differences between human performance and an optimal LMS learning algorithm. Still, Bayesian learning models may provide an alternative approach.

Capacity restrictions in learning

Why is there a benefit for the capacity model over all other LMS models? The capacity model has been motivated by research on capacity limits in judgment and decision making and cue competition (Birnbaum, 1976; Busemeyer et al., 1993a; Busemeyer et al., 1993b; Juslin et al., 2008). It assumes that a capacity limit restricts how much importance people assign to the cues. This slows down updating of the cue weights in each single step 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, setting weights to zero for which this substraction would change the predictive direction. This implies that people will have problems to track cue weights, if the number of cues increases. One prediction that follows from this, is, for instance, that problems with one predictive cue should be learned at a faster speed than problems with five predictive cues if the learning rate is high.

If the capacity restriction does not allow to consider all predictive cues, however, strict capacity limitations may change from trial to trial which cue is considered most important. Consequently, a lower capacity limit can induce a more inconsistent weighting of the cues and thereby decrease judgment consistency – a finding in line with previous work suggesting that constraints in working memory capacity may limit how consistently people pursue a strategy at the end of learning (Hoffmann et al., 2014). Finally, 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 both cues are affected by cue competition to the same degree (a symmetric effect). Yet, 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. Further research may specify the conditions under which cue competition effects are likely to occur in judgment.

Interestingly, the model proposes that the capacity restriction may possess an adaptive value. Specifically, in the LMS rule high learning rates are not always beneficial because the cue weights can strongly change in a single trial. In our study, this overadaptation led the LMS rule to predict high error in some blocks, whereas the majority of participants did not adjust their predictions to the same degree. A capacity limit close to the optimal sum of weights limits the possibility that people update the cue weights too strongly, which in turn prevents overly high errors during training and as a result may also enable higher learning rates.

Does the capacity model capture judgment phenomena beyond learning of linear, additive functions with positive cue-criterion relationships? All LMS models considered here are limited to learning linear, additive functions and do not postulate in advance any prior relationship between the cues and the criterion. As a result, the capacity model would learn functions with a negative correlation between each cue and the criterion at the same speed as positive relationships. However, past research has shown that participants solve judgment problems including only positive relationships between the cues and the criterion faster than judgment problems with negative or mixed relationships (Rolison et al., 2012). One way to capture this phenomenon, could be to assume positive starting values of the cue weights or to make more sophisticated assumptions about the prior beliefs people may have regarding the cue-criterion relations (for recent approaches to this problem see Lucas et al., 2015; Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017).

In addition, the current version of the capacity model only considers linear additive relationships. In the present task this posed no problem, as the task we used did not require learning nonlinear effects. Although people are usually biased to learn positive linear functions (for a review Lucas et al., 2015), they are able to learn more complex relationships (Bott & Heit, 2004; Mellers, 1980). Thus, it seems important to consider how our model could be extended to describe learning processes of nonlinear relationships between cues and a criterion. One possibility would be to include nonlinear terms (for instance quadratic or multiplicative ones) as further predictors. In this case, centering and scaling these predictors should be considered because the capacity limit in combination with high correlations between linear and nonlinear predictors may punish nonlinear predictors too harshly. Thus, how a capacity limit may influence the ability to learn nonlinear effects would be an important area for future research.

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 they learned the original task (Betsch et al., 1999; Betsch et al., 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 environment. Whereas some people rather quickly adapted to the task environment, 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 why decay only played a minor role in our experiment could be that we informed participants about a potential change in the task environment and introduced a rather salient shift in the cues' importance. This shift in task environment 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 occurred. Accordingly, including a mechanism 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 associative learning (Denton & Kruschke, 2006; Kruschke et al., 2005; Le Pelley et al., 2016) and past research in fields closely related to judgment research has repeatedly provided evidence for the idea that attention influences learning processes (Le Pelley et al., 2016). Furthermore, attentional shifts have been used to explain learning phenomena such as blocking (Kruschke et al., 2005), overshadowing (Denton & Kruschke, 2006) and they may also explain cue competition (Kruschke, 2001). Thus, attentional learning seemed a promising candidate to explain how people deviate from optimal learning. Here, we implemented an attentional learning mechanism that adjusts the learning rate for each cue weight over time, thereby considering changes in recent trials and difference in visual salience (Sutton, 1992).

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 also did not perform well in the generalization test indicating strong overfitting. These misfits are potentially caused by an overly strong tuning of the learning rates to the judgment problem and may result in an inability to predict future judgments. Possibly, limiting the learning rates 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 feedback. 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. We took this approach to (1) be able to investigate learning of cue validities in a task where people have been shown to learn the appropriate function and (2) to identify the psychological mechanisms that need to be assumed to explain how people learn cue weights within a single modelling framework.

However, further successful frameworks that describe how people learn to make judgments exist. For one, 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 our study, the Bayesian Linear Filter on average successfully described and predicted human judgments, too. Comparing the Bayesian Linear Filter to the capacity model suggested that the Bayesian Linear Filter fares somewhat better than the capacity model in describing participants' judgments. However, both models were equally successful at predicting generalization performance, yielding similar predictions for how people should weigh the cues within our relearning design. These results suggest that the capacity model may be a suitable competitor to further contrast with different updating rules. Moreover, the capacity model specifies a mechanism to explain why the weights people assign to cues may deviate from optimal cue validities, which is grounded in psychological theory.

Assuming a psychological mechanism such as the capacity limit has the advantage that it allows predicting under which conditions learning of cue validities depends on further factors such as the number of predictive cues. In contrast, it seems less clear how a similar mechanism could be incorporated into a Bayesian learning model. Thus, one way to tease the models apart could be to vary how difficult it is to detect the predictive cues.

Besides the Bayesian approach, association-based approaches might also do well in describing the learning process for judgment tasks. In particular, it has been argued that people may learn the associations between specific patterns of cue and criterion values, instead of solving judgment tasks by relying on explicit linear rules. In this vein, the associative learning model (ALM, Busemeyer et al., 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). Nevertheless, our current work cannot and did not aim to rule out these models, but identifies a more promising competitor model than the original LMS rule within the framework of LMS models. Beyond the question which of those frameworks describes learning processes best and in which task environments, future research may seek to embed similar psychological constraints across different frameworks to test their psychological plausibility.

Lastly, recent approaches to function learning have suggested that rule- and similarity-based approaches can be integrated within a Gaussian process model and have highlighted inductive biases people have for different types of functions (Lucas et al., 2015) as well as the compositional nature of human function learning (Schulz et al., 2017). This work has made an important contribution to understanding how people can construct highly complex representations from simple basis functions, but did not focus on distinct learning mechanisms. Thus, understanding how cue weights are learned and the psychological mechanisms describing the learning process may complement this work.

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 within learning models may inform our understanding of human behavior.

References

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

Awh, E., Vogel, E. K., & Oh, S. H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208. doi:10.1016/j.neuroscience.2005.08.023

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), 417–429. doi:10.2307/1421615

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38–50. doi:10.1037/0278-7393.30.1.38

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*, *32*(4), 904–918. 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*, 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

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* (pp. 272–279).

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), 227–247. 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

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations.* New York, NY: CRC Press.

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), 319–340. 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

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

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychological Bulletin & Review*, *22*, 1193–1215.

Mellers, B. A. (1980). Configurality in multiple-cue probability learning. *American Journal of Psychology*, *93*(3), 429–443. doi:10.2307/1422722

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), 660–663. 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

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

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

Shalev-Shwartz, S., & Singer, Y. (2006). Efficient learning of label ranking by soft projections onto polyhedra. *The Journal of Machine Learning Research*, *7*, 1567–1599.

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), 266–298. 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. (1992). Adapting bias by gradient descent: an incremental version of delta-bar-delta. In *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 171–176). MIT Press.

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), 301–310. 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 Parentheses

Model	Training					Test			
	BIC	BIC_w	n	RMSD	$RMSD_{LB}$	D	D_w	n	RMSD
Baseline	1833 (78)	0.04 (0.18)	12	9.2 (1.2)	9.2 (1.6)	543 (145)	0.03 (0.16)	9	14.2 (6.2)
LMS rule	1782 (144)	0.03 (0.15)	12	8.4 (2.4)	7.8 (2.5)	435 (43)	0.04 (0.13)	10	9.1 (3.1)
Decay	1781 (146)	0.1 (0.29)	29	8.3 (2.5)	7.9 (2.6)	431 (44)	0.09 (0.23)	26	8.9 (3)
Capacity	1641 (116)	0.65 (0.46)	185	5.9 (1.6)	5.4 (2)	393 (40)	0.52 (0.45)	152	6.3 (2)
Attention	1715 (123)	0.17 (0.36)	49	6.6 (1.8)	5.7 (2.1)	413 (239)	0.32 (0.41)	90	6.5 (2.4)
Regression	—	—	—	4.7 (1.3)	4.3 (1.5)	—	—	—	5.4 (1.7)

Note. BIC = Bayesian Information Criterion; BIC_w = 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	2989 (77)	0 (0)	0	10 (0.9)	1514 (49)	0.16 (0.36)	8	10.5 (1.2)
LMS rule	2858 (181)	0.03 (0.12)	1	8.7 (1.9)	1467 (109)	0.1 (0.25)	5	9.2 (2.0)
Decay	2846 (186)	0.26 (0.43)	13	8.5 (2)	1477 (98)	0.16 (0.34)	8	9.5 (2.2)
Capacity	2715 (182)	0.65 (0.46)	34	7.2 (1.5)	1439 (180)	0.44 (0.48)	23	8.0 (2.4)
Attention	2798 (157)	0.06 (0.21)	3	7.9 (1.5)	3480 (3656)	0.14 (0.34)	7	12.5 (7.4)

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

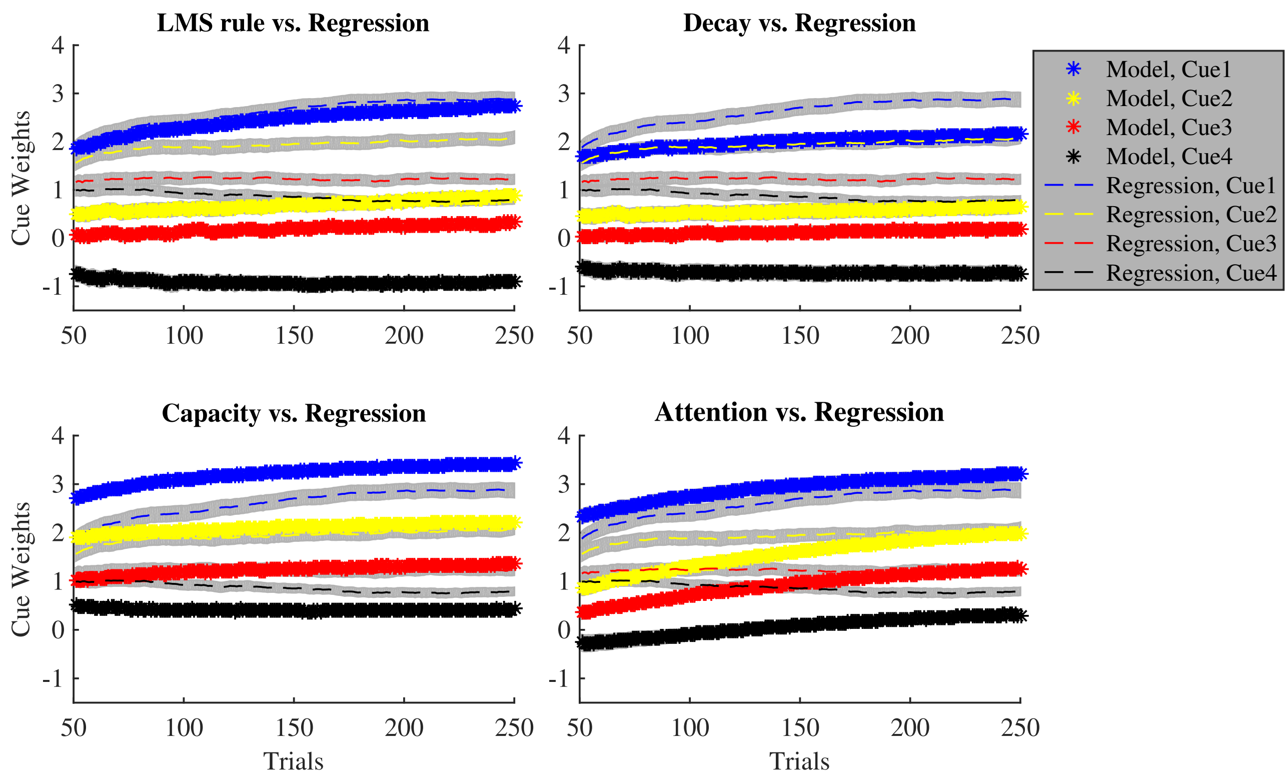

Figure 1. Cue weights predicted by each model in the reanalysis compared to cue weights from a rolling regression. Grey shaded areas show confidence intervals.

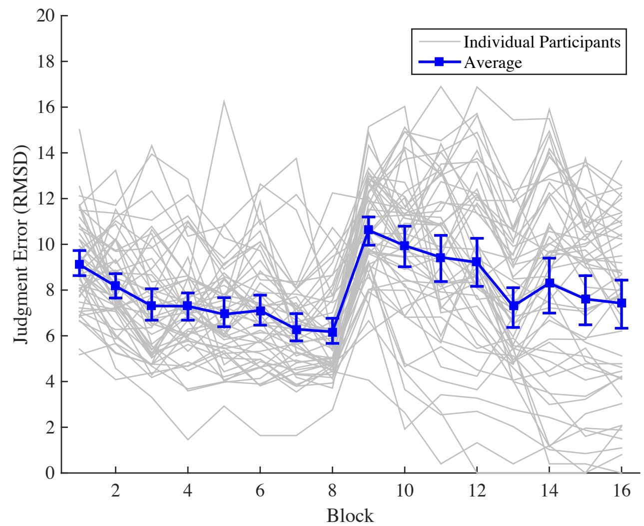

Figure 2. Judgment error (Root Mean Square Deviation, RMSD) in the Relearning Experiment. The black line shows the average judgment error, gray lines show judgment error of individual participants in each learning block. Error bars plot bootstrapped confidence intervals.

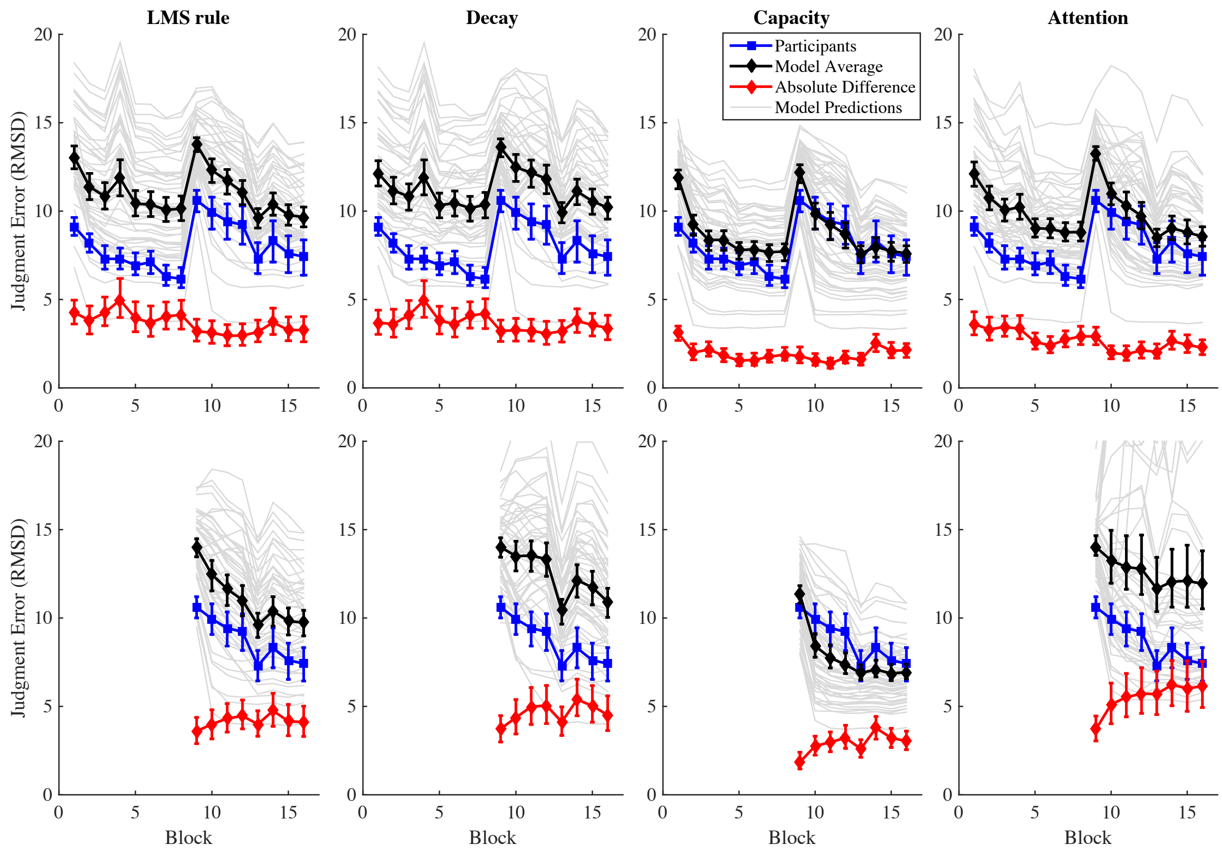

Figure 3. Judgment error (Root Mean Square Deviations, RMSD) averaged across all participants (blue lines) and judgment error predicted on average by each model (black lines) in each learning block. Red diamonds illustrate the absolute difference between both learning paths; light grey lines depict model 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.

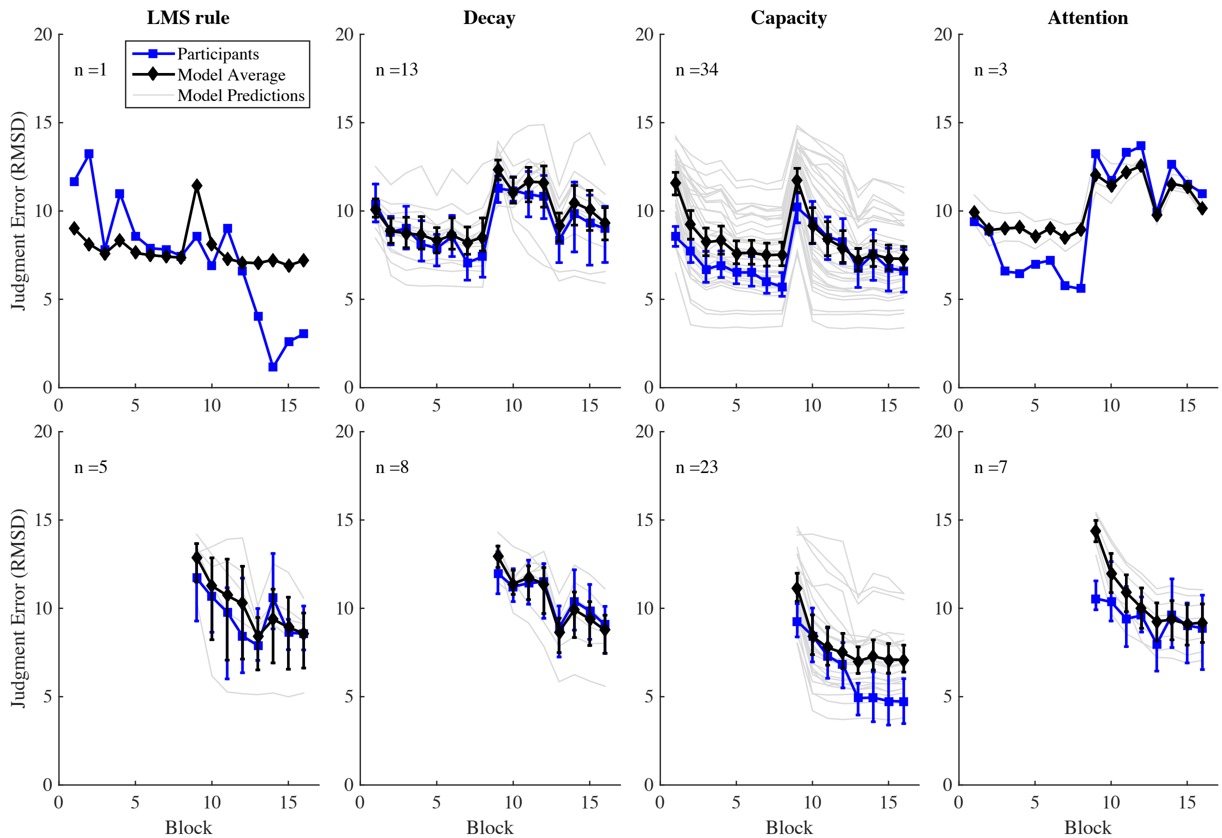

Figure 4. Judgment error in Root Mean Square Deviations (RMSD) for participants classified to each model (blue lines) and judgment error predicted by the model on average for those participants (black lines) in each learning block. Light gray lines depict model predictions for each single participant. 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.

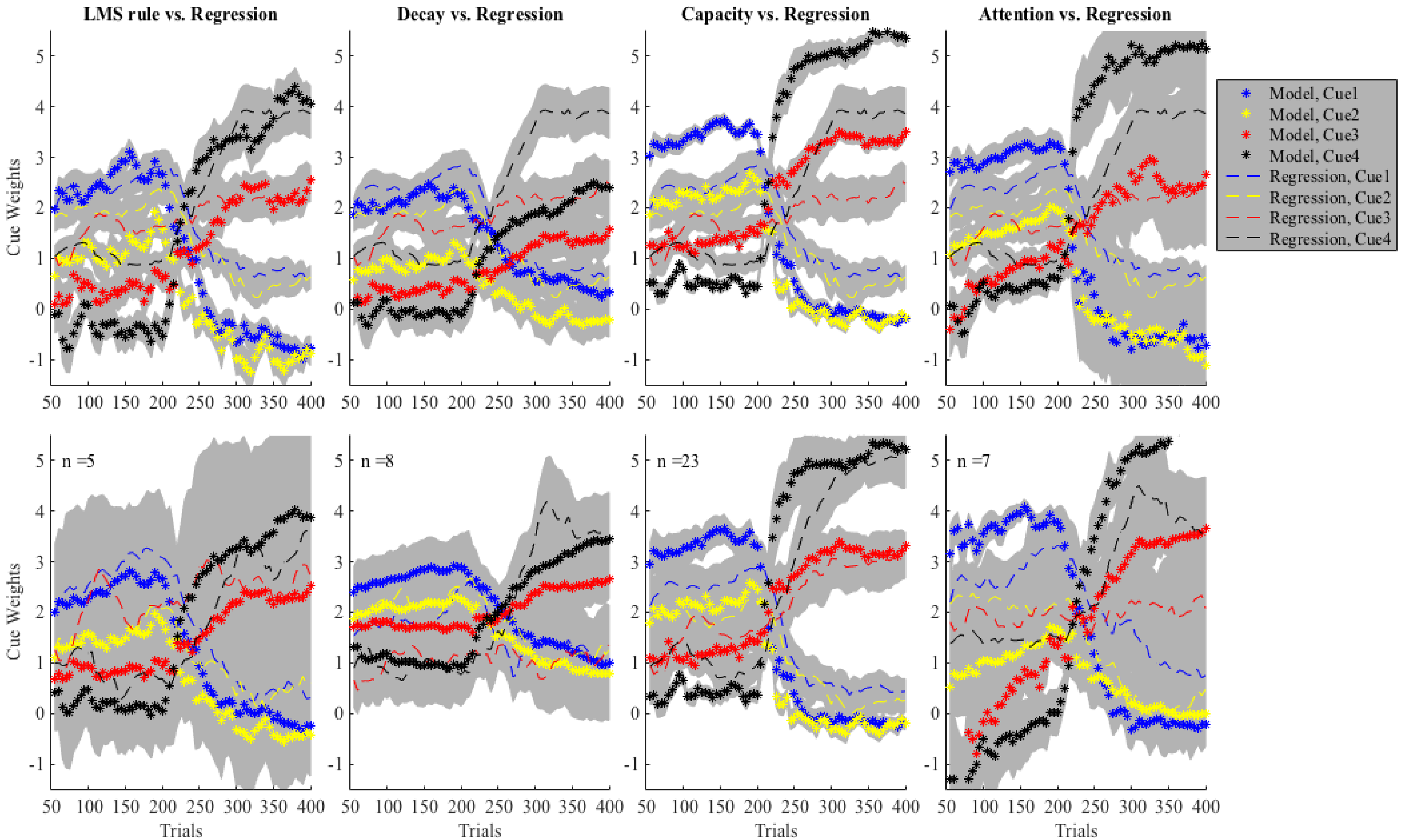

Figure 5. Cue weights predicted by each model (stars) in the relearning experiment compared to cue weights (lines) from a rolling regression. The upper row shows cue weights for all participants based on generalization; the lower row shows cue weights for those participants best described by each model in generalization. Grey shaded areas show confidence intervals.

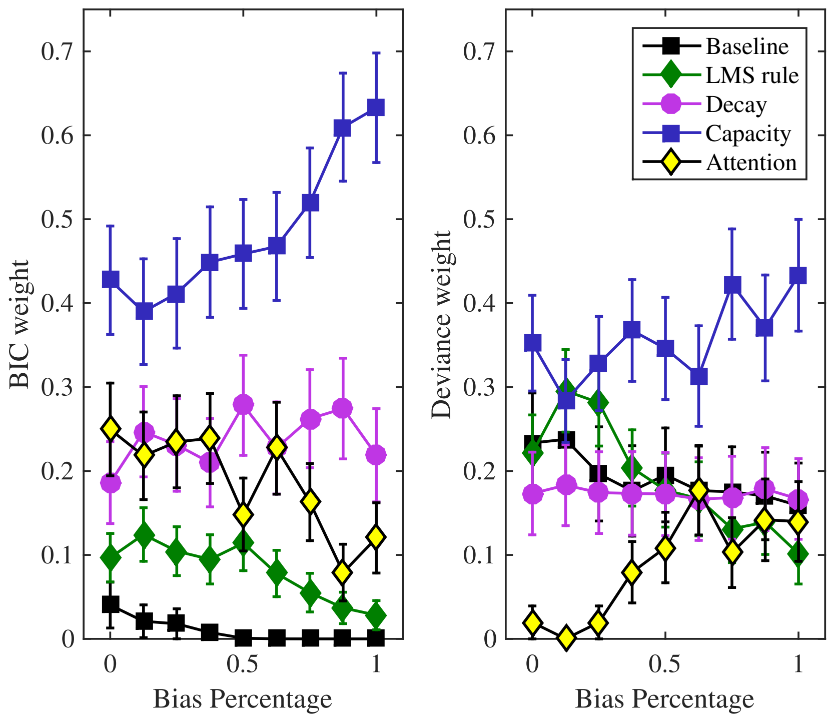

Figure 6. BIC and deviance weights for each model (separate lines) as a function of the starting value for the cue weights (expressed in percentage of bias). The left panel shows BIC weights from estimating the models based on all learning trials. The right panel shows deviance weights from generalization for the second half of the experiment. Error bars show ± standard error.

Appendix A

Updating of cue weights in the capacity model

The capacity model proposes that the updated cue weights sum up to a capacity restriction r, $\sum_i |w_{t+1,i}| \leq r$ (Busemeyer et al., 1993a). Using this restriction within the learning model is similar to learning a LASSO regression that optimizes the regression weights under the constraint that $\sum_i |w_{t+1,i}| \leq r$. One optimization technique that efficiently solves the lasso regression problem is projected gradient descent (Hastie, Tibshirani, & Wainwright, 2015). In projected gradient descent, the algorithm first performs an unconstrained gradient step, updates the regression weights, and next projects the updated regression weights back onto the constrained space. To learn a lasso regression, thus it is possible to first update the cue weights using the LMS rule and in a next step to project the updated cue weights back onto the L1-ball. This is the approach we also pursued here.

Specifically, the problem reduces then to projecting onto the L1-ball only if the capacity restriction is violated and, consequently, one can solve instead the simpler problem (Duchi et al., 2008)

$$\text{minimize}\, \|\beta - u\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1 \leq r \quad \text{and} \quad \beta \geq 0 \tag{12}$$

with $u_i = |v_i|$ and $w_{t+1,i} = \text{sign}(v_i)\beta_i$. Within this problem, there exists a unique θ such that

$$\beta_i = \left[u_i - \theta\right]^+ \tag{13}$$

where $[x]^+$ is x for $x > 0$, else 0. Accordingly, θ needs to be found that fulfills $\sum_i^I \left[u_i - \theta\right]^+ = r$ with only those u_i contributing to the sum that are not set to 0 during the projection. Reformulating this equation yields

$$\theta = \frac{1}{p}\sum_{i=1}^{p} \mu_i - r \tag{14}$$

with p indicating the non-zero elements after the update and μ_i indicating that u_i is sorted in descending order of magnitude (Shalev-Shwartz & Singer, 2006). Finally, p is then the index for which

$$p = \max\{j \in [n] : \mu_j - \frac{1}{j}(\sum_{q=1}^{j} \mu_q - r) > 0\}, \tag{15}$$

that is, the index p is selected for which the absolute weight μ_j is still higher than the difference between the averaged ordered absolute weights up to index p and the capacity limit.

Appendix B

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.

$$-2LL = -2 \cdot \sum \ln(L) \tag{16}$$

We calculated the likelihood as the probability density of participants' judgments j assuming a normal distribution, with the models' predicted responses $\hat{j}_t$ as the mean of the normal distribution and a fitted standard deviation σ.

Bayesian Information Criterion

To calculate the Bayesian Information Criterion (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:

$$\text{BIC} = -2LL + k \ln n, \tag{17}$$

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).

$$\text{BIC}w_M = \frac{e^{-.5\Delta\text{BIC}_M}}{\sum\limits_i e^{-.5\Delta\text{BIC}_i}} \tag{18}$$

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 1000 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 environment) and predicted participants' learning performance in the second half of the learning blocks (after the task environment changed). In accordance with the BIC weights, we computed deviance weights (D_w) 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 the predicted RMSD for each learning block in the validation trials, we generated model predictions for all validation trials 1000 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 C

Model parameters for the reanalysis and the relearning experiment

Appendix C lists the estimated mean parameter values for the reanalysis (Table C1) and the relearning experiment (Table C2) 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 B). Parameter estimates for the generalization test were estimated based on the first 200 trials in the experiment.

Table C1

Model Parameter in the Reanalysis. SD in Parentheses

Model	Parameter				
	λ	δ	r	b_0	*SD*
Baseline	0.082 (0.076)	–	–	–	9.4 (1.2)
LMS rule	0.008 (0.005)	–	–	–	8.7 (2.5)
Decay	0.043 (0.102)	7.1 (20.6)	–	–	8.6 (2.5)
Capacity	0.018 (0.016)	–	17.3 (6.2)	–	6.4 (1.6)
Attention	0.021 (0.28)	–	–	-5.091 (3.25)	7.5 (1.8)

Note. λ = Learning rate; δ = Decay rate; r = Capacity restriction; b_0 = Initial learning rate for the single cues; *SD* = Standard Deviation.

Table C2

Model Parameter in the Relearning Experiment. SD in Parentheses

Criterion	Model	Parameter				
		λ	δ	r	b_0	*SD*
BIC	Baseline	0.087 (0.089)	–	–	–	10 (0.9)
	LMS rule	0.009 (0.007)	–	–	–	8.7 (1.9)
	Decay	0.032 (0.036)	0.5 (0.5)	–	–	8.5 (2)
	Capacity	0.012 (0.015)	–	17.8 (7.3)	–	7.2 (1.5)
	Attention	0.0001 (0.009)	–	–	-4.171 (1.281)	8 (1.5)
Generalization	Baseline	0.118 (0.108)	–	–	–	9.6 (1)
	LMS rule	0.008 (0.006)	–	–	–	8.7 (2.3)
	Decay	0.036 (0.049)	2.8 (12)	–	–	8.5 (2.3)
	Capacity	0.02 (0.016)	–	18.1 (7.2)	–	6.8 (1.5)
	Attention	0.009 (0.02)	–	–	-4.589 (2.186)	8 (1.8)

Note. BIC = Bayesian Information Criterion; λ = Learning rate; δ = Decay rate; r = Capacity restriction; b_0 = Initial learning rate for the single cues; *SD* = Standard Deviation.

Appendix D

Specification of the full model using the combined psychological mechanisms

In line with the isolated psychological mechanisms, the full model including all psychological mechanisms proposes that the judgment $\hat{j}_t$ is a linear, additive combinations

$$\hat{j}_t = \sum_i w_{t,i} \cdot x_{t,i} \tag{19}$$

with $x_{t,*}=\begin{bmatrix} x_{t,1} & \dots & x_{t,n} & 1 \end{bmatrix}$ where n denotes the number of cues and 1 denotes the constant intercept. As in the attention model, the learning rates are adjusted before any cue weight is updated by considering the global learning rate λ, the salience of the cues, and a decaying memory trace, $h_{t,i}$, that stores previous weight updates.

$$b_{t,i} = b_{t-1,i} + \frac{\lambda \cdot x_{t,i} \cdot (y_t - \hat{j}_t) \cdot h_{t,i}}{t^{\delta}} \tag{20}$$

Compared to the decay model, the decay in learning rates here affects the global learning rate λ. In a next step, the cue weights are updated with $\alpha_{t,i} = e^{b_{t,i}}$, resulting in the unconstrained cue weights $v_{t,i}$.

$$v_{t,i} = w_{t,i} + \alpha_{t,i} \cdot x_{t,i} \cdot (y_t - \hat{j}_t) \tag{21}$$

In case the capacity limit is not reached during the update, the unconstrained cue weights are used in the next trial, $w_{t+1,i} = v_{t,i}$ and the degenerated memory trace $h_{t,i}$ is updated with this current change in cue weights as in the attention model.

$$h_{t+1,i} = h_{t,i}\left[1 - \alpha_{t,i} \cdot x_{t,i}^2\right]^+ + \alpha_{t,i} \cdot x_{t,i} \cdot (y_t - \hat{j}_t) \tag{22}$$

In case, the capacity limit is reached during the update, however, the unconstrained cue weights are adjusted to adhere to this limit without changing their initial directions, $w_{t+1,i} = \text{sgn}(v_{t,i})\beta_{t,i}$, with the restricted weights $\beta_{t,i}$ calculated as the difference to the capacity restriction,

$$\beta_{t,i} = \left[|v_{t,i}| - \theta\right]^+ = \left[|v_{t,i}| - \frac{1}{p} * \Big(\sum_p |v_{t,p}| - r\Big)\right]^+ \tag{23}$$

with $[x]^+$ is x for $x > 0$, else 0. As a result, the memory trace $h_{t+1,i}$ that stores the recent changes in weights with respect to the learning rates $b_{t,i}$ (Sutton, 1992) likewise has to reflect this restriction.

$$h_{t+1,i} = \frac{\delta w_{t+1,i}}{\delta b_{t,i}} = \frac{\delta(\mathrm{sgn}(v_{t,i})\beta_{t,i})}{\delta b_{t,i}} \tag{24}$$

If the absolute unconstrained weight $|v_{t,i}|$ falls below or equal θ, $|v_{t,i}| \leq \theta$,the memory trace $h_{t+1,i}$ is set to 0. If the absolute unconstrained weight $|v_{t,i}|$ lies above θ, the memory trace decays with

$$h_{t+1,i} = h_{t,i}\left[1-\alpha_{t,i}\cdot x_{t,i}^2-\tfrac{1}{p}+\tfrac{1}{p}\,\mathrm{sgn}(v_{t,i})x_{t,i}\sum_p \mathrm{sgn}(v_{t,p})\alpha_{t,p}\cdot x_{t,p}\right]^+ + \Big(1-\tfrac{1}{p}\Big)\alpha_{t,i}\cdot x_{t,i}\cdot(y_t-\hat{j}_t) \tag{25}$$

The decay in the memory trace reflects the adjustment with respect to the other cue weights (in the first term of the equation) and is considered as strictly positive (Sutton, 1992). The update for the degenerated memory trace then similarly has to account for the adjustment in the current change of cue weights, $\Big(1 - \frac{1}{p}\Big)\alpha_{t,i} \cdot x_{t,i} \cdot (y_t - \hat{j}_t)$.

Appendix E

Model fits for the combined mechanisms in the relearning experiment

Appendix E lists the model fits for the combined psychological mechanisms in the relearning experiment (Table E1) with standard deviations in parentheses. BIC values were determined based on all trials in the experiment (cf. Appendix B). *D* values for the generalization test were calculated based on the last 200 trials in the experiment.

Table E1

Model Fits for the Combined Mechanisms in the Relearning Experiment. SD in Parentheses

Model	BIC				Generalization			
	BIC	BIC_w	n	RMSD	D	D_w	n	RMSD
Capacity + Decay	2681 (178)	0.62 (0.46)	31	6.9 (1.5)	1472 (137)	0.37 0.48)	19	8.3 (1.7)
Attention + Decay	2789 (157)	0.06 (0.22)	3	7.7 (1.4)	2942 (3302)	0.29 0.46)	15	11.1 (6.9)
Capacity + Attention	2702 (182)	0.22 (0.38)	12	7 (1.5)	3244 (3010)	0.1 0.29)	5	12.1 (6.1)
Full Model	2707 (179)	0.1 (0.28)	5	7 (1.5)	1815 (1151)	0.24 0.42)	12	9 (3.6)

Note. BIC = Bayesian Information Criterion; BIC_w = Bayesian Information Criterion weight; RMSD = Root Mean Square Deviation; D = Deviance; D_w = Deviance weight; Full Model = Capacity + Attention + Decay. BIC_w, D_w and n only consider comparisons within the set of combined learning mechanisms.