Reconstructing the Einstellung effect

Marcel Binz
Max Planck Institute for Biological Cybernetics

Eric Schulz
Max Planck Institute for Biological Cybernetics

The Einstellung effect was first described by Abraham Luchins in his doctoral thesis published in 1942. The effect occurs when a repeated solution to old problems is applied to a new problem even though a more appropriate response is available. In Luchins' so-called water jar task, participants had to measure a specific amount of water using three jars of different capacities. Luchins found that subjects kept using methods they had applied in previous trials, even if a more efficient solution for the current trial was available: an Einstellung effect. Moreover, Luchins studied the different conditions that could possibly mediate this effect, including telling participants to pay more attention, changing the number of tasks, alternating between different types of tasks, as well as putting participants under time pressure. In the current work, we reconstruct and reanalyze the data of the various experimental conditions published in Luchins' thesis. We furthermore show that a model of resource-rational decision-making can explain all of the observed effects. This model assumes that people transform prior preferences into a posterior policy to maximize rewards under time constraints. Taken together, our reconstructive and modeling results put the Einstellung effect under the lens of modern-day psychology and show how resource-rational models can explain effects that have historically been seen as deficiencies of human problem-solving.

Keywords: Einstellung Effect; Resource Rationality; Problem-Solving; Reasoning; Decision-Making

Correspondence concerning this article should be addressed to Eric Schulz, MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics, Max Planck Ring 8, 72076 Tübingen, Germany. E-mail: eric.schulz@tue.mpg.de

"In the beginner's mind there are many possibilities, in the expert's mind there are few." – Shunryu Suzuki

A general rule of human intelligence seems to be that improving at a task simply requires repeatedly performing it (Rock, 1957; Saxe, 2013; Tolman, 1934). The sort of tasks at which people can improve range from the exceptionally low-level, such as visual discrimination tasks (Poggio et al., 1992), to the high-level, such as complex problem-solving tasks (Funke, 2012). The usual explanation of this ability is that repetition leads to learning which in turn leads to improved performance.

Yet this rule has been repeatedly challenged (Birch & Rabinowitz, 1951; N. R. Maier, 1931). For example, people can be limited to use an object only in the way it is traditionally used. This effect was observed and described by Karl Duncker who gave participants a candle, a box of thumbtacks, and a book of matches, and asked them to attach the candle to the wall so that it did not drip onto the table below. Participants tried to attach the candle directly to the wall with the tacks, or to glue it to the wall by melting it. However, only very few of them thought of using the inside of the box as a candle holder. Duncker argued that participants were fixated on the box's usual function of holding thumbtacks and could not re-conceptualize it in a manner that allowed them to solve the problem at hand (Duncker & Lees, 1945).

One of the main psychological findings on how repeated encounters with a task can lead to suboptimal behavior is the *Einstellung effect* (A. S. Luchins, 1951). The Einstellung effect describes the development of a mechanized state of mind and refers to a person's predisposition to solve a problem in a specific manner even though better or more appropriate solutions are available. The effect itself has been observed in many contexts. One of them is expertise, where, even though expertise frequently leads to superior decision-making (Ericsson & Charness, 1994; Ericsson et al., 2007), people can sometimes get worse at particular tasks with experience. For example, when expert chess players were confronted with problems that could be solved with either a common sequence of moves or with a less common but shorter sequence, they tended to overlook the simpler solution (Bilalić et al., 2008a, 2008b, 2010; Reingold et al., 2001). A similar observation has been made in medical decision-making, where professional doctors can sometimes diagnose rare illnesses less reliably than novices (Croskerry, 2003; Graaf, 1989).

The most famous set of experiments on the Einstellung effect, however, are the ones that introduced it. These experiments were conducted by Abraham Luchins and published in his doctoral thesis in 1942 (A. S. Luchins, 1942). In these experiments, participants were asked to assume that there were three water jars, each of which had the capacity to hold a different amount of water. Participants had to figure out how to measure a certain amount of water using these jars. When participants solved multiple such problems in succession, Luchins found that subjects kept using methods they had applied in previous trials, even if a more efficient solution for the current trial was available. This means that participants were influenced by their previous experience to be less efficient at solving current problems, establishing the original Einstellung effect.

Luchins conducted an impressive set of empirical studies, collecting data of over 1000 participants in different conditions and thereby mapping out the key determinants of the Einstellung effect. We believe that such a remarkable collection of knowledge should be preserved for future generations. Therefore, we have transcribed the experimental data reported in Luchins' thesis into a contemporary format, analyzed them using modern statistical tools, and made them publicly available for further investigation.[1] We furthermore provide a simple computational model that can capture the rich behavioral pattern found in Luchins' data. Whereas previous research has largely characterized the Einstellung effect as maladaptive behavior, our analysis provides a resource-rational account of the effect. In particular, our model assumes that people attempt to find solutions with maximum utility but that they are subject to information processing constraints. These results enrich our understanding of human problem-solving and shine new light on a historic finding using the lens of current psychological theories.

The remainder of this paper is structured as follows. First, we briefly recapitulate Luchins' experimental setup. We then introduce our resource-rational model of decision-making and provide full simulation results for Luchins' water jar task. Subsequently, we go through the individual experiments reported in Luchins' thesis. For each experiment, we reanalyze the transcribed data using modern statistical tools and validate that our proposed model exhibits similar behavior. To round up our analysis, we also conduct an ablation analysis demonstrating that all components of the model are indeed necessary to capture the richness in Luchins' data. Finally, we summarize our results, highlight the limitations of our approach, and propose directions for future research.

Luchins' experiments

Luchins adapted his water jar task from tasks that were initially designed by Duncker and Zener[2] and described by N. R. F. Maier (1930):

> *"Zener, in some preliminary experiments at the Psychological Institute of the University of Berlin, in 1927, habituated his subjects to solve certain types of problems in the same way. A*

[1] The complete code for this project, including the reconstructed data and all model simulations, can be accessed under https://github.com/marcelbinz/Einstellung.

[2] Note that the results from Duncker and Zener were never published but only described anecdotally by N. R. F. Maier (1930).

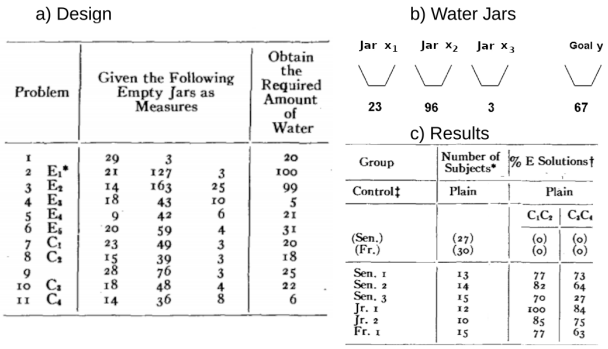
test problem was then given. He found that an obvious and simple solution of the test problem was usually over-looked because the characteristic method of solution, set up in the preceding problems, was used in the test problem. Control groups tended to solve the problem in the obvious and simple manner."

A. S. Luchins (1942) argued that it was important to further test this effect:

"It seemed important to conduct further experiments of this kind because the quoted findings of these preliminary experiments appeared to show clearly an interesting result: The successive, repetitious use of the same method mechanized many of the subjects—blinded them to the possibility of a more direct and simple procedure."

After consulting with Max Wertheimer, Luchins ended up using similar problems to those utilized by Zener and Duncker. The resulting task is now known as *Luchins' water jar task*, and it asks people to write down, using pen and paper, how they would obtain a target quantity of water using three jars which each hold a specific quantity.[3] We refer to the quantity each jar holds as x_1, x_2, x_3 and the target quantity as y (Fig. 1b). Participants first get a problem with only two jars in which $x_1 = 29$, $x_2 = 3$ and $y = 20$. This problem can be easily solved by $x_1 - 3x_2$, that is removing the quanitity measued by the second jar from the first jar. The next five problems are referred to as the E-problems and are used to induce the Einstellung. This is because all of these problems can be solved using the same method: $x_2 - x_1 - 2x_3$. The following two problems are referred to as the C1- and C2-problem. They can be solved both by the previous method but also by a simpler method, such as $x_1 + x_2$. In C1, for example, the available jars are $x_1 = 23$, $x_2 = 49$, $x_3 = 3$ and the target quantity is $y = 20$. This means that this problem can be solved by $x_2 - x_1 - 2x_3$ but also by using the shorter $x_1 - x_3$. Applying $x_2 - x_1 - 2x_3$ to these problems is called an E-solution (where the E stands for Einstellung), whereas the shorter solutions, i.e. $x_1 + x_2$ or $x_1 - x_3$, are called D-solutions (where the D stands for direct). C1 and C2 are followed by a problem that can not be solved using the E-solution but only by using a shorter D-solution. Finally, participants are confronted with two problems, C3 and C4, that can again be solved by both E- and D-solutions. Luchins' original design is shown in Fig. 1a.

Luchins mostly focused on participants' responses to the problems C1/C2 and C3/C4. If participants were more likely to solve these problems using the E-solution than the D-solution, an Einstellung effect was observed. He normally compared a "plain" group going through all of the problems



Problem		Given the Following Empty Jars as Measures			Obtain the Required Amount of Water
1		29	3		20
2	E_1*	21	127	3	100
3	E_2	14	163	25	99
4	E_3	18	43	10	5
5	E_4	9	42	6	21
6	E_5	20	59	4	31
7	C_1	23	49	3	20
8	C_2	15	39	3	18
9		28	76	3	25
10	C_3	18	48	4	22
11	C_4	14	36	8	6

Group	Number of Subjects*	% E Solutions†	
Control‡	Plain	Plain	
		C_1C_2	C_3C_4
(Sen.)	(27)	(0)	(0)
(Fr.)	(30)	(0)	(0)
Sen. 1	13	77	73
Sen. 2	14	82	64
Sen. 3	15	70	27
Jr. 1	12	100	84
Jr. 2	10	85	75
Fr. 1	15	77	03

Figure 1
Overview of Luchins' water jar task. ***a:*** *Experimental design. After a warm-up problem, participants observe five problems that require the complicated E-solution, followed by five problems that can be solved using a shorter D-solution.* ***b:*** *Water jars. Participants were asked to combine the quantities of three jars to reach at a target quantity.* ***c:*** *Results of first experiment reported in A. S. Luchins (1942). Luchins reported all results in tables showing the percentage of participants who solved the presented problems using E- or D-solutions.*

shown in Fig. 1a to other groups, for example, participants who did not have to solve problems E1-E5. For example, in Luchins' very first study, he compared a control group of students not receiving E1-E5 with the plain group who received these problems. His results showed that the plain group solved C1 and C2 using the E-solution between 77% and 82% of the time, and C3 and C4 between 27% and 84% of the time. The control group, however, never applied any E-solutions. Luchins' very first results, taken from the first table of his thesis, are shown (as displayed in his thesis) in Fig. 1c.

In his doctoral thesis, Luchins also reported many other experiments assessing how different manipulations influenced the presence and strength of the Einstellung effect. These additional experiments show an impressive data collection effort. In particular, Luchins' results showed that the Einstellung effect was stronger for earlier (C1/C2) than later (C3/C4) test trials, decreased when participants were asked to pay more attention, increased with the number of training tasks, vanished when E- and D-problems were interleaved during training and increased when participants were put under time pressure.

Luchins reported all of his results in tabular form (see Fig. 1c), indicating how many people were assigned to which group as well as the frequency of E-solutions and –sometimes– D-solutions. For him, this format was sufficient

[3] Assuming that subjects have access to an infinite source of water.

to draw conclusions about what conditions induced an Einstellung effect. He did, however, not analyze his data quantitatively as current statistical software and analysis practices did not exist in the 1940s. To fill this gap, we transcribed Luchins' results into usable data files by going through his tables and coding the reported percentages as the number of participants who showed a certain response. This reconstruction allows us to analyze his results using modern-day statistics, as well as to reinterpret the Einstellung effect from the perspective of current approaches to reasoning and decision-making.

Explaining the Einstellung effect

To improve our understanding of the Einstellung effect, we have to ask ourselves why people would show the effect in the first place. In this section, we present a computational model that allows us to interpret Luchins' findings from the perspective of resource-rational decision-making (Bhui et al., 2021; S. J. Gershman et al., 2015; Lieder & Griffiths, 2019). This model assumes that people attempt to maximize their performance in Luchins' water jar task but do so while spending as little physical and mental effort as possible. In the subsequent section, we will then show that these simple principles are sufficient to explain the rich set of behavioral phenomena observed by Luchins.

Problem setting

Recall that $x = (x_1, x_2, x_3)$ denotes the capacity of each water jar, and y the total amount of water to be measured. The goal of a decision-making agent is to find a combination $w = (w_1, w_2, w_3)$ of integers, such that $w^\top x = y$. To each setting of w, we assign the following utility U:

$$U(w, x, y) = \begin{cases} 1 - \lambda||w||^2 & \text{if } w^\top x = y \\ -\lambda||w||^2 & \text{if } w^\top x \neq y \end{cases} \tag{1}$$

Equation 1 assigns a utility of 1 to valid solutions, and a utility of 0 to invalid solutions. In addition, it also contains a second term that favors simple solutions, i.e., those with a low squared Euclidean norm. Intuitively, this term captures that solutions which require fewer steps to execute, i.e., those that require less physical effort, should be preferred. The importance of the latter term (relative to finding a valid solution) is controlled by a hyperparameter λ, which we set to 0.05 in all of our simulations unless stated otherwise.

A perfectly rational decision-maker should simply search through all candidate solutions, and select the solution w^* with the highest utility. However, as it has been repeatedly pointed out in the literature (e.g. Gigerenzer & Selten, 2002; Simon, 1990), searching through the entire space of possible solutions is normally not feasible beyond simple toy problems. In the present task it would, for example, require nearly 10000 evaluations even if we restricted ourselves to $w_i \in \{-10, -9, \ldots, 9, 10\}$.

Resource-rational decision-making

How can we make progress if searching the entire solution space is not feasible? The framework of resource rationality offers a solution to this problem. Like perfectly rational decision-makers, resource-rational decision-makers attempt to find an optimal solution, but do so while taking limited computational resources into account. There exist a number of implementations of this idea (for example, S. J. Gershman, 2020; Lieder & Griffiths, 2017; Sanborn et al., 2010; Vul et al., 2014). The approach we pursue in this work is an extension of the information-theoretic model proposed by Ortega et al. (2015). In this model, the decision-maker starts with a prior preference over solutions $p(w)$, which is then transformed into a posterior policy $q(w|x, y)$ once a decision has to be made. More specifically, this transformation is done such that the posterior policy maximizes excepted utility, while keeping the cost of the transformation minimal:

$$q = \underset{\hat{q}}{\operatorname{argmax}} \underbrace{\sum_w \hat{q}(w|x, y)U(w, x, y)}_{\text{expected utility}} - \underbrace{\frac{1}{\beta}\mathrm{KL}\left[\hat{q}||p\right]}_{\substack{\text{transformation} \\ \text{cost}}} \tag{2}$$

Equation 2 uses the Kullback-Leibler (KL) divergence to measure the cost of transforming prior preferences into the posterior policy. We will return to how this cost can be interpreted from a psychological perspective below. The trade-off between this transformation cost and the excepted utility is controlled by an inverse temperature parameter β. For $\beta \to \infty$, Equation 2 recovers the solution that maximizes expected utility without considering computational resources, i.e., $q(w|x, y) = \delta_{ww^*}$. For $\beta \to 0$, it corresponds to the trivial solution of not transforming the prior at all, i.e., $q(w|x, y) = p(w)$. For any arbitrary β, optimal solution to Equation 2 is given by the Gibbs distribution (Ortega et al., 2015):

$$q(w|x, y) = \frac{1}{Z}p(w)e^{\beta U(w,x,y)} \tag{3}$$

$$Z = \sum_w p(w)e^{\beta U(w,x,y)} \tag{4}$$

Looking at Equations 3 and 4, we can observe a close connection to Bayesian inference. In particular, we can interpret the exponentiated utilities $e^{\beta U(w,x,y)}$ as a likelihood term in the standard Bayesian framework, with the hyperparameter β acting as a scaling factor for utilities.

Importance sampling

To act optimally, a resource-rational decision-maker has to draw a sample from its posterior policy. The naive method

of obtaining such a sample is to first compute $q(w|x, y)$ according to Equations 3 and 4, followed by sampling from it. However, this would still require to iterate over the entire search space, making it seem like we have gained nothing from the resource-rational problem formulation. Luckily, it turns out that it is possible to obtain a sample from the posterior policy *without* explicitly computing it. Havasi et al. (2018) proposed a simple importance sampling procedure for doing exactly that. Their algorithm draws M samples $w_1, \dots, w_M \sim p(w)$ and selects w_m with probability $p_{\text{accept}} \propto \frac{q(w_m|x,y)}{p(w_m)}$. In the limit of $M \to \infty$, this procedure will generate an unbiased sample from $q(w|x, y)$. But how accurate is this procedure when the number of samples is limited? Havasi et al. (2018) showed that one only has to draw $M = \left\lceil e^{KL[q||p]} \right\rceil$ samples to ensure that the bias remains low. In our context, this result allows us to interpret the minimization of $KL\,[\hat{q}||p]$ as a reduction in the number of candidate solutions to be inspected for obtaining a sample from the posterior policy. Therefore, this term measures the number of thinking steps, i.e., the required mental effort to solve a problem. We may say that an agent that does not care about the transformation cost (i.e., an agent with high β) has to inspect many candidate solutions before setting on one of them, while an agent that does care about the transformation cost (i.e., an agent with low β) only has to inspect a few candidate solutions.

The choice of prior preferences

We have remained agnostic about the choice of prior preferences up to now. However, for a full model specification, we have to ask: what prior should a resource-rational agent use? A reasonable modeling choice is to assume that the agent employs the most economic prior for its decision-making environment, i.e., the prior that on average leads to the least costly decisions. Tishby et al. (2000) showed that this prior is given by the marginal distribution $p^*(w) = \sum_{x,y} q(w|x, y)p(x, y)$. The problem with this line of reasoning is that an agent in Luchins' water jar task does not have access to the distribution of decision-making problems that can be encountered $p(x, y)$, and thus it can not compute $p^*(w)$. Instead of using $p^*(w)$ directly, we suggest a heuristic strategy to approximate it. In particular, we start from an initial prior distribution $p(w|\theta)$ that is parametrized through a softmax function:

$$p(w|\theta) = \prod_{i=1} p(w_i = k) = \prod_{i=1} \frac{e^{\theta_{ik}}}{\sum_{k'} e^{\theta_{ik'}}} \tag{5}$$

and then adjust this prior with a single gradient step after each trial, such that the probability of selecting the previous solution w_t is increased:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log p(w_t|\theta) \tag{6}$$

This updating procedure approximates the optimal prior $p^*(w)$ after sufficiently many interactions with the environment. We treat the learning rate α as a free parameter in our upcoming modeling simulations, and initialize $\theta_{ik} = 0$ for all i and k, leading to a uniform initial prior preference. For convenience, we restrict the space of possible settings to $w_i \in \{-10, -9, \dots, 9, 10\}$.

Model simulations

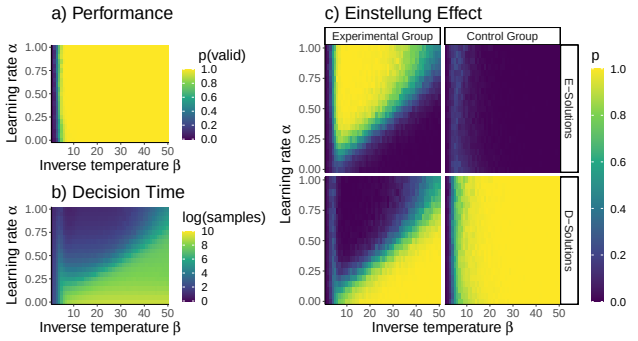


Figure 2

Model simulation results from the resource-rational decision-maker for different learning rates α and inverse temperatures β. ***a:*** *Percentage of valid solutions in the problems C1-C4, regardless of their complexity.* ***b:*** *Number of samples to be inspected in order to make a decision (shown on a logarithmic scale).* ***c:*** *Einstellung effect for different model parameters. The upper panels show the proportions of E-solutions for C1-C4, whereas lower panels show the proportions of D-solutions for the same tasks.*

We simulated our proposed model when performing Luchins' original task for different values of the learning rate α and the inverse temperature parameter β. For this simulation, we varied the learning rate α from 0 to 1, and the inverse temperature β from 1 to 50. The results of letting agents perform the task 100 times and averaging per unique (α, β)-constellation are summarized in Fig. 2. Fig. 2a shows the percentage of valid solutions in the problems C1-C4. We observed that any agent with an inverse temperature $\beta \gtrsim 8$ solves the task reliably and interpret this as a validation that our model exhibits the expected behavior. For lower β-values, the influence of the transformation cost dominates, causing agents to stick sampling from their prior preferences. Fig. 2b displays the number of samples to be inspected in order to make a decision. We observed, for example, that a resource-rational decision-maker with $\alpha = 0.6$ and $\beta = 25$ only needs to inspect 10 samples to make a decision, which is much less demanding from a computational perspective compared to the perfectly rational decision-maker, who has to inspect each of the $21^3 = 9261$ possible choices. In gen-

eral, we see that as β decreases, the number of required samples also decreases. The number of samples required is furthermore reduced drastically for medium to large learning rates, indicating that an agent can save a substantial amount of computational resources by learning a prior that is appropriate for its decision-making environment. Finally, we show the percentage of D- and E-solutions in Fig. 2c for both the experimental group (i.e., plain) and the control group. While most of the agents in the experimental group were able to solve the task reliably, they did so in very different ways: resource-rational decision-makers with low inverse temperatures tended to apply the E-solution, whereas those with high inverse temperatures tended to apply the D-solution. The point of transition from E- to D-solution was mediated by the learning rate. In contrast to this, we found that agents that were not exposed to any E-problems (the control group) exclusively relied on D-solutions, regardless of the chosen parameters. We will next use this model to take a closer look at several of Luchins' experimental findings.

The main Einstellung effect

For his very first experiment on the Einstellung effect, Luchins recruited several subjects from New York University as well as New York high schools and public schools. There were 310 participants in total whose responses we reconstructed from Luchins' thesis (A. S. Luchins, 1942). From these participants, 155 were in the plain group and therefore went through all of the problems shown in Fig. 1a, and 155 were in the control group and therefore did not experience problems E1-E5.

Reanalyzing this data (see Fig. 3a), we found that 128 out of 155 participants in the plain group responded by using an E-solutions to problems C1 and C2, whereas none of the 155 participants in the control group responded to these problems by using E-solution. The plain group, therefore, showed a stronger Einstellung effect than the control group on C1 and C2 (0.83 vs. 0.00; $\chi^2 = 218.02$, $p < .001$, Bayes Factor: $BF \approx 9.8 \times 10^{57}$). Moreover, 100 of the 155 participants in the plain group responded by using E-solutions to problems C3 and C4, whereas again none of the 155 participants in the control group did. Thus, the plain group also showed a stronger Einstellung effect than the control group on C3 and C4 (0.65 vs. 0.00; $\chi^2 = 147.62$, $p < .001$, $BF \approx 7.9 \times 10^{38}$). Finally, participants in the plain group showed a stronger Einstellung effect when responding to C1 and C2 than when responding to the later problems C3 and C4 (0.83 vs. 0.65; $\chi^2 = 13.00$, $p < .001$, $BF \approx 109.8$).

We next attempted to reproduce these response patterns using our model of resource-rational decision-making (see Fig. 3b). We picked a model with a medium learning rate of $\alpha = 0.5$ and an inverse temperature parameter of $\beta = 25$. We then let this model run for 100 simulations for both conditions and tracked whether or not it responded by using the

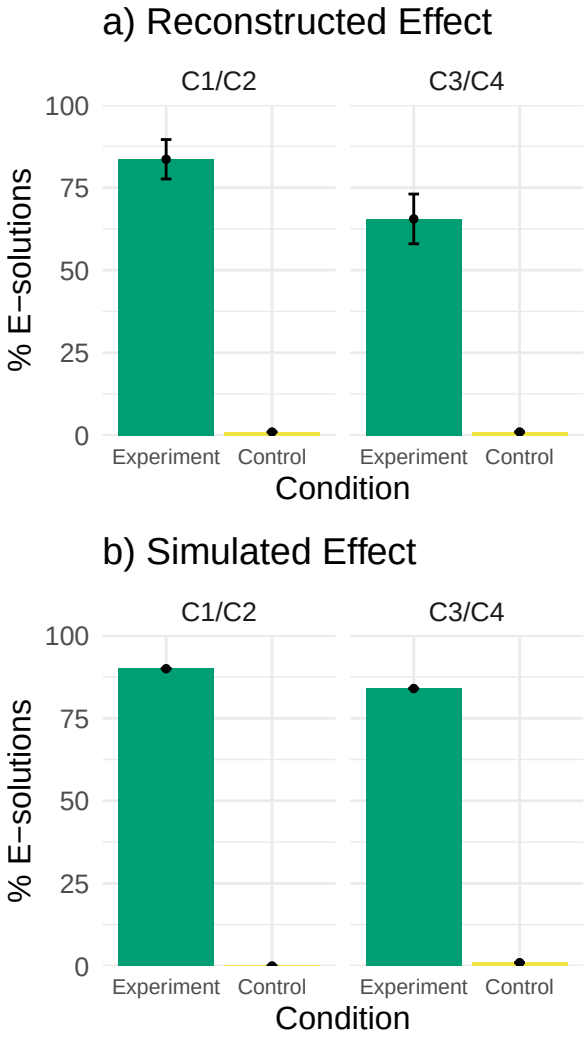


Figure 3
Results of Luchins' main experiment. ***A*** *Reconstructed experimental effect. Participants in the experimental condition showed a strong Einstellung effect, whereas participants in the control group did not show an Einstellung effect.* ***b:*** *Simulated effects. Our resource-rational decision-making model can reproduce the observed effects. Error bars indicate the standard error of the mean.*

E-solution for the problems C1-C4. We do not report significance tests for these simulations, since any simulated difference can become significant if only repeated for long enough. Looking at the simulated behavior, we saw that in 90 out of 100 simulations the agents responded by using E-solutions to problems C1 and C2 in the plain group, whereas 0 out of 100 agents used E-solutions for the same problems in the control group. The simulated behavior of agents in the plain group, therefore, showed a stronger Einstellung effect on C1 and C2

than the simulated behavior of agents in the control group (0.90 vs. 0.00). Furthermore, 84 out of 100 agents in the plain group responded by using the E-solution to problems C3 and C4, while again 0 out of 100 agents in the control group responded by using E-solutions. Thus, agents in the plain group showed a stronger Einstellung effect on C3 and C4 than agents in the control group (0.84 vs. 0.00). Finally, simulated agents also applied more E-solutions to C1 and C2 than to C3 and C4 (0.90 vs. 0.84). Based on these results, we conclude that our model can fully reproduce the observed human behavior in Luchins' original experiment.

Don't be blind

Already in his very first experiments, Luchins added another condition that had not been part of Duncker and Zener's original design. In this condition, called the "Don't be blind!" –or short DBB– condition, the following procedure was applied:

> *"Before any problems were presented, other members had been taken outside of the classroom and had been told, in the absence of the other subjects 'After returning to the classroom you will get a number of problems. After you will have completed Problem Six, write on your papers the words Don't be blind!'."*

This condition became a standard manipulation in Luchins' tasks. It is akin to debiasing participants by directly telling them to pay attention to possibly occurring flaws in their decision-making (Morewedge et al., 2015). In these studies, it is normally assumed that asking participants to pay more attention will drive them to perform better and therefore solve the task using the shorter D-solution. For example, Lane and Jensen (1993) showed that simply telling participants about the Einstellung effect can reduce the proportion of E-solutions significantly. Chrysikou and Weisberg (2005) also found that using de-fixation instructions (i.e., explicitly asking people to avoid using previously provided example solutions) eliminated set effect of the example solutions in problem-solving. Here, we report the main results of the DBB condition, taken from tables 1-5 of Luchins' thesis, comparing participants in the DBB group with participants from the plain group from before.

Reanalyzing this data (see Fig. 4a), we found that 87 out of 153 participants from the DBB group responded by using E-solutions to problems C1 and C2, showing a weaker Einstellung effect than the plain group on C1 and C2 (0.57 vs. 0.83; $\chi^2 = 24.16$, $p < .001$, $BF \approx 32156.9$). Moreover, 57 of the 153 participants from the DBB group responded by using E-solutions to problems C3 and C4, thereby also showing a weaker Einstellung effect than the plain group on these problems (0.37 vs. 0.65; $\chi^2 = 24.16$, $p < .001$, $BF \approx 14250.9$). Finally, participants in the DBB group showed a stronger

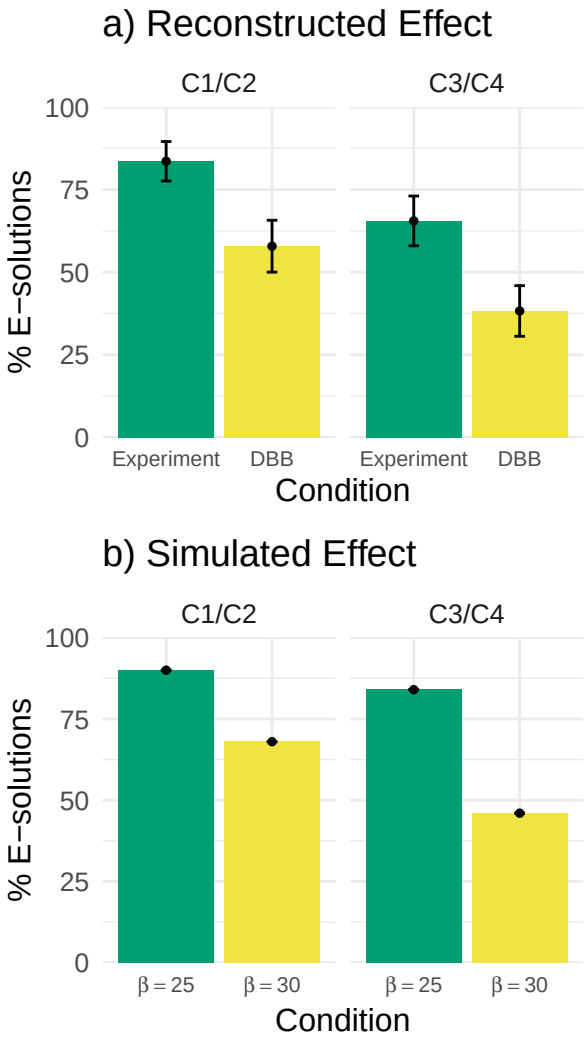


Figure 4
Results of the "Don't be blind!" condition. ***a:*** *Empirical results as reconstructed from Luchins' thesis.* ***b:*** *Simulated results as produced by our model of resource-rational decision-making. Error bars indicate the standard error of the mean.*

Einstellung effect when responding to C1 and C2 than when responding to the later problems C3 and C4 (0.57 vs. 0.37; $\chi^2 = 11.81$, $p < .001$, $BF = 52.1$).

We then again reproduced these effects by simulating from our model (see Fig. 4b). In our model, paying more attention corresponds to an increase in the inverse temperature parameter β, which leads to exerting more mental effort. We, therefore, increased this parameter to $\beta = 30$ while keeping the learning rate at $\alpha = 0.5$. In 68 out of 100 simulations for the DBB condition, the model solved C1 and C2 using the E-solution. For problems C3 and C4, the model used the E-solution in 46 out of 100 simulations. Thus, our model was

able to reproduce the Einstellung effect observed in the DBB condition. These simulation results show that the reduction of the Einstellung effect in the DBB condition can be seen as a result of more extensive, and therefore more effortful, updating of the distribution over possible solutions.

Varying the number of training tasks

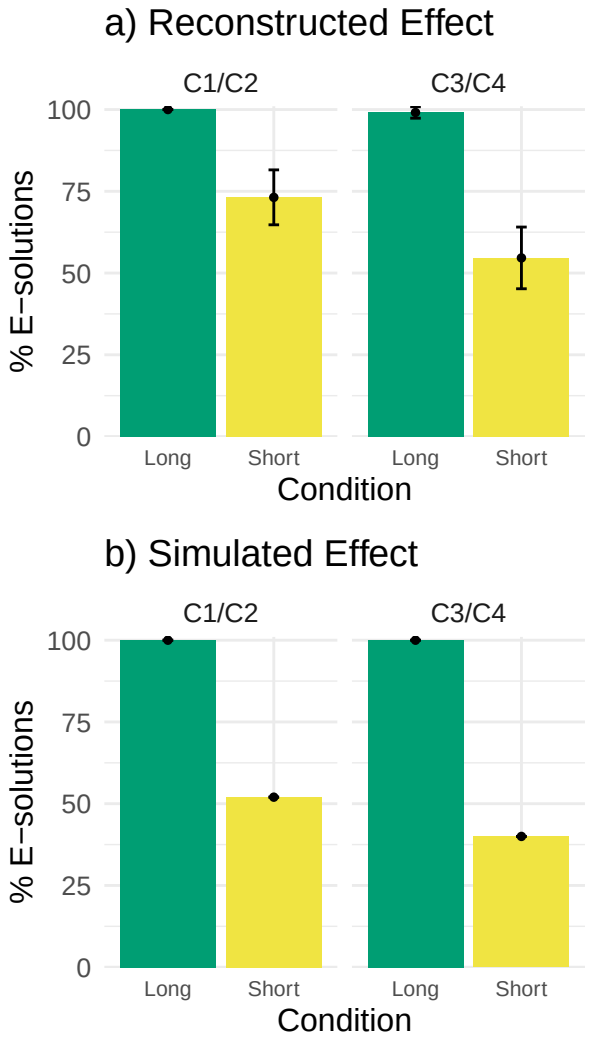


Figure 5
Effect of number of training tasks. ***a:*** *Empirical results as reconstructed from Luchins' thesis.* ***b:*** *Simulated results as produced by our model of resource-rational decision-making. Error bars indicate the standard error of the mean.*

Another variation that Luchins tried was to give one group of participants two E-problems and another group of participants ten E-problems. He did this to assess how the number of E-problems influenced the Einstellung effect. We transcribed this data from tables 12-13 in Luchins' thesis. The participants were students from public schools recruited around the New York area.

Reanalyzing this data (see Fig. 5a), we found that 112 out of 112 participants receiving ten E-problems responded by using E-solutions to problems C1 and C2, whereas only 79 of the 108 participants receiving two E-problems responded to these problems by using E-solutions. The group receiving more E-problems, therefore, showed a stronger Einstellung effect on C1 and C2 than the group receiving fewer E-problems (1 vs. 0.73; $\chi^2 = 34.64$, $p < .001$, $BF \approx 3.2{\times}10^8$). Furthermore, whereas 111 of the 112 participants receiving ten E-problems responded by using E-solutions to problems C3 and C4, only 59 of the 108 participants receiving two E-problems did. Thus, participants receiving more E-problems also showed a stronger Einstellung effect on C3 and C4 (0.99 vs. 0.55; $\chi^2 = 61.93$, $p < .001$, $BF \approx 1.5 \times 10^{15}$).

We also analyzed the behavior of simulated agents for this task. While we kept the learning rate the same as before, we had to decrease the inverse temperature parameter to $\beta = 15$ to fully capture human behavior in this task. This was because participants already showed a strong Einstellung effect even after only having seen two E-problems. Whether this increase in savings of mental effort was specific to the population collected for this task (i.e. public school students) or something that is generally the case for shorter experiments remains an open question for future investigations. We then trained the resulting agents on either ten or two E-problems as described above. When trained on ten E-problems, 100 of the 100 agents applied E-solutions to all problems C1-C4, thereby showing an increased Einstellung effect. When trained on only two E-problems, 52 of 100 agents responded by using E-solutions to problems C1 and C2, and 40 of 100 agents responded by using E-solutions to problems C3 and C4. Thus, agents trained on fewer E-problems showed a weaker Einstellung effect. Taking these results together, we can conclude that our model reproduces the effect of the number of training tasks on the Einstellung effect as described by Luchins.

Interleaving E- and D-problems during training

Another experimental condition that Luchins had run, was to present one group of participants with alternating E- and D-problems during training, whereas the other group only received E-problems. In this manipulation, one group of participants received seven problems that alternated between E- and D-problems, whereas the other group only received E-problems. Both groups were then tested on the problems C1 and C2 from the original design (Fig. 1a). That alternated training can make heuristic decision-making disappear has previously been shown in other domains. For example, Koehler (1996) argued that base rate neglect only appears when base rates are manipulated between rather than within subjects, and several studies have found support for

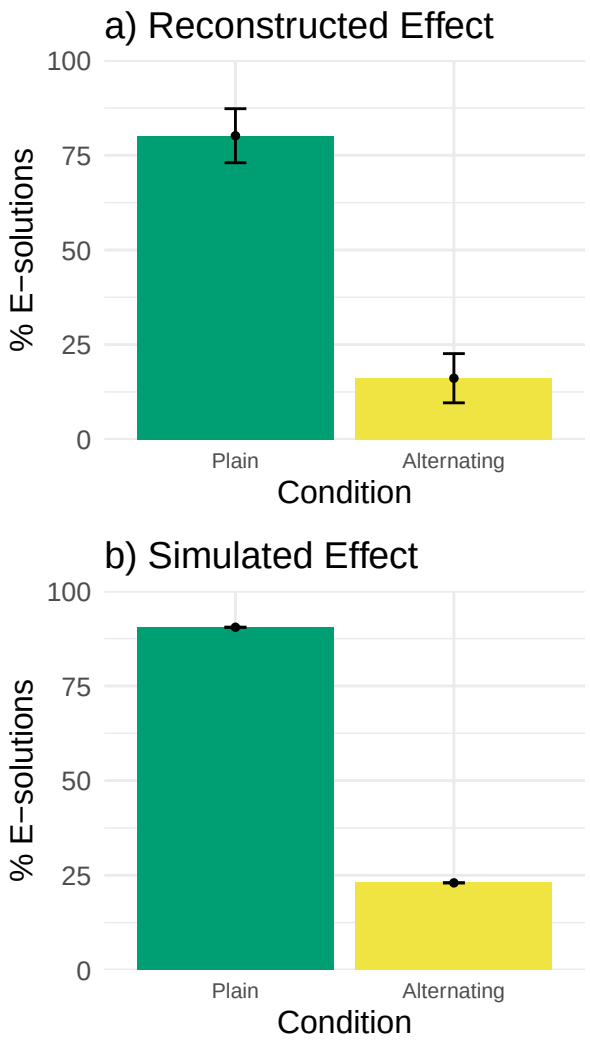


Figure 6
Effect of alternating E- and D-problems during training. ***a:*** *Empirical results as reconstructed from Luchins' thesis.* ***b:*** *Simulated results as produced by our model of resource-rational decision-making. Error bars indicate the standard error of the mean.*

Koehler's argument for example, see Fischhoff et al., 1979. Additionally, other studies have shown that interleaved training improves mathematical problem-solving (Rohrer et al., 2015) and category learning (Kornell & Bjork, 2008), among others. However, there have also been studies showing that blocked training can be better for both humans and neural networks (Flesch et al., 2018).

We transcribed the data from tables 14-15 in Luchins' thesis. There were 124 participants in the alternating condition and 121 participants in the plain condition. Participants were college students recruited from New York universities. When being trained on alternating problems, only 20 of 124 participants responded by using E-solutions to problems C1 and C2, while 97 of 121 participants receiving the plain training used E-solutions for the same problems. Thus, the group receiving alternate training showed a weaker Einstellung effect than the group receiving the normal training tasks (0.16 vs. 0.80; $\chi^2 = 100.65$ $p < .001$, $BF \approx 4.5 \times 10^{22}$).

We next assessed the behavior of simulated agents when E- and D-problems were either alternated or not (again keeping the learning rate $\alpha = 0.5$ and the inverse temperature $\beta = 25$). The results of this simulation showed that 90 of 100 agents responded by using E-solutions to problems C1 and C2 when trained only on E-problems, whereas only 23 of 100 agents responded by using E-solutions to the same problems. Thus, our model was able to reproduce the differences of the Einstellung effect when either trained on plain or alternating tasks.

Interestingly, multiple past studies have also shown that one way to reduce set effects more than by just telling people to pay more attention is to show them different solutions to the preceding problems (Crilly, 2015; Neroni & Crilly, 2020). Likewise, our model reduces the proportion of E-solutions if it has seen different solutions in the past, because there will be more probability mass on these different solutions.

The Einstellung effect under time pressure

Luchins also investigated how the Einstellung effect changed when people were put under time pressure. He started the description of these experiments with the observation that many students were relatively tense when they had to perform the task, and that they seemed to be behave almost as if they had to pass an exam. This led him to the following observation:

> *"It may even be an intelligent response to such conditions, the subject reasoning that he will get through quickly, or will finish first, etc., by repeating a previously mastered process, and that more time will be consumed if he stops to look for new methods."*

This observation is actually relatively close to our proposed model, which can save time by not updating its policy too much. To further test whether increasing this pressure could lead to a stronger Einstellung effect, Luchins put his subjects under time pressure by telling them that their responses will be evaluated based on how fast they were and emphasized the time pressure further "by a large laboratory clock at the front of the room, by the recording of the minutes on the blackboard, and by three stopwatches on the instructor's desk."

Nowadays, we know that people will frequently repeat previous actions despite intending to choose an alternative

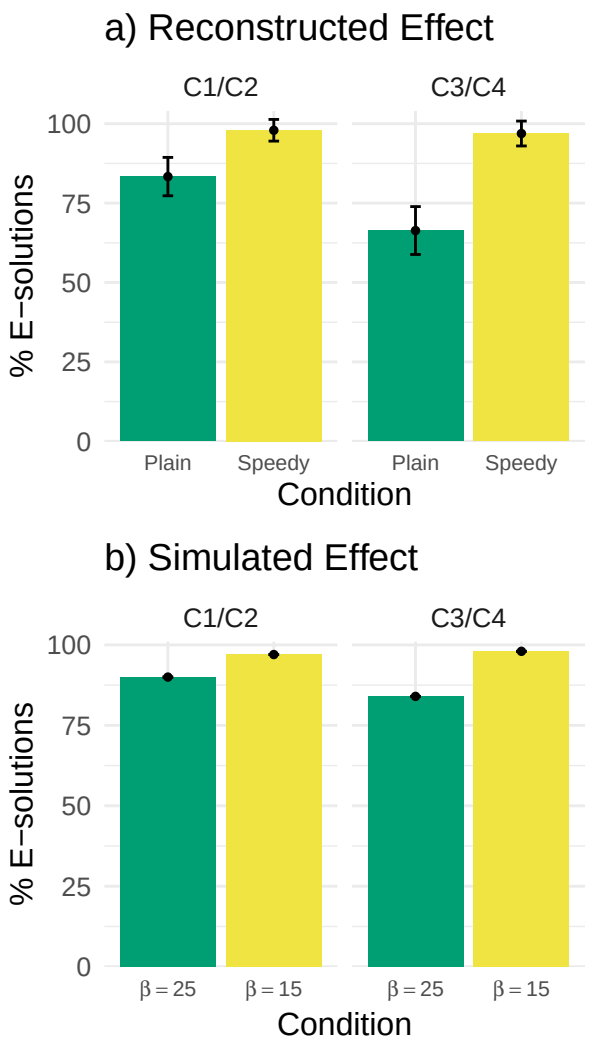


Figure 7

Effect of time pressure. ***a:*** *Empirical results as reconstructed from Luchins' thesis.* ***b:*** *Simulated results as produced by our model of resource-rational decision-making. Error bars indicate the standard error of the mean.*

action when put under time pressure (Betsch et al., 2004). Reaction times are also facilitated for response repetitions in serial choice reaction (Bertelson, 1965) and bandit tasks (Wu et al., 2019). Moreover, it has been argued that time pressure causes participants to rely more on intuitive decision-making (Kahneman & Frederick, 2002), making immediate outcomes more salient (Ariely & Zakay, 2001) such that people rely more on fast, recognition-based processes rather than slower, more analytical processes (Klein, 1993).

We transcribed the data from these speed experiments from tables 28-30 in Luchins' thesis. Participants were New York college and senior high school students. There were 153 participants in the plain group and 98 participants in the speedy group. Whereas 126 out of 153 participants from the plain group responded using E-solutions to problems C1 and C2, 95 out of 98 participants from the speedy group used E-solutions for the same problems. The speedy group, therefore, showed a stronger Einstellung effect than the plain group on C1 and C2 (0.97 vs. 0.82; $\chi^2 = 12.08$, $p < .001$, $BF \approx 187.8$). Furthermore, whereas 100 of the 153 participants in the plain group responded by using E-solutions to problems C3 and C4, 94 of the 98 participants of the speedy group did. Thus, participants in the speedy group also showed a stronger Einstellung effect on C3 and C4 (0.99 vs. 0.55; $\chi^2 = 31.8$, $p < .001$, $BF \approx 1.8 \times 10^7$).

We then again tried to reproduce the observed effects using our proposed resource-rational decision-making model. Time pressure can be induced in our models by lowering the inverse temperature parameter. We, therefore, set this parameter to $\beta = 25$ (for the plain group) and to $\beta = 15$ (for the speedy group) to match the assumption of both conditions. For the plain condition, we found that 90 of 100 simulated agents responded using E-solutions to problems C1 and C2, and 84 of 100 simulated agents responded using E-solutions to problems C3 and C4. For the speedy condition, 97 of 100 simulated agents solved problems C1 and C2 by applying E-solutions, and 98 of 100 of applied E-solutions to problems C3 and C4. Thus, our model can reproduce the observation of an increased Einstellung effect when put under time pressure.

Developmental differences

Luchins also tested several children using his experiment. His main interest was to assess how education and intelligence influenced the observed effect. After many studies in different schools, Luchins did not find any strong influence of general intelligence or education on the general size of the Einstellung effect. However, Luchins did not check how the effect changed with age itself. Only later, in the 1950s, Luchins and his colleagues could show that the Effect increased with age (A. S. Luchins & Luchins, 1959; Ross, 1952), even when controlled for IQ. Recent research has shown that children can sometimes explore more than adults (Schulz et al., 2019), which can even lead to better performance when different hypotheses need to be considered (Gopnik et al., 2017). It has also been shown that 5-year-old children show no signs of functional fixedness (German & Defeyter, 2000).

Interestingly, Luchins' original data already contained the necessary information to compare the Einstellung effect across different age groups. Specifically, there were four different age groups in Luchins' data, which can be separated into primary school children, junior high school students, senior high school students, and freshmen college students. In total, there were 504 different subjects in this data set, which

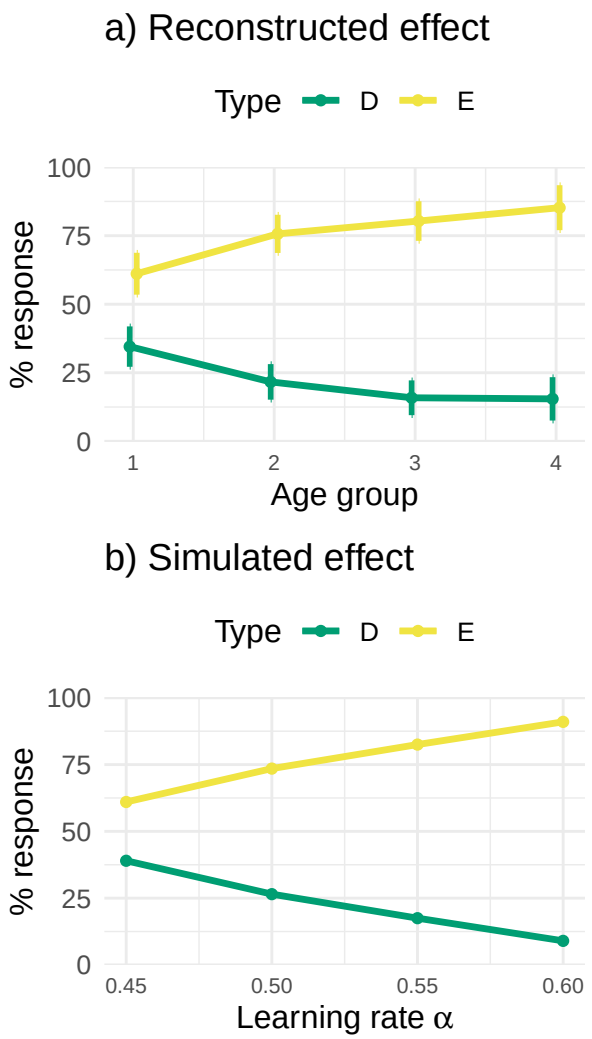


Figure 8
Developmental differences in the Einstellung effect. ***a:*** *Empirical results as reconstructed from Luchins' thesis.* ***b:*** *Simulated results as produced by our model of resource-rational decision-making. Error bars indicate the standard error of the mean.*

we split into the aforementioned groups by creating four different age ranks, from very young children to adults. Reanalyzing this data, we found that the proportion of responding to C1 and C2 by using an E-solution increased with age ($r(503) = 0.19$, $t = 4.39$, $p < .001$, $BF \approx 1207$), while the proportion applying D-solutions in the same tasks decreased ($r(503) = -0.18$, $t = -3.99$, $p < .001$, $BF \approx 242.7$).

We tried to replicate this effect using our model of resource-rational decision-making. To account for the differences in the age groups' Einstellung effects, we manipulated the learning rate α to model differences in learning for the different age groups. This can be justified by previous studies that have shown an increase in learning rates for older children and adolescents (Davidow et al., 2016; Master et al., 2020), and research that argued learning rates become more adaptive to task demands as children get older (Meder et al., 2021; Nussenbaum & Hartley, 2019). In particular, we used four different learning rates $\alpha = \{0.45, 0.5, 0.55, 0.66\}$ to account for differences in learning. We found that the agents responded more frequently by using E-solutions (61, 75, 82 and 91 out of 100 simulations) and less frequently by using D-solutions (39, 26, 18 and 9 out of 100 simulations) to C1 and C2 as the learning rate increased. Thus, our model was able to reproduce the observed developmental differences through a manipulation of the learning rate.

Ablation analysis

Finally, we wanted to investigate which parts of our model are crucial for capturing Luchins' effects and therefore conducted an ablation analysis in which we removed individual components of the model. The following ablations were considered in our analysis:

1. A fully-rational model that does not include a cost for transforming prior preferences into posterior policies, and therefore does not take any mental effort in consideration. This model is a limiting case of the full model in which the inverse temperature parameter β goes to infinity. Because this model does not use prior information at all, it is invariant to changes in the learning rate α.

2. A model that does not adjust its prior preferences to the decision-making environment, i.e., one that does not learn. This model is a special case of the full model in which we restrict the learning rate to $\alpha = 0$.

3. A model that does not include a term for penalizing complex solutions in its utility function, and therefore does not take any physical effort into consideration. This model is a special case of the full model in which we set the complexity parameter to $\lambda = 0$.

As before, we simulated these models for 100 runs in each of Luchins' conditions. The outcome of this ablation analysis is summarized in Table 1. We discuss each individual result in more detail below. The complete simulation results can be found in the accompanying code repository: https://github.com/marcelbinz/Einstellung.

No cost for mental effort ($\beta \to \infty$)

Let us first consider the fully-rational model. This model always selects the option with the highest utility regardless of the observation history. Therefore, it will solve C1-C4 using the D-solution in both the plain and control condition, and

Table 1
Model ablation analysis. If a model can reproduce an effect, then this is indicated by a check mark (✓), whereas a failure to do so is indicated by a cross (✗).

	Full model	No mental effort $\beta \to \infty$	No prior adaptation $\alpha = 0$	No physical effort $\lambda = 0$
Einstellung effect	✓	✗	✗	✓
C1/C2 versus C3/C4	✓	✗	✗	✗
Don't be blind	✓	✗	✓	✗
Number of tasks	✓	✗	✗	✓
Interleaved tasks	✓	✗	✗	✓
Time pressure	✓	✗	✓	✗
Developmental differences	✓	✗	✗	✓

consequently not show an Einstellung effect. The same reasoning can be applied to recognize that this model does not exhibit a difference in the percentage of E-solutions between C1/C2 and C3/C4 (in both cases it applies the D-solution without exception). Because the fully-rational model is invariant to the observation history, it also does not capture changes that arise from varying the number of tasks or from interleaving E- and D-problems during training. Furthermore, the fully-rational model does not contain any parameter that could be manipulated to emulate the changes Luchins observed in the DBB condition, under time pressure, or when inspecting participants in different age groups.

No adaptation of prior preferences ($\alpha = 0$)

We may also consider a model that does not adjust its prior preferences to the encountered decision-making environment. In this model, the probability of selecting the E-solution in C1-C4 increases as the inverse temperature parameter decreases. However, it is dominated by the probability of selecting the D-solution for all values of β (see Fig. 9). Therefore, this model is unable to account for the main Einstellung effect. In addition, we can note that the average utilities for C1/C2 and C3/C4 are identical, and thus the model shows no difference in the percentage of E-solutions in those two phases. Because prior preferences remain constant over time, the model is invariant to the observation history. Therefore, it can not capture any effect from varying the number of tasks or from alternating between E- and D-problems during training. Furthermore, the model can not account for developmental differences, which we have previously done by manipulating the learning rate.

There are, however, two of Luchins' findings that this model can capture by manipulating the inverse temperature parameter. Increasing this parameter will increase the percentage of D-solutions, mirroring the results of the DBB condition. Decreasing this parameter, on the other hand, will increase the percentage of E-solutions, mirroring the results of Luchins' studies under time pressure. Taken together, this model only captures a very limited set of Luchins' results.

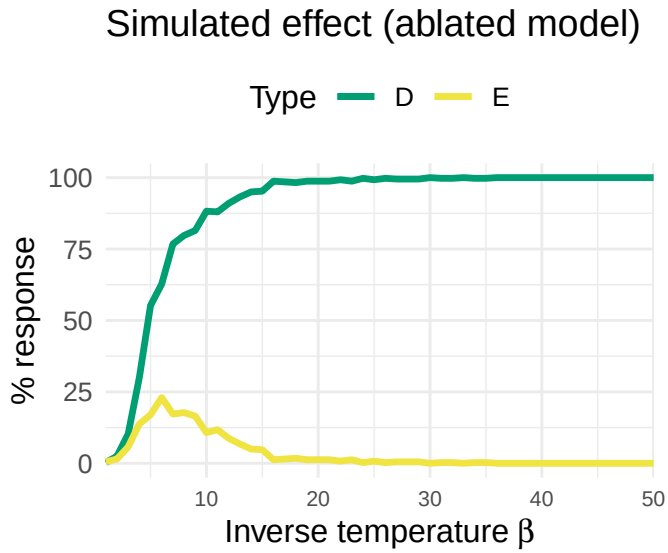


Figure 9
Percentage of D- and E-solutions for the model that does not adjust its prior preferences to the decision-making environment.

No cost for physical effort ($\lambda = 0$)

From the ablated models, only the one without the term for penalizing complex solutions captures the main Einstellung effect. This model still assigns the majority of its probability mass to the E-solution in C1-C4. However, the probability of selecting such solutions is lower than in the full model because a larger number of alternatives with equal utility has to be considered (e.g., the solution $w = [9, -4, 3]$ for $x = [23, 49, 3]$ and $y = 20$ is just as likely as $w =$

$[1, 0, -1]$). In contrast to the other two ablated models, this model adjusts its prior preferences to the encountered environment and is hence able to capture an increase in E-solutions after previous encounters with such problems, as well as a decrease in E-solutions when E- and D-problems are interleaved during training. Finally, we can also relate increases in learning rate within this model to the developmental differences found in Luchins' studies as we did for the full model.

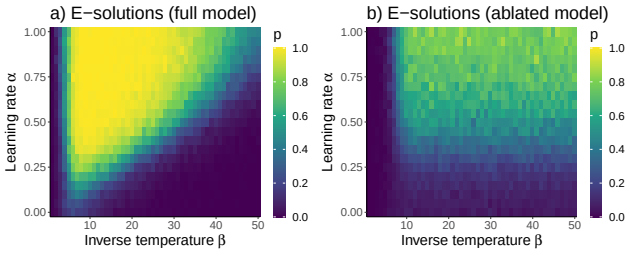


Figure 10

Percentage of E-solutions. ***a:*** *For the full model (reproduced from Fig. 2).* ***b:*** *For the model that does not include the term for penalizing complex solutions.*

However, we did not observe any difference in the percentage of E-solutions between C1/C2 and C3/C4. We also observed that the percentage of E-solutions in C1-C4 is largely independent of the inverse temperature parameter β (see Fig. 10b). Therefore, this model does not show an increase in E-solutions under time pressure and neither does it show an increase in D-solutions in the DBB condition. We speculate that this is the case because a higher number of solutions is equally valuable in the initial training phase, and hence an Einstellung effect is not developed for a fixed percentage of runs even when increasing the inverse temperature parameter β.

Summary

The previous ablation analyses demonstrated that each component of our model is crucial to capture the complete set of effects described by A. S. Luchins (1942). Not including a cost for transforming prior preferences into posterior policies prevented the model from showing any effect. If we remove the ability to adjust to the encountered decision-making environment, the model becomes invariant to the observation history and it is hence not able to replicate history-dependent effects. If we remove the term that favors simple solutions, the percentage of E-solutions becomes independent of the inverse temperature parameter, and hence the model can no longer account for human choices under time pressure and in the DBB condition. Finally, it is important to note that it is not just the sum of the components but the interaction between them that gives rise to the complete set of effects: the observation that the Einstellung effect is stronger in C1/C2 than in C3/C4 is captured by none of the ablated models but only arises from the combination of all components.

Discussion

The Einstellung effect is an empirical phenomenon in which previous experience with a problem leads to apparently inefficient solutions for a current problem. This effect was first established by Abraham Luchins and published in his doctoral thesis in 1942. In this doctoral thesis, Luchins collected data from over 1000 participants in different conditions. To make this data amenable to modern statistical analysis, we have reconstructed Luchins' results by transcribing his tables into digital data formats. Doing so, we were not only able to reanalyze the original Einstellung effect, but also many of Luchins' additional experiments that probed various factors influencing the effect. This showed that the Einstellung effect is higher for earlier than for later test problems, decreases when participants are told to not be blind, increases with the number of training problems of the same type, diminishes when problems that require different solutions are interleaved during training, increases under time pressure, and increases with participants' age.

It is typically assumed that the best solution for any particular problem is necessarily the shortest, and thus previous research has largely characterized the Einstellung effect as maladaptive behavior. In the present paper, we have challenged this assumption and provided a resource-rational interpretation of the effect. We did so with the help of an information-theoretic model of decision-making. The central premise of this model is to transform prior preferences into posterior policies in a way that trade off expected utility with the time it takes to make a decision. The resulting model incorporates three basic principles: (1) people prefer simple solutions, i.e., they attempt to spend as little physical effort as possible, (2) they avoid costly computations, i.e., those that require high mental effort, and (3) they adapt to their environment, i.e., they learn about statistics of the problem they interact with. We found that these simple principles are sufficient to capture the rich characteristics found in Luchins' data. An additional ablation analysis confirmed that all of these principles are necessary to reproduce the entire set of phenomena reported in Luchins' thesis.

Limitations and future directions

There are several limitations of our work that deserve to be mentioned. The first shortcoming concerns the data reported in Luchins' thesis. Luchins only collected population-level statistics, which prevented us from investigating individual differences. He also did not record participants' reaction times, which potentially could provide another window into how people approach the task. The model we have proposed especially makes strong predictions concerning the

latter, and it would be interesting to examine whether these predictions can be confirmed experimentally.

We have also only studied the Einstellung effect in a limited domain, namely Luchins' original water jar task. Luchins' doctoral thesis from 1942 already encompassed other domains in which people showed similar behavioral patterns, such as a maze navigation problem and a word puzzle task. Following Luchins' seminal work, these ideas have been extended to an even wider range of domains, including chess (Bilalić et al., 2008a, 2008b, 2010; Reingold et al., 2001) and medical diagnostics (Croskerry, 2003; Graaf, 1989). To further validate our model, we plan to apply it to these domains in future studies. In this context, it is noteworthy to mention that earlier work of Braun et al. (2011) has already extended the modeling principles discussed here to the setting of path integral control, which could be used to build agents for Luchins' maze navigation problem. Moreover, our model could also be adapted to explain human behavior in more complex planning tasks, including the observed Einstellung effect of professional chess players, by equipping deep learning models' planning capacity with resource-rational decision-making.

Furthermore, our analysis focused on Marr's computational level of analysis (Marr, 1982), meaning that it answers the question of what goal a decision-maker is trying to achieve. It does not provide a concrete mechanistic description of the decision-making process. In particular, we have suggested that a resource-rational decision-maker can use importance sampling to draw a low-bias sample from its posterior policy. In the outlined importance sampling procedure, the decision-maker draws multiple samples from the prior in parallel and then selects one in proportion to their importance weights. Doing so is theoretically appealing as it allows us to link the KL divergence between prior preferences and posterior policy to the time it takes to make a decision. However, we imagine that this process looks quite different within the human mind. For example, it might be possible that people do inspect samples sequentially instead of in parallel, or that they are willing to accept some bias in order to make even quicker decisions.

Finally, we do not claim that the proposed model is the only way to explain the data from Luchins' experiments. There are, for example, several alternative approaches to constrain computational resources that could induce similar behavior. In this context, the frameworks of rate-distortion theory (Genewein et al., 2015; S. J. Gershman, 2020) and rational meta-reasoning (Lieder & Griffiths, 2020; Russell & Wefald, 1991) provide two obvious alternative theories. Another interpretation of the presented results is that people selectively reuse outcomes from previous computations, i.e., that they engage in amortized inference (Dasgupta et al., 2018; S. Gershman & Goodman, 2014). It might also be possible that similar decision-making strategies can be meta-learned and implemented via deep neural networks (Binz et al., 2020; Dasgupta et al., 2020).

Related work

Our proposed model can be connected with several lines of historical findings. First of all, it has been well-documented that animals sometimes repeat actions in a maladaptive fashion, leading to fixation (Krechevsky & Honzik, 1932) as well as stereotyped behavior (Dantzer, 1986). Furthermore, the human lack to change one's line of thinking has also been studied in the domain of rigidity (Schultz & Searleman, 2002), which is normally seen as the inability to change one's habits or attitude. We believe that our model of resource-rational decision-making could also explain some of these past findings as an adaptation to an environment that requires agents to save mental and physical effort.

From a modeling perspective, there exists a range of resource-rational approaches that are closely connected to the one we have employed (Bhui et al., 2021; S. J. Gershman et al., 2015; Lieder & Griffiths, 2019). Most relevant to our work are approaches based on rate-distortion theory (Tishby et al., 2000). In this framework, one attempts to maximize some measure of performance, while simultaneously placing an upper bound on the number of bits required to store an object of interest. S. J. Gershman (2020) applied this idea to limit an agent's policy complexity and showed that this accounts for perseveration effects in human learning. Mathematically, this can be realized by maximizing the objective from Equation 2 averaged over all data-points. While, on the surface, this account seems quite similar to the one we have suggested, the resulting interpretation is quite different. A rate-distortion theoretic approach stores a policy for each data-point in memory and simply queries it whenever a particular data-point is encountered. In contrast, our approach does not keep a policy in memory but instead searches for a solution from scratch once a new data-point is encountered. We may ask ourselves which of the two approaches is appropriate under which conditions. If a problem is frequently encountered, it might make sense to cache the policy for that problem in memory (as in the rate-distortion theoretic approach). If, however, the specific problem is rarely encountered, or storing a policy for it is expensive, one is better off by searching for new solutions from scratch upon interacting with the problem (as in the approach employed in this work). In Luchins' water jar task, a problem is never encountered twice, and thus we believe that it is more appropriate to search for solutions from scratch in this setting instead of storing each of them in memory.

Conclusion

We have combined the reconstruction of historical data with the use of modern computational tools to reinterpret a classic effect of human problem-solving. To the best of our

knowledge, we are not aware of any prior work that involved such a combination of methods. In particular, we have explained the results of Luchins' doctoral thesis (A. S. Luchins, 1942) using a computational model based on three simple principles: (1) people prefer simple solutions, i.e., they want to reduce physical effort, (2) they avoid costly computations, i.e., they want to reduce mental effort, and (3) they adapt to the environment, i.e., they learn over time. Ever since Herbert Simon's work bounded rationality (Simon, 1972), psychologists have argued that models of human behavior should be based on these principles. However, only recently researchers have been able to translate these principles into computational models. Having access to these new models allows us to reevaluate historical data from a fresh perspective, not only in the context of the Einstellung effect but also in domains beyond.

References

Ariely, D., & Zakay, D. (2001). A timely account of the role of duration in decision making. *Acta psychologica*, *108*(2), 187–207.

Bertelson, P. (1965). Serial choice reaction-time as a function of response versus signal-and-response repetition. *Nature*, *206*(4980), 217–218.

Betsch, T., Haberstroh, S., Molter, B., & Glöckner, A. (2004). Oops, i did it again—relapse errors in routinized decision making. *Organizational behavior and human decision processes*, *93*(1), 62–74.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Bilalić, M., McLeod, P., & Gobet, F. (2008a). Inflexibility of experts—reality or myth? quantifying the einstellung effect in chess masters. *Cognitive psychology*, *56*(2), 73–102.

Bilalić, M., McLeod, P., & Gobet, F. (2008b). Why good thoughts block better ones: The mechanism of the pernicious einstellung (set) effect. *Cognition*, *108*(3), 652–661.

Bilalić, M., McLeod, P., & Gobet, F. (2010). The mechanism of the einstellung (set) effect: A pervasive source of cognitive bias. *Current Directions in Psychological Science*, *19*(2), 111–115.

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2020). Heuristics from bounded meta-learned inference.

Birch, H. G., & Rabinowitz, H. S. (1951). The negative effect of previous experience on productive thinking. *Journal of experimental psychology*, *41*(2), 121.

Braun, D. A., Ortega, P. A., Theodorou, E., & Schaal, S. (2011). Path integral control and bounded rationality. *2011 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*, 202–209.

Chrysikou, E. G., & Weisberg, R. W. (2005). Following the wrong footsteps: Fixation effects of pictorial examples in a design problem-solving task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1134.

Crilly, N. (2015). Fixation and creativity in concept development: The attitudes and practices of expert designers. *Design studies*, *38*, 54–91.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, *78*(8), 775–780.

Dantzer, R. (1986). Behavioral, physiological and functional aspects of stereotyped behavior: A review and a re-interpretation. *Journal of Animal Science*, *62*(6), 1776–1786.

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67–81.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological review*, *127*(3), 412.

Davidow, J. Y., Foerde, K., Galván, A., & Shohamy, D. (2016). An upside to reward sensitivity: The hippocampus supports enhanced reinforcement learning in adolescence. *Neuron*, *92*(1), 93–99.

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychological monographs*, *58*(5), i.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American psychologist*, *49*(8), 725.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard business review*, *85*(7/8), 114.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational behavior and human performance*, *23*(3), 339–359.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322.

Funke, J. (2012). Complex problem solving. *Encyclopedia of the Sciences of Learning (682-685). Heidelberg: Springer*.

Genewein, T., Leibfried, F., Grau-Moya, J., & Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, *2*, 27.

German, T. P., & Defeyter, M. A. (2000). Immunity to functional fixedness in young children. *Psychonomic Bulletin & Review*, *7*(4), 707–712.

Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. *Proceedings of the annual meeting of the cognitive science society*, *36*(36).

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

Graaf, E. D. (1989). A test of medical problem-solving scored by nurses and doctors: The handicap of expertise. *Medical education*, *23*(4), 381–386.

Havasi, M., Peharz, R., & Hernández-Lobato, J. M. (2018). Minimal random code learning: Getting bits back from compressed model parameters. *arXiv preprint arXiv:1810.00440*.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, *49*, 81.

Klein, G. (1993). Sources of error in naturalistic decision making tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *37*(4), 368–371.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Normative, descriptive and methodological challenges. *Behavioral & Brain Science*, *19*, 1.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological science*, *19*(6), 585–592.

Krechevsky, I., & Honzik, C. H. (1932). Fixation in the rat. *University of California Publications in Psychology*.

Lane, D. M., & Jensen, D. G. (1993). Einstellung: Knowledge of the phenomenon facilitates problem solving. *Proceedings of the human factors and ergonomics society annual meeting*, *37*(18), 1277–1280.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 1–85.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of einstellung. *Psychological monographs*, *54*(6), i.

Luchins, A. S. (1951). On recent usage of the einstellung-effect as a test of rigidity. *Journal of Consulting Psychology*, *15*(2), 89.

Luchins, A. S., & Luchins, E. H. (1959). Rigidity of behavior: A variational approach to the effect of einstellung.

Maier, N. R. F. (1930). Reasoning in humans. i. on direction. *Journal of Comparative Psychology*, *10*(2), 115–143. https://doi.org/10.1037/h0073232

Maier, N. R. (1931). Reasoning in humans. ii. the solution of a problem and its appearance in consciousness. *Journal of comparative Psychology*, *12*(2), 181.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt; Co., Inc.

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. (2020). Distentangling the systems contributing to changes in learning during adolescence. *Developmental cognitive neuroscience*, *41*, 100732.

Meder, B., Wu, C. M., Schulz, E., & Ruggeri, A. (2021). Development of directed and random exploration in children. *Developmental Science*, e13095. https://doi.org/10.1111/desc.13095

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140.

Neroni, M. A., & Crilly, N. (2020). How to guard against fixation? demonstrating individual vulnerability is more effective than warning of general risk. *The Journal of Creative Behavior*.

Nussenbaum, K., & Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental cognitive neuroscience*, *40*, 100733.

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E., & Tishby, N. (2015). Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*.

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, *256*(5059), 1018–1021.

Reingold, E. M., Charness, N., Schultetus, R. S., & Stampe, D. M. (2001). Perceptual automaticity in expert chess players: Parallel encoding of chess relations. *Psychonomic Bulletin & Review*, *8*(3), 504–510.

Rock, I. (1957). The role of repetition in associative learning. *The American journal of psychology*, *70*(2), 186–193.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900.

Ross, V. M. (1952). A comparison of the effect of einstellung in different age groups.

Russell, S., & Wefald, E. (1991). Principles of metareasoning. *Artificial intelligence*, *49*(1-3), 361–395.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Saxe, A. M. (2013). Precis of deep linear neural networks: A theory of learning in the brain and mind.

Schultz, P. W., & Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genetic, social, and general psychology monographs*, *128*(2), 165.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological science*, *30*(11), 1561–1572.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Simon, H. A. (1990). Bounded rationality. *Utility and probability* (pp. 15–18). Springer.

Suzuki, S. (2020). *Zen mind, beginner's mind*. Shambhala Publications.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Tolman, E. C. (1934). Theories of learning.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Wu, C., Schulz, E., Gerbaulet, K., Pleskac, T., & Speekenbrink, M. (2019). Under pressure: The influence of time limits on human exploration. *41st Annual Conference of the Cognitive Science Society (CogSci 2019)*, 1219–1225.