The Cognitive Complexity of Rule Changes

Sara Todorovikj (sara.todorovikj@hsw.tu-chemnitz.de)
Daniel Brand (daniel.brand@hsw.tu-chemnitz.de)
Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology
Straße der Nationen 62, 09111 Chemnitz, Germany

Abstract

Concept change is a fundamental cognitive process that enables individuals to adapt to new and conflicting information. To investigate the mechanisms underlying such adaptations, we introduce the *Counting Game*, an abstract rule-based paradigm. In this article, we evaluate whether the paradigm effectively captures complexity differences between different types of manipulations. Our experiment involved counting tasks where participants had to apply rules that modify how certain objects are counted, enabling us to examine the effects of perceptual complexity, rule operations, rule interactions, temporal dependencies and scope. We analyzed accuracy and response times to assess whether these manipulations elicit desired effects. Additionally, we constructed predictive models to identify key features influencing task difficulty. By evaluating the theoretical soundness of the *Counting Game*, we establish an empirical foundations for future studies on forgetting operations, among other concept changes.

Keywords: Cognitive Complexity; Concept Change; Rule Updating; Forgetting Operations

Introduction

Rule changes involve adapting to new or conflicting information and play a crucial role in problem solving, decision making and learning. Whether adjusting to a new environment or integrating new knowledge, humans continuously modify existing and previously known concepts and rules in order to navigate the world. Consider an everyday action, such as crossing the street. In Paris, you must first look left to check for oncoming traffic. However, should you travel to London, you must immediately switch to looking right to avoid dangerous situations. Applying the wrong rule at the wrong time can have serious consequences. Fortunately, most of our daily activities involving rule adjustments – such as adapting to a new operating system or learning to play a new board game – occur smoothly. Other changes, such as breaking an addictive habit, like smoking, can be significantly more challenging.

Humans continuously encounter new objects and situations, requiring them to use a range of cognitive mechanisms to categorize and interpret information. Understanding how we acquire and update such categorization rules is central in cognitive science. Experimental approaches in this field often use rule-based systems to analyze how individuals update concepts. For example, Brand, Dames, Puricelli, and Ragni (2022) conducted a study to investigate the cognitive costs of adding new categorization rules or altering previously learned rules. They found that modifying an existing rule was less cognitively demanding than adding a new one indicating the presence of processes depending on the number or complexity of the rules. Additionally, they found that categorization performance improves when new rules align with old ones, but in case of conflicting situations, this was no longer apparent. Ultimately, they show how rule-based categorization involves cognitive costs that depend on rule complexity. At their core, rules function as if-then conditions: if *a precondition is satisfied* then *a corresponding action follows*. This structure underlies processes such as learning to categorize (Maddox, Ashby, Ing, & Pickering, 2004; Maddox, Ashby, & Bohil, 2003), implementing intentions (Oettingen & Gollwitzer, 2010; Gollwitzer & Brandstätter, 1997) and is central to cognitive architectures like ACT-R (Anderson, 2007), among others. Studying how rules are acquired, updated and modified is essential for understanding how individuals adapt to changing environments.

However, rule change is not always a simple rule update. Sometimes information and knowledge is simply removed or made conditional depending on the context. These operations can be recognized as forgetting operations, where individuals adapt their knowledge by removing, modifying or replacing learned information. Forgetting, as a phenomenon of everyday life, involves cognitive flexibility in dynamic environments. Various forgetting operations explain how information is selectively discarded or restructured. These processes have been formalized by Beierle, Kern-Isberner, Sauerwald, Bock, and Ragni (2019) as a structured classification of different types of forgetting in common-sense belief management. The framework has been primarily explored from a logical perspective, an empirical investigation of the cognitive costs associated with these forgetting operations remains an open question.

Towards empirically testing formal frameworks of rule change, we present a new experimental paradigm - the *Counting Game*, that allows us to test rule changes in an abstract environment. Given 16 shapes in different colors, individuals are tasked with finding out which color has the most elements as quickly as possible. Throughout the game, they are presented with different rules that change the counting system, such as "Blue stars count twice.". The paradigm allows for easy manipulation of the rule-based counting system. For example, an intentional removal of a specific rule can be simply expressed by "Blue stars count as one again.". By analyzing

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

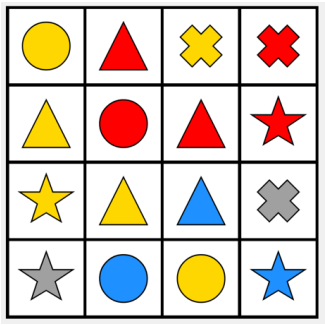

Figure 1: A representation of the Counting Game showing a trial whose winner is yellow. A rule "Red triangles count twice" changes the winner to red.

the individuals' correctness and response times, we get insight into the cognitive demands and costs of such operators. In order to isolate the effect of forgetting operations, we must first examine the complexity of the experimental tasks themselves. With that, we can establish what factors need to be accounted for and omitted in future investigations of the forgetting operations. In this article, we analyze the paradigm from two perspectives. First, we look into *content-level* features, examining the effect of color distributions in the tasks and the effect of different rule types. On the other hand, we investigate *process-level* features, focusing on the effect of rule interactions, the proximity between rules and the scope of rules. With this we capture various aspects of task complexity. Ultimately, our goal is to understand better the cognitive mechanisms underlying rule changes and task complexity within this paradigm. Therefore, we construct predictive models incorporating perceptual and rule-based features, evaluating how trial complexity, rule operations and structural factors contribute to performance. We interpret accuracy as an indicator of correct rule application and response time (RT) as a representation of cognitive load and retrieval difficulty. Differences in these measures reflect either increased cognitive load or interference. We compare different feature combinations and identify the key predictors of accuracy and RTs. We aim to answer the following research questions:

[RQ1] Is the counting paradigm suitable for investigating complexity differences between operations?

[RQ2] What factors contribute to the cognitive complexity of rule changes in our counting paradigm?

Method

The goal of this study was to investigate the effect of altering a set rule system on participants' performance measured by their accuracy and response times. In order to achieve that, we developed an abstract experimental paradigm based on shapes and colors, called the *Counting Game*. The game consists of *trials* - grids of 16 objects that vary in their *shape* (circle, triangle, cross, star) and *color* (red, yellow, blue, gray), and its winner, i.e. the color that has the largest number of objects (Fig. 1). The winner can be red, yellow or blue, while gray elements are used as optional filler objects. At most half of the total number of elements can be of the same color, to prevent a single color from dominating the trial.

Participants are presented with multiple trials whose winners they need to determine as fast as possible. There is a scoring system that awards them points for fast and accurate answers. Each correct answer brings them 10 points and they can get 1 additional point per second left on a timer that counts down from 20s. Incorrect answers do not award points, regardless of the response time.

Rules

The baseline when starting the game is that each object, regardless of color and shape, counts as one. We introduce two types of rules that change how certain objects are counted:
1. *Count-twice* rules state that an object of a certain color and shape will count twice from now on - e.g., "Blue triangles count twice."
2. *Count-as* rules state that an object of a certain color and shape will now count as if it was a different color - e.g. "Yellow stars count as red."

For all rules there are certain trials that are *critical* - a successful rule application changes their winner. E.g., applying the rule "Red triangles count twice" to the trial in Fig. 1 changes the winner from yellow to red Using these rules we implement the following four intentional forgetting operators (see Beierle et al., 2019) with the aim to alter the individuals' set of rules. Table 1 shows example contents for the below explained rule implementations.

Contraction describes the intentional removal of a specific rule from a rule set, resulting in a state where the contracted rule is no longer accepted. We implement this using one *count-twice* rule that is explicitly withdrawn later.

Revision involves updating a rule set with new information that should be prioritized over existing rules, which may involve *implicitly* forgetting conflicting rules. We implement this with two *count-as* rules, where the affected objects are the same, but the new target color is different, implicitly forcing the individual to forget the first rule.

Conditionalization restricts rules by omitting ones that do not satisfy a given precondition. We implement this using specific *count-as* rules where the initial color is gray (optional elements), introducing the precondition that *if* there are gray shapes, they will count as another color.

Fading out is a gradual forgetting process where information becomes increasingly difficult to retrieve over manipulations. We implement this using one *count-twice* rule followed by other filler rules that make it more difficult to retrieve the first rule. This aligns with the concept of *retroactive interference*, as new rules make previous ones vulnerable to decay (Ricker, Nieuwenstein, Bayliss, & Barrouillet, 2018; Dames & Oberauer, 2022).

In the study, a *Phase* refers to a distinct stage within a round where a specific rule configuration is active. Each round starts with *Phase 0* and progresses through multiple

Table 1: Example contents for the two rounds of each forgetting operator and the manipulations introduced in the second rounds that aim to increase the complexity. Filler rules are typed out in *italics* and are included to increase the complexity or to "fill" the rounds that need less than 4 rules to implement a forgetting operator.
Notation: FO - Forgetting Operators; Contr. - Contraction; Rev. - Revision; Cond. - Conditionalization; Fade - Fading Out.

FO	Round 1	Round 2	Manipulation
Contr.	Blue stars count twice. Blue stars count as one again. Yellow circles count twice. Yellow circles count as one again.	Blue stars count twice. Yellow circles count twice. Blue stars count as one again. Yellow circles count as one again.	Multiple rules active simultaneously & contraction of a previous rule that is not the last one presented.
Rev.	Blue circles count as yellow. Blue circles count as red. *Blue circles count as blue again.* *Yellow crosses count twice.*	Blue circles count as yellow. *Blue stars count as yellow.* Blue circles count as red. *Yellow crosses count twice.*	Additional filler rule between the original rule and its revision.
Cond.	Gray shapes count as red. *Blue stars count twice.* *Blue stars count as one again.* *Yellow triangles count twice.*	Gray triangles count as red. *Blue stars count twice.* *Blue stars count as one again.* *Yellow triangles count twice.*	Contrast between rules applied to all shapes of a color (general) and specific shapes.
Fade	Red triangles count twice. *Blue crosses count twice.* *Yellow stars count twice.* *Blue crosses count as one again.*	Red triangles count twice. *Blue crosses count twice.* *Blue stars count as yellow.* *Blue crosses count as one again*	Effect of *count-twice* vs. *count-as* filler rules on fading out.

phases, with a new phase beginning whenever a rule is introduced, updated or removed. So, *Phase 1* starts after the first rule is applied, *Phase 2* after the second, and so on. During each phase, the participants are presented with the same trials in a randomized order. This structure allows for us to examine how participants adapt to increasing rule complexity and how prior rules influence the application of new ones.

Complexity

In order to analyze the cognitive complexity, we will analyze the participants' answer accuracy and response times, based on hypothesized influencing features.

Content-level Features The content-level features describe the nature of individual trials and rules. We characterize the color distribution of the trials in terms of how dominant the winner is, how closely contested the trial is and how evenly the colors are distributed. These aspects capture the perceptual complexity of the trials, determining how much visual comparison is needed to identify the correct response when using fast-and-frugal strategies to estimate the winner. On the other hand, we differentiate between rules that require individuals to perform numerical/quantitative adjustments (*count-twice*) and rules that require categorical reasoning and thinking (*count-as*).

Hypothesis (Perceptual Complexity): Trials with clearly dominant winners should be accurate and fast, as they require minimal comparison. If the winner is only slightly ahead of another color, the difficulty increases, requiring a closer evaluation. The most difficult trials should be those where the competing colors are similarly distributed, as they do not provide an obvious visual advantage.

Hypothesis (Operational Complexity): Count-as rules are more difficult than *count-twice* rules as they alter the "identity" of certain shapes, forcing participants to override and mentally reassign objects according to the given rule, while the latter only change the "weight" of the shapes when they are encountered within the game.

Process-level Features We implement and investigate process-level features in our paradigm through certain manipulations: *interaction complexity* - the effect of simultaneous rule applications on task difficulty; *temporal complexity* - the influence of temporal proximity between a rule and its contraction/revision; and *scope complexity* - the difficulty of applying rules with specific vs. general scopes. Examples of these manipulations are shown in Table 1.

Hypothesis (Interaction Complexity): In the case of two rules, the application of the one rule should be more difficult when another one is simultaneously applicable, as managing overlapping rules should increase the cognitive load.

Hypothesis (Temporal Complexity): The contraction or revision of a rule should be more difficult when there are additional rules between them and the rule they aim to change, as a greater temporal and structural separation would make it harder to retrieve relevant information.

Hypothesis (Scope Complexity): Applying rules to specific shapes of a certain color should be more difficult than to all shapes of the color in general, as general rules require less

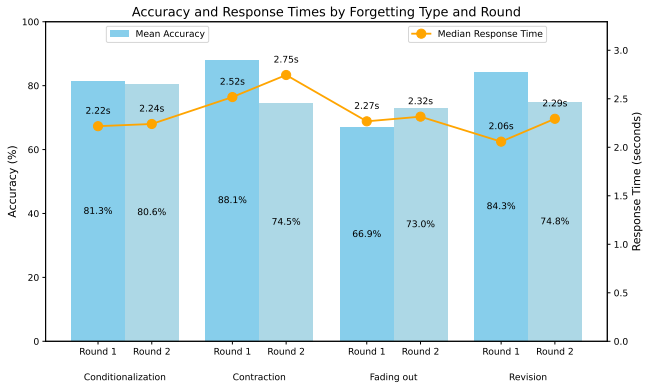

Figure 2: Average accuracy and response time for each forgetting operation and round.

information to be remembered and verified.

Procedure

The four forgetting operators translate into four participant groups, each being presented with two rounds, whose presentation order is randomized. This structure enables a controlled comparison of manipulations, and lays the basis for future work exploring individual differences in rule adaptation. In order to ensure that no unwanted effects or bias are introduced, each participant gets assigned their colors and shapes randomly. That means that a rule "$color_1$ $shape_1$ count twice" might be about red stars for one person and blue circles for another. The structure of the rounds and relevant rules is preserved otherwise. The position of specific objects in a trial is also randomized each time it is presented. Participants responded by clicking on one of three buttons colored with the target colors (red, yellow, blue) using their mouse.

We obtained data from 98 participants (age 18-74, 51% female) recruited on Prolific[1]. The experiment was performed online as a web-experiment. As compensation, the participants received 4.5 GBP after completing the experiment. All participants were native English speakers.

After giving consent, participants were introduced to the possible shapes and colors and their goal to find the winner among the red, yellow and blue elements. They were then introduced to the scoring system, explaining that they aim to get as many points as possible by providing correct answers fast. Afterwards, they had a practice round with a few tasks and rules, in order to get accustomed to the game. Once finished, they started with the study. In between the two rounds they were presented with their score and challenged to beat it in the second round. In the following analysis, we focus on RT and accuracy as indicators of cognitive processes. A more targeted analysis of score-based strategies will be explored in future work. The data, analysis and modeling scripts developed for this article are available on GitHub[2].

[1]https://www.prolific.com

[2]https://github.com/saratdr/CogSci2025-CountingGame

Analysis

We identified 3 participants with average accuracy scores diverging by more than 2 standard deviations away from the mean as outliers and excluded them from the following analysis. The final number of participants is 95. The overall accuracy across all participants and conditions was 78.91% (median: 100.00%, SD = 40.79%). The mean accuracy of trials critical for at least one rule was 68.12%, in contrast to 90.15% for non-critical trials. The median response time across all trials was 2.91s. Critical trials had a negligible increase in median response time (2.35s) compared to non-critical trials (2.30s). Figure 2 displays the average accuracy and response time for each forgetting type and round. Note that these results capture not only the relevant operators but also any filler rules that might be present. In the case of contraction and revision, the accuracy drops in the second rounds, while the RT is increased. Fading out shows a slightly better accuracy in round 2 with a similar RT and conditionalization shows similar results for both rounds.

Trial Color Distribution Features

To describe the trial structures, we consider three features that quantify the color distribution. **Winner dominance** (WinDom) is the ratio between the winner and all elements, indicating if a winner is clearly dominant with a higher value. **Lead** is the difference between the winner and the second-place color, a smaller value indicating "tight" competition. **Imbalance** (Imbal) quantifies how evenly distributed the three eligible colors are by considering the difference between the highest and lowest counts with a lower value suggesting a more balanced trial. When describing a trial, these features interact with each other, e.g., a trial with a dominant winner can have one dominant color, or also be in close competition with the second-place color.

We examined the relationship between perceptual trial features and task performance using point-biserial correlations for accuracy and Spearman rank correlations for RT. This analysis was conducted only on trials where no rules were active, ensuring that the observed effects reflect purely perceptual complexity rather than rule-based transformations. Holm-Bonferroni corrections were applied to account for multiple comparisons, all reported p-values are corrected. The results are presented in Table 2. To analyze the effect of clearly dominant winners with no close competition, we consider the product between WinDom and Lead in the following. It is positively correlated with accuracy and negatively with RT, suggesting that participants responded quickly and accurately in trials with a clear dominant winner. Lead is positively correlated with accuracy but negatively with RT, which tells us that participants responded faster and more accurately for trials that had a bigger difference between the winner and second-place color. Finally, Imbal has negative correlations with both accuracy and RT, indicating that participants answered slower and with higher accuracy when trials had more balanced color distributions. Our hypothesis regarding Win-

Table 2: Correlation values between the three trial distribution features Winner Dominance (WinDom), Lead and Imbalance (Imbal), and the accuracy (point-biserial) and RT (Spearman) of participant answers. Significant (corrected) p-values are marked in bold.

Feature	Accuracy	p	RT	p
WinDom × Lead	0.11	$<$ **.001**	−0.12	$<$ **.001**
Lead	0.14	$<$ **.001**	−0.09	$<$ **.001**
Imbal	−0.11	$<$ **.001**	−0.14	$<$ **.001**

Dom and Lead is confirmed, however Imbal reveals that while similarly distributed colors were more difficult and time consuming, that time was spent assessing the winner correctly.

Rule Operations

We investigate how *count-twice* and *count-as* rules affect task difficulty. These rules introduce different cognitive demands – *count-twice* requires a quantitative adjustment and *count-as* a categorical remapping – so, we expect their impact on accuracy and RT to be different. Since we are interested in the immediate impact of rule application, we restrict our analysis to *Phase 1*, where rules are first introduced. We conducted logistic regression analyses on the accuracy and Mann-Whitney U tests for RT, with Holm-Bonferroni applied to account for multiple comparisons. We found that, in critical trials, *count-twice* rules lead to significantly lower accuracy than *count-as* ($\beta = -0.367, p = .005$), suggesting that *count-twice* is more difficult. We also examined whether rule application specifically affects critical trials. We confirmed that non-critical trials were significantly easier than critical trials for both *count-twice* ($\beta = 1.323, p < .001$) and *count-as* rules ($\beta = 1.622, p < .001$). In terms of RT, we found that applying *count-twice* took significantly longer ($U = 142895.00, p < .001$), suggesting higher cognitive load. The comparisons of RTs between critical and non-critical trials within each rule type revealed no significance ($p = .078$ for both), suggesting that RT differences are primarily driven by rule type rather than trial criticality. Contrary to our initial hypothesis, *count-twice* rules caused greater cognitive load, as reflected in both accuracy and RT. This suggests quantitative adjustments may be more demanding than overriding categorical mappings.

Rule Interaction

Here, we focus on the two contraction rounds, specifically on Rule 2 (e.g., "Yellow circles count twice" in Table 1). In the first round, Rule 2 is first active in *Phase 3* by itself and in the second round it is already active in *Phase 2* with Rule 1 also present. By examining the performance in these cases, we assess the interaction complexity – how cognitive load changes when multiple rules are applicable sequentially vs. simultaneously. We analyze three types of trials: R2-Crit - critical *only* for Rule 2, R1R2-Crit - critical for both rules and non-critical trials, as a baseline. We conducted logistic regression analyses on the accuracy and Mann-Whitney U tests on RTs, both corrected with Holm-Bonferroni. The results show a significant drop in accuracy in Round 2 for R2-Crit ($\beta = -3.482, p < .001$), R1R2-Crit ($\beta = -1.735, p < .001$) and non-critical trials ($\beta = -1.150, p < .001$). The RT analysis showed a marginally significant difference between rounds for R2-Crit and R1R2-Crit trials ($p = .056$ for both). However, the RTs for non-critical trials was significantly longer in the second round ($p = .003$). These results confirm our hypothesis that applying a rule simultaneously with another one is more difficult, as reflected through the accuracy, while the RT analysis suggests that though that task is more difficult, there is no difference in the cognitive effort. Moreover, we find out that even non-critical trials are affected, suggesting an increased cognitive load effect in general.

Rule Interaction and Operations Additionally, we investigate the difference between rule types (*count-twice* vs. *count-as*) in the case when other rules are simultaneously active, specifically Rule 3 in *Phase 3* of the Fading Out rounds. Even in these circumstances, we confirm the previously found results that *count-as* rules are easier to apply than *count-twice* rules ($\beta = 0.597, p = .042$), however, this time with no significant difference in RTs ($p = .387$).

Temporal Effects

In this section we report how temporal separation between a rule and its revision affects task difficulty. We focus on revision Round 1 with an immediate revision (*Phase 2*) and Round 2 with a delayed revision and a filler rule in between (*Phase 3*). The relevant trials are R1R2-Crit, critical for both the original rule and its revision. A logistic regression analysis revealed a significant decrease in accuracy in Round 2 ($\beta = -0.865, p = .018$) and a Mann-Whitney U test showed that RTs are significantly longer in Round 2 ($p < .001$). This confirms our hypothesis that a longer delay including filler rules leads to more difficulty with revising previous information, as they may disrupt memory retrieval.

Scope of Rules

We test whether applying rules to specific subsets (e.g., "Gray stars count as red") is more difficult than applying them to a broader category (e.g., "Gray shapes count as red"). We focus on conditionalization Round 1 with a general rule and Round 2 with a specific one, both active in their respective *Phase 1*. A logistic regression model on trials critical for the relevant rules shows that accuracy is significantly lower for specific rules than for general ones ($\beta = -0.987, p = .008$), confirming our hypothesis. A Mann-Whitney U test on RTs showed that there were no significant differences ($p = .740$).

Modeling

In this section, we construct predictive models to determine which features best explain task complexity across the whole dataset. By analyzing a range of content-level and process-level features we aim to identify the key predictors of accu-

racy and response time. We include seven features, as described and analyzed above. The three trial color distribution features (**WinDom** - winnner dominance, **Lead**, and **Imbal** - imbalance), **RuleType** - *count-twice* vs. *count-as* vs. None, **CurrActive** - number of currently applicable rules, **RuleProximity** - if current rule is a contraction/revision, the distance to the relevant rule, **Scope** - specific or general rule. Each of these features captures different aspects of complexity, from purely perceptual influences to cognitive demands imposed by rule application.

We evaluated all possible combinations of the seven features. This resulted in a total of 127 tested models for each accuracy (linear classification) and RTs (linear regression)[3]. We determined the best models using the lowest Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) values. These two metrics provide us with a threshold after which the addition of parameters does not lead to a significant performance improvement, yet increases the tendency to overfit.

For accuracy, the best-performing model, as determined by the lowest AIC (-36167.33) and BIC (-36135.07) achieved an accuracy of 78.5%. For RTs, on the other hand, the best-performing model with lowest AIC (30763.10) and BIC (30803.43) achieved a mean absolute error of MAE = 1.32(s). The coefficients associated with the relevant features for both models are displayed in Table 3. Both, accuracy and RT, measures are affected by Lead, Imbal and RuleType. Lead likely encourages more careful processing, improving accuracy and reducing RT. Balanced color distributions make it harder and slower to determine the winner. WinDom only affects RT, with dominating colors leading to faster responses. RuleProximity also only affects RT, with greater distances between related rules leading to longer RTs. CurrActive, on the other hand, is only relevant for accuracy - a larger number of simultaneously active rules decreases the accuracy.

Table 3: Feature coefficients for the selected models predicting accuracy (LDA) and response time (Linear Regression).

Feature	Acc. Model (β)	RT Model (β)
WinDom	×	−1.19
Lead	0.41	−0.10
Imbal	−0.24	−0.06
RType-CountTwice	0.11	0.33
RType-CountAs	0.34	0.01
CurrActive	−0.66	×
RuleProximity	×	0.15
Scope	×	×

Discussion

We introduced the *Counting Game* paradigm developed as an abstract environment for testing forgetting operators. In this article, we focused on investigating the factors contributing to the complexity of rule changes in the paradigm, specifically how *content-level* and *process-level* features influence accuracy and response time (RT) during rule-based counting tasks. We tested five hypotheses, each targeting a specific aspect of task complexity in order to answer our research questions.

Our first research question, **RQ1**, addresses the suitability of our presented paradigm for investigating complexity differences in rule changing scenarios, specifically forgetting operations. The study we conducted implemented four forgetting operators and included different manipulations. With our analysis we showed that the difference in performance is meaningful and that the paradigm provides a suitable "playground", adequately sensitive to manipulations. In **RQ2** we address which factors contribute to the complexity of rule changes in our paradigm. We found that perceptual complexity influences task performance with strongly *dominant winners* making responses faster but slightly less accurate, while *balanced distributions* increase both RT and error rates. Operational complexity (*rule type*) also plays a major role, as *count-twice* rules were more difficult than *count-as* rules, suggesting that numerical transformations cause greater cognitive load than categorical remapping. Interaction complexity shows that when multiple rules are *simultaneously active*, accuracy drops significantly, indicating increased cognitive burden. Temporal complexity describes how the longer a revision or contraction is *separated* from the relevant rule, the more difficult it becomes to apply it, emphasizing the role of memory demands. Scope complexity confirms that rules applied to a *subset* of a category are harder to apply than rules applied to the *whole* category. In summary, these results confirm that rule change difficulty is not uniform – it depends on how the new information is structured perceptually, how it interacts with previous knowledge, and how memory demands and the ability to adapt, inhibit and apply rules influence performance. We identified the smallest feature subsets that influence rule change complexity, using LDA as classification models for accuracy and linear regression models for RTs.

In this article we identified how trial features and rule applications in the counting game paradigm contribute to task difficulty. This work lays the foundation for future studies within the paradigm that will explicitly investigate the cognitive costs of concept changes, including forgetting operations. These findings help estimate the impact on tasks that should be omitted and accounted for in effects and trend analyses, but also aid the task design for various other future investigations. This is preliminary work that enables not only further empirical investigation on the logical framework of forgetting operations, as presented by Beierle et al. (2019), but also cognitive costs of mental operations, like rule updating, as the current dataset contains instances of adding, modifying and deletion of information.

[3]We use `scikit-learn`'s `LinearDiscriminantAnalysis` and `LinearRegression` for which we converted the categorical variables (RuleType and Scope) using `panda`'s `get_dummies` to one less 0/1-variables than the total amount of different feature values. E.g., the RuleType feature is divided in RuleType-CountTwice and RuleType-CountAs which account for no rules as well.

Acknowledgments

This project has been funded by a grant to MR in the DFG-projects 529624975, 427257555 and 318378366.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Beierle, C., Kern-Isberner, G., Sauerwald, K., Bock, T., & Ragni, M. (2019). Towards a general framework for kinds of forgetting in common-sense belief management. *Künstliche Intelligenz*, *33*, 57-68. doi: https://doi.org/10.1007/s13218-018-0567-3

Brand, D., Dames, H., Puricelli, L., & Ragni, M. (2022). Rule-based categorization: Measuring the cognitive costs of intentional rule updating. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society* (pp. 2810–2817).

Dames, H., & Oberauer, K. (2022). Directed forgetting in working memory. *Journal of Experimental Psychology*, *151*(12).

Gollwitzer, P. M., & Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology*, *73*(1), 186–199. doi: 10.1037/0022-3514.73.1.186

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650.

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, *32*(4), 582–591.

Oettingen, G., & Gollwitzer, P. M. (2010). Strategies of setting and implementing goals: Mental contrasting and implementation intentions. In J. E. Maddux & J. P. Tangney (Eds.), *Social psychological foundations of clinical psychology* (pp. 114–135). New York: Guilford Press.

Ricker, T., Nieuwenstein, M. R., Bayliss, D. M., & Barrouillet, P. (2018). Working memory consolidation: Insights from studies on attention and working memory: An overview of working memory consolidation. *Annals of the New York Academy of Sciences*, *1424*(1), 8–18. doi: 10.1111/nyas.13633

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