Predicting Complex Problem Solving Performance in the Tailorshop Scenario

Daniel Brand (daniel.brand@metech.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Sara Todorovikj (sara.todorovikj@metech.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Abstract

Complex problem solving (CPS) is a fundamental capability of humans. It is often studied through microworlds, with the Tailorshop-scenario as a well-investigated prominent example. This paper addresses several research questions for CPS in the Tailorshop scenario: Firstly, it examines the impact of background knowledge vs. understanding underlying dynamics. Secondly, it investigates the predictability of a participants' performance, particularly when considering their assumptions about the scenario's mechanisms. Finally, it discusses the suitability of the Tailorshop as a scenario for cognitive modeling of CPS. Thereby, we discuss some of the measures that have been proposed to assess CPS performance, considering CPS from an perspective of predictive modeling. Based on our results, we conclude that effective prediction of outcomes in complex tasks necessitates uniform impact of actions throughout, facilitating comprehension of both overarching strategies and smaller adjustments crucial in real-world problem-solving domains.

Keywords: Complex Problem Solving; Causal Map; Mental Representations; Cognitive Modeling; Tailorshop

Introduction

In our everyday life, individuals regularly encounter complex systems spanning societal, economic, and environmental realms with many latent variables, requiring adept problem-solving and decision-making skills. However, traditional decision-making research often occurs in small controlled settings, raising concerns about its relevance to real-world complexities (Pitz & Sachs, 1984). To address this, complex dynamic tasks, known as dynamic decision-making (DDM), have been used to study Complex Problem Solving (CPS) behavior. DDM involves participants making decisions within dynamic environments, observed as outcomes that may or may not be affected by decisions made (Edwards, 1962). Computer simulations, called *microworlds*, provide realistic environments for studying complex problem-solving and decision-making processes. These studies challenge cognitive demands regarding goal elaboration, information search, hypothesis formation and forecasting, which ultimately rely on an individual's planning and decision making capabilities, but also creativity (Dörner & Wearing, 1995; Gonzalez, Vanyukov, & Martin, 2005; Funke, 2014). The microworld Tailorshop (e.g., Putz-Osterloh, 1981, 1983; Funke, 1988; Danner et al., 2011; Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015) is an extensively studied computer-based dynamic decision-making scenario for CPS. Participants assume the role of a tailorshop manager for 12 months, tasked with purchasing raw materials, managing production capacity, and maximizing profit by selling shirts. The environment comprises 24 variables, with 21 visible to participants and 12 directly manipulable. These variables are interconnected, with modifications to one potentially impacting others in subsequent simulated months (e.g., advertising influences customer interest, which then affects sales). Tailorshop has been utilized to explore problem-solving processes, intelligence, and professional performance among others (Danner et al., 2011). Success in Tailorshop is typically defined as a consistent increase in company value over months, with the first month excluded from scoring to enhance consistency with the 2-12 months score being a reliable predictor for success, as found by previous studies (Danner et al., 2011; Greiff et al., 2015). However, Greiff and Funke (2009) criticize the "one-item-testing" of one large, complicated scenario as a severe shortcoming of CPS research. They propose that the detection of individual differences could be facilitated by a formal framework of linear structural equation systems — the MicroDYN approach. Instead of a single, complex system, subjects engage with 8-12 items to explore, detect causal relations between variables, draw connections between them to represent their mental model, and then adjust values to achieve target outcomes.

In the light of the discussion in the current state of the art, this paper presents a rigorous analysis of the Tailorshop scenario from a predictive modeling perspective: (1) Investigating how prior knowledge and individual characteristics influences behavior and assess their worth as a predictor; (2) Search for action patterns that can serve as a base for modeling endeavours; and (3) discuss the predictability of participants' performance and the suitability of the Tailorshop scenario for predictive modeling of CPS as a whole. Thereby, the structure of the paper is as follows: the next section presents the experimental data, followed by an introduction to the causal map analysis in Section 3. Section 4 outlines initial implications drawn from our analyses and tests the relationship between causal map information, strategies, participant actions, and performance. Finally, a discussion addressing the aforementioned key issues concludes the paper.

Experiment

Participants and Materials. We conducted an onsite study in German in our lab involving 52 students at the Chemnitz

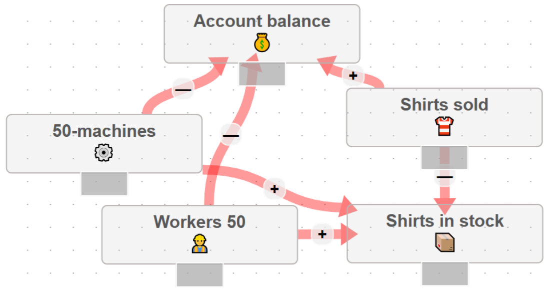


Figure 1: Illustrative example for a causal map created in the graphical user interface used by the participants to represent the relationships between Tailorshop variables (cp. Table 1).

University of Technology. Participants were compensated with either course credits or monetary rewards. The Tailorshop simulation was based on the implementation by Danner et al. (2011)[1]. Similar to the drawing of variable connections in MicroDyn (e.g., Greiff & Funke, 2009), we aimed at obtaining information about the understanding of the relationships between variables (cp. Table 1) in the scenario. Therefore, we developed a graphical interface that allowed participants to represent their understanding in the form of a causal map (Figure 1 shows an illustrative example).

Procedure. Prior to the Tailorshop experiment, participants completed the German version of the Need for Cognition questionnaire (NFC; Beißert, Köhler, Rempel, & Beierlein, 2015) and a 7-question version of the Cognitive Reflection Task (Toplak, West, & Stanovich, 2014). They were then introduced to the Tailorshop topic without explaining any of its mechanisms. Subsequently, participants were presented with variables within the causal map tool and asked to delineate connections denoting relationships between them labelling these connections as positive or negative. Afterwards, participants had an exploration phase of 6 simulated months with the Tailorshop simulator. Following the exploration phase, the scenario was reset, and participants performed a 12-month testing phase. Post testing, participants were asked to construct another causal map to assess their comprehension. Then they were asked for their specific strategies and rated variable relevance using a 5-point Likert scale. All collected data and associated scripts are publicly accessible on GitHub[2].

Analyzing the Causal Maps

For the analysis 4 participants had to be excluded, since they skipped a causal map, leading to a dataset containing the responses of 48 participants (30 female, 17 male, 1 diverse).

Causal Map Properties

Figure 2 shows the aggregated graphs from participants' causal maps both before and after engaging with the Tailorshop simulation. Only edges reported by at least 5% of participants are depicted. Additionally, this figure includes relationships derived from the Tailorshop implementation for comparative analysis. For simplicity, the graphs representing the causal maps of participants before and after interaction with the Tailorshop will be referred to as *Before* and *After* for the remainder of this paper.

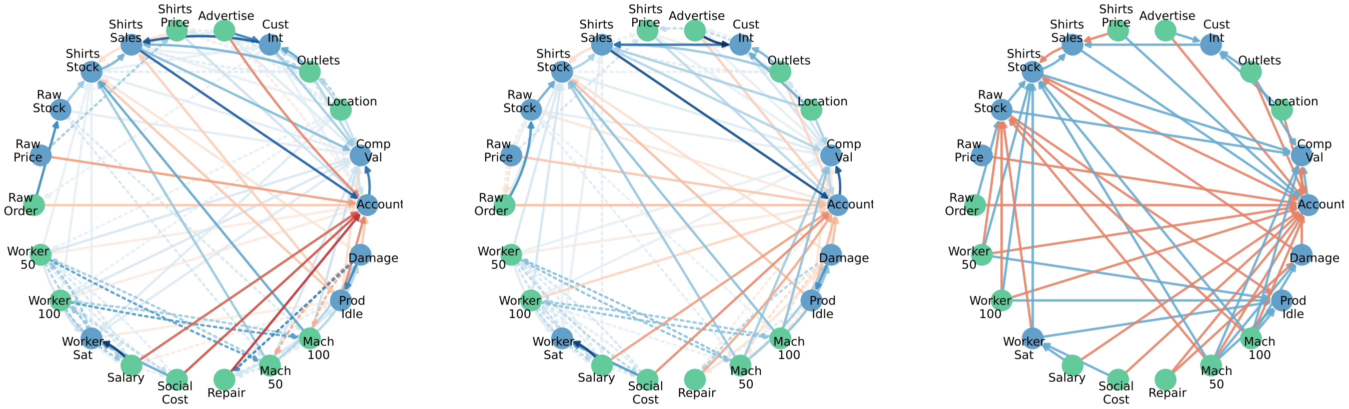
Variables controllable by participants, denoted in lightblue, are intentionally designed to be not influenced by other variables within the Tailorshop – in contrast to potential interconnections in the real world. Therefore, edges towards these variables are represented as dotted lines in the graph. This presentation form aims to highlight other edges directly comparable to the Tailorshop simulator.

The core discrepancy is between the Tailorshop graph and participants' causal maps. The simulator graph demonstrates mostly direct connections to few key variables such as account, shirt, and material stock. However, participants' causal maps exhibit higher levels of indirection and interconnection: For instance, while workers are not directly linked to costs, they are indirectly influenced by factors like salary, even when denoted on a per-person basis. Moreover, the connections in participants' causal maps also cover real-world connections that go beyond the scope of the simulation. For instance, the influence of location on worker satisfaction is identified, a *soft factor* relationship not covered by the simulator. While the first differences are most likely caused by a less formal understanding of the concepts, the latter is an expected problem of a real-world based simulation, since a simulation will automatically fall short in some aspects, which can lead to some false assumptions by the participants. The difference in interconnectivity is also visible with respect to the number of incoming and outgoing edges (see Table 2). The Tailorshop simulation has a few central nodes (e.g., the bank account) where everything comes together, while other nodes have no incoming edges at all (i.e., the variables controlled by the participants), whereas no such extremes are visible in the participants' graphs. The table also shows that the differences between *Before* and *After* are slim, indicating that no substantial structural changes occurred. Although subtle, with some adjustments to the aforementioned problems (i.e., interactions between location and worker satisfaction is no longer present) seemed to have taken place.

In order to quantify the changes between *Before* and *After*, we calculated the similarity between the participants' graphs, and the tailorshop graph. If participants adjusted their assumptions based on experiences with the tailorshop simulation, the changes between *Before* and *After* should lead to an increased similarity with the tailorshop graph. We used the average of the cosine similarities between the adjacency vectors for each node, leading to an overall similarity of .247 for *Before* and .255 for *After*. This change was not significant (Mann-Whitney-U: $U = 1113.5$, $p = 0.781$), which confirms the observation that participants overall did not revise their assumptions to a greater extend.

[1] https://www.psychologie.uni-heidelberg.de/ae/allg/tools/tailorshop/index.html

[2] https://github.com/brand-d/iccm2024-tailorshop



(a) Before playing the scenario (b) After playing the scenario (c) Tailorshop Simulation

Figure 2: Causal maps before (a) and after (b) playing the tailorshop scenario alongside the graph depicting actual dependencies in the simulation. Blue/Red edges indicate positive/negative relationships, respectively. Darker shades indicate a higher proportion of the respective edge. Green nodes denote controllable variables, while blue nodes represent derived variables. Edges to controllable variables are dotted. Edges reported by less than 5% of participants are omitted.

Table 1: Importance of a variable defined as the average number of paths leading to *Company Value* in the individual *before* and *after* graphs and in the Tailorshop implementation (TS). Average relevance (Rel) of the respective variable reported by the participants is included, variables controllable by participants are excluded.

	Var. Importance			Rel.
	Before	After	TS	
Company Value	(40.58)	(38.35)	(111)	-
Bank Account	24.19	25.31	67	-
Customer Interest	9.06	8.69	9	4.27
Shirts Sales	15.58	15.62	36	4.69
Shirts in Stock	7.19	9.77	72	3.79
Raw Material Price	0.46	1.46	0	3.56
Raw Material Stock	0.96	4.23	32	3.98
Worker Satisfaction	11.0	7.15	14	2.85
Production Idle	2.9	3.83	0	3.30
Damage	3.33	3.1	12	3.25

Table 2: Overview of the graph connectivity comparing the number of incoming and outgoing edges for the graphs from the causal map and the implementation of the tailorshop.

		Mean	SD	Min	Max
Incoming	Before	1.46	1.01	0.46	5.19
	After	1.39	1.18	0.29	5.79
	Tailorshop	2.41	3.77	0	15
Outgoing	Before	1.46	0.36	0.52	2.31
	After	1.39	0.33	0.52	1.96
	Tailorshop	2.41	1.53	0	6

The graphs obtained from the causal map can also allow for estimates of a variables importance. Since the maximization of the company value was the goal of the tailorshop scenario and participants were instructed to try to do so, we evaluate the importance of a variable with respect to company value. As an importance metric for a variable, we used the number of occurrences in all (cycle-free) paths leading to company value, excluding those starting at the respective variable. Put differently, since edges in the graph denote a positive or negative relationship between variables, the metric gives an estimate of the number of ways a variable influences the company value indirectly when another variable is changed. Table 1 shows the importance values for all derived variables (i.e., variables not directly controllable) as well as the relevance that participants provided at the end of the experiment. Note that the bank account was excluded from the relevance, since it was directly explained to be a part of the company value, rendering its relevance trivial. Unsurprisingly given the scenario, shirt sales was assigned the highest relevance, which was also reflected by the importance (15.62 for *After*). Apart from that, no clear correspondence between importance and relevance was visible. However, the importances of the participants' graphs are generally in line with the importances of the variables in the actual tailorshop simulation (Kendall's Tau between *Before* and *TS*: $\tau_b = 0.556$, $p = .029$), indicating that the general concepts are comparable. The relevance, on the other hand, seemed to be mostly focusing on directly sales-related concepts (i.e., shirt sales, and the customer interest), rating variables for production generally lower.

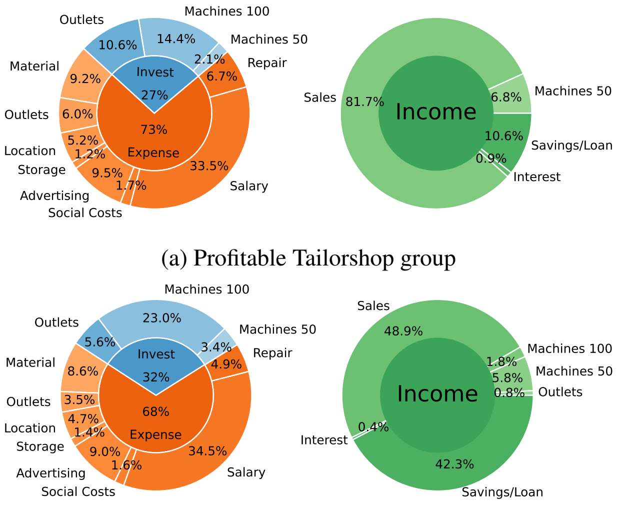
Causal Map and Performance

Since the assumptions participants have about the mechanisms underlying the tailorshop scenario are likely to influence their actions, we investigated the connection between the causal maps and the performance in the tailorshop, assuming that participants with a *Before* graph more similar to the actual tailorshop graph will achieve a better performance. To assess performance, we considered the 11th month as a reference point for the final performance, since participants can skew the results by selling everything in the last month (25% of participants stated that they considered that strategy). Unlike Danner et al. (2011), we use the total difference in company value, since participants were instructed to maximize it until the end of the run (and not consistently each month). We argue that modeling should focus on a task as closely related to the actual instructions as possible. Additionally, to normalize the values of the Tailorshop, we represented the performance as a proportion of the company value change (i.e., by calculating $perf = (cv_{11} - cv_0)/cv_0$, where cv_{11} is the company value at the end of month 11 and cv_0 the initial company value). Subsequently, we proceeded by splitting the participants into two groups based on the median difference in company value between the beginning and the last month. Here, the differences are more apparent: The high performing group had an average similarity of .279 between *Before* and the tailorshop graph, which increased to .319 for *After*. In comparison, the low performing group started with a similarity of .215, which decreased to .191. This indicates that participants that already started out in line with the tailorshops mechanisms were able to further adjust their assumptions, while the low performing group seemed to struggle to grasp the mechanisms. Based on these findings, we aimed to predict the performance in two ways: 1) We used a Support Vector Regression (SVR; for an overview, see Awad & Khanna, 2015) as a simple general-purpose model to predict the performance in the tailorshop for all individual participants based on the *Before* graph, and 2) used the similarity directly as an estimate for the tailorshop performance. First, the SVR was trained and tested using a leave-one-out cross-validation, to ensure that the limited number of participants for a machine learning method is used efficiently. The adjacency matrix of the *Before* graph was used as the input, while the performance value described beforehand was used as the target. The Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2) were used to measure the performance. The median and mean of the target values were added as baseline models, since they represent the optimal constants to minimize MAE and RMSE, respectively.

The results are shown in Table 3. The results show that the SVR was not able to leverage any of the information available in the graph, achieving a similar performance than the mean and median. Including additional individual information (CRT and NFC) did not improve the performance. Two possible explanations for that hinge on fundamental attributes of the present data: First, the coarse structure of the present causal maps only reflect relationships, but do not capture the meaning or importance of certain connections. Second, the tailorshop simulation is a complex, non-linear scenario, that can provide greatly differing experiences even for participants with a rather similar overall behavior.

Table 3: Results of a leave-one-out cross-validation analysis for predicting the tailorshop performance. The table shows the MAE, RMSE and R^2 for the Support Vector Regression (SVR) based on the causal map graph provided before the tailorshop, the SVR based on actions in the first month and the mean and median target value as baseline predictors.

Predictor	MAE	RMSE	R^2
Performance Mean	0.298	0.378	0
Performance Median	0.295	0.381	-0.018
SVR (*Before* graph)	0.293	0.379	-0.007
SVR (First Month Actions)	0.255	0.328	0.247



(a) Profitable Tailorshop group

(b) Unprofitable Tailorshop group

Figure 3: Comparison of the income, expense, and investment proportions of the two participant groups.

However, even when direct predictions based on the causal map graphs were not possible, the similarity to the tailorshop graph can still serve as a predictor of performance in terms of correlation: If participants started with a graph more similar to the actual simulation, they should be able to make better informed decisions, thereby increasing their performance. The one-sided Spearman correlation between the similarity of the *Before* graph to the tailorshop graph showed a significant moderate correlation ($r = .264$; $p = .035$). Still, it does not seem to provide enough information for the models for individual predictions, but can be a useful utility metric. Similarily, the CRT showed a significant correlation with performance, while the NFC did not (One-sided Spearman rank correlation: CRT: $r = .245$, $p = .047$; NFC: .062, $p = .337$).

Analyzing Strategies and Actions

Since the tailorshop scenario is a dynamic simulation, participants experienced different situations depending on their initial decisions, making it hard to be captured and predicted by the limited causal map information. Therefore, we will now turn to the analysis of actions and strategies that participants used with respect to the resulting performance.

General Properties

Overall, 31.25% of the participants ended in debt, while only 10 participants (20.83%) had profitable tailorshops. For the following analyses, we focus on the difference between the profitable and all unprofitable tailorshops, not covering the differences to the subgroup of unprofitable tailorshops that ended up in debt specifically. First, the two groups are compared in terms of their expense, investment and income strategies. A breakdown thereof is shown in Figure 3. While the difference for the income is mostly due to the necessity of taking a loan or using up the savings, it becomes apparent that the profitable group very rarely relied on their savings over the course of the simulation. When considering the expenses and investments, the only major difference appears to be the investment in machines, which takes up a substantially larger proportion of the investments for the unprofitable group, and was invested in additional sales outlets instead by the profitable group. Overall, investment and expenses are rather similar, hinting at a problem of finding the right point in time: While comparable over the course of the run, the profitable group seems to make better decisions from the beginning (as indicated by the low proportion of used savings).

To gain a deeper understanding of the mechanisms causing the differences, we investigated the behavior of both groups on the level of actions and respective effects on the derived variables. Additionally, we included the exploration phase into the investigation, in order to see if participants with a better performance used the exploration phase to learn a strategy or had a better approach right from the beginning. Figure 4 shows the actions performed by both groups in each month as well as the resulting changes to the observable derived variables. From this, several findings are noteworthy: First, both groups had a negative outcome in the exploration phase, but used the phase by performing more drastic changes compared to the test phase, which was possibly the cause for the worse overall performance. Second, the first month showed by far the biggest changes and seemed to contain all the initial investments and adjustments that were planned to set the tone for the remaining months. Especially in the test phase, no substantial adjustments to the extend of the first month are made to any variable afterwards (besides selling everything right at the end to boost the final results). Third, the actions performed in the first month resembled the behavior already present in the exploration phase, with minor adjustments. This is corroborated by a significant strong correlation between the performance in the exploration phase with the performance in the test phase (One-sided Spearman's rank correlation: $r = .879$, $p < .001$).

Overall, a few clear but subtle differences between both groups emerged: For one, both groups switch to the machines with more capacity, but the profitable group sells the old machines more decisive. For another, the profitable group seemed to avoid running in a supply shortage by investing more in raw material, machine maintenance, new machines and workers as well as salary compared to the unprofitable group. Finally, the unprofitable group starts with expanding outlets right away, which the profitable group is more hesitant to do. After the first month, the actions reflect the general situation: While the profitable group performs minor adjustments to advertisement and shirt price, the unprofitable group is forced to make cuts. While this is important for investigating the participants' ability to perform small-scale adjustments, it is mostly a product by the decisions that were made in the first month.

Predicting Performance

To predict the performance based on the actions, we rely on the same methods as in the causal map analysis. Again, we use the SVR, this time using the actions performed in the first month as inputs. The results (see Table 3) show that the SVR is now able to outperform ($MAE = 0.255$, $RMSE = 0.328$) the baseline models ($MAE = 0.295$ for the median baseline and $RMSE = 0.378$ for the mean baseline, respectively). Furthermore, it now achieves a positive coefficient of determination ($R^2 = 0.247$), indicating that, even for a simple general model, the first month provides easily accessible information.

Similarly to the correlation between causal map similarity and performance, we developed metrics aiming to correlate well with performance based on the actions. We used two simple heuristic strategies as metrics:

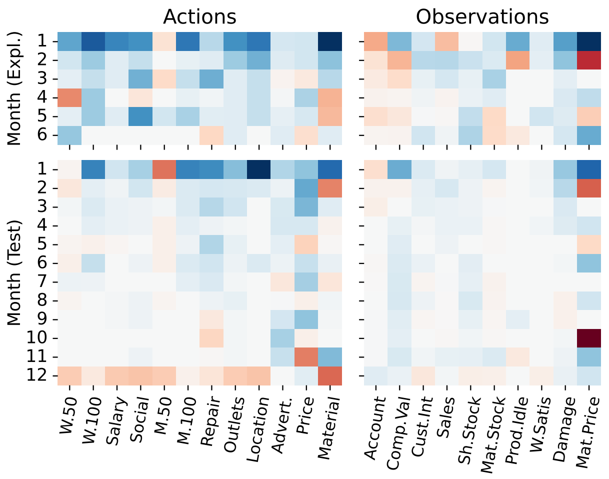
1. *Upgrade machines* (buy better machines, hire the respective workers, and sell the old machines), was calculated as follows:
   $strategy1 = sign(\Delta M100 + \Delta W100) * sign(-\Delta M50)$, where *sign* is the signum function and $\Delta M100$, $\Delta W100$ and $\Delta M50$ are the changes of the number of machines and workers.

2. *Avoid production loss* (buy raw material and invest in repair/maintenance), calculated as follows:
   $strategy2 = Material + Repair$

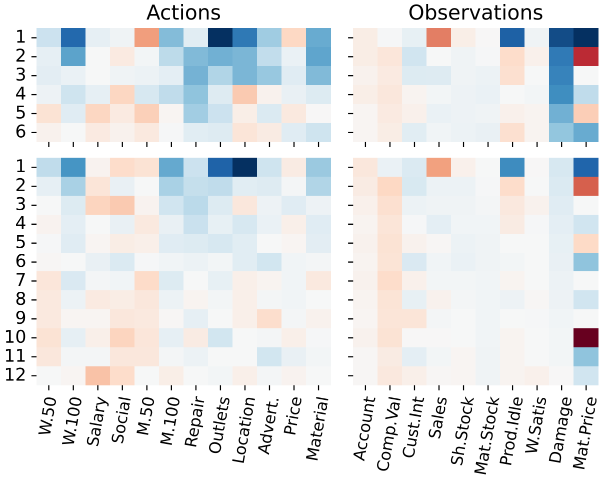
Both of the metrics correlated significantly with performance (Spearman's rank correlation for S1: $r = .310$, $p = .016$; for S2: $r = .656$, $p < .001$), indicating that a few actions in the first month are already good predictors for the performance. The present results suggest another explanation for the limited success of using causal maps: Due to the task's reliance on initial actions, many assumptions about variable interplay become irrelevant, whereas later decisions hard to predict due to the self-reinforcing and complex nature of the scenario.

Discussion

In the present article, we assessed three issues: First, we assessed the importance of the knowledge and assumptions that



(a) Profitable



(b) Unprofitable

Figure 4: Average actions (changes to controllable variables) and the resulting observed variables for profitable/unprofitable tailorshops during the exploration and test phase. Darker shades of blue/red denote higher increments/decrements, respectively.

participants have before interacting with the tailorshop simulation are for the performance and if a revision of the assumptions is observable. For that, we obtained a causal map representing the relationships between the variables of the tailorshop before and after participants were interacting with the simulation. Thereby, the causal maps showed no significant signs that the initial assumptions were updated and could, despite correlating moderately, not be used as predictors of tailorshop performance when modeled using support vector regression. Similarly, the results of the Cognitive Reflection Task and the participants' Need for Cognition had no significant influence on the model. Finally, we found that the performance during the exploration phase was strongly correlated with the performance in the test phase, which further supports that no substantial changes to the assumptions were made. While it was expected that a real-world inspired scenario would be substantially impacted by real-world knowledge, part of the results could be explained by a limitation of the causal maps: The restriction to only represent positive or negative dependencies is too coarse to describe the dependencies that participants actually expect, introducing noise due to ambiguity and the lack of expressiveness.

Second, the actions selected by the participants were investigated. The analysis showed that the first month was by far the most dominant month, setting the tone for the whole run. This is likely to cause most other factors to become irrelevant, especially since the scenario itself is highly dynamic. Participants that made less fortunate actions in the first month rarely recovered, which in turn can likely alter their strategies. Although the first month is often excluded (Danner et al., 2011; Greiff et al., 2015), which is a reasonable means if the focus lies on the micromanagement feedback-loop during the other months, we argue that this is not ideal for cognitive modeling of complex problem solving. On the one hand, participants were instructed that the scenario has a time limit of 12 months, where only the company value at the end mattered. This implies that each intermediate steps on its own does not necessarily reflect the actual thought processes. The fact that 25% of participants considered selling everything at the end to boost the final result further corroborates that they had, in fact, an overarching strategy. Since excluding the most impactful month from the performance evaluation strips the tailorshop almost entirely of its investment phase, in which planning and the strategies of participants arguably matter the most. Even when excluded, the first month will still alter the whole scenario and thereby the behavior of the participants, making it near impossible to predict. When predicting based on the initial actions, the support vector regression performed substantially better and outperformed the baselines. Furthermore, we were able to formulate simple strategies based on the first month alone that can serve as highly correlating predictors for success in the tailorshop. For predictive modeling endeavours, this leaves the tailorshop scenario in a tricky state, since most of the planning and adaption processes will be hidden by the dominant initial decisions, which can set the tone for the complete run in a complex non-linear scenario.

In conclusion, while we deem the use of complex problem solving in cognitive modeling important to extend its boundaries further into the area of real-world scenarios, we argue that the tailorshop gets trapped in its complexity, which makes it prone for snowball effects based on early actions. To this end, our results align with the critique by Greiff and Funke (2009) on "one-item-testing". Especially for cognitive modeling, it is essential to rely on a complex tasks that is either easily repeatable (i.e., by having multiple items), or is less self-reinforcing, so that the actions performed by participants across all steps of the tasks have a similar impact. In the end, overarching strategies have to be observed at the same time as small step-to-step adjustments — since both are essential components of real-world complex problem solving.

Acknowledgements

This project has been partially funded by a grant to MR in the DFG-projects 529624975 and 283135041 and by Saxony State Ministry of Science and Art (SMWK3-7304/35/3-2021/4819) research initiative "Instant Teaming between Humans and Production Systems".

We extend gratitude to Prof. Dr. Joachim Funke and Dr. Daniel Holt for providing us with the original Tailorshop implementation.

References

Awad, M., & Khanna, R. (2015). Support vector regression. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 67–80). Berkeley, CA: Apress. doi: 10.1007/978-1-4302-5990-9_4

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2015). Deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition NFC-K [German short scale for measuring the construct Need for Cognition NFC-K]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*.

Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., & Funke, J. (2011). Measuring performance in dynamic decision making. *Journal of individual differences*, *32*, 225-233. doi: 10.1027/1614-0001/a000055

Dörner, D., & Wearing, A. J. (1995). Complex problem solving: Toward a (computersimulated) theory. In *Complex problem solving* (pp. 65–99). Psychology Press.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processings. *Human factors*, *4*(2), 59–74.

Funke, J. (1988). Using simulation to study complex problem solving: A review of studies in the FRG. *Simulation & Games*, *19*(3), 277–303.

Funke, J. (2014). Problem solving: what are the important questions? In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 493–498).

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in human behavior*, *21*(2), 273–286.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving: The microdyn approach. Office for Official Publications of the European Communities.

Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence*, *50*, 100-113. doi: https://doi.org/10.1016/j.intell.2015.02.007

Pitz, G. F., & Sachs, N. J. (1984). Judgment and decision: Theory and application. *Annual review of psychology*, *35*(1), 139–164.

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [The relation between test intelligence and problem solving success]. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*, *189*(1), 79-100.

Putz-Osterloh, W. (1983). Über Determinanten komplexer Problemlöseleistungen und Möglichkeiten zu ihrer Erfassung [On factors for complex problem solving and possibilities of their diagnosis]. *Sprache & Kognition*, *2*, 100–116.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147-168. doi: 10.1080/13546783.2013.844729