Journal of Artificial General Intelligence 2(2) 38-51, 2010 Submitted 2009-12-15; Revised 2010-06-10
10.2478/v10229-011-0008-2

Keep it simple – A case study of model development in the context of the Dynamic Stocks and Flows (DSF) task

Marc Halbrügge HALBRUEGGE@GMAIL.COM

Editor: Christian Lebiere, Cleotilde Gonzalez, and Walter Warwick

Abstract

This paper describes the creation of a cognitive model submitted to the 'Dynamic Stocks and Flows' (DSF) modeling challenge. This challenge aims at comparing computational cognitive models for human behavior during an open ended control task. Participants in the modeling competition were provided with a simulation environment and training data for benchmarking their models while the actual specification of the competition task was withheld. To meet this challenge, the cognitive model described here was designed and optimized for generalizability. Only two simple assumptions about human problem solving were used to explain the empirical findings of the training data. In-depth analysis of the data set prior to the development of the model led to the dismissal of correlations or other parametric statistics as goodness-of-fit indicators. A new statistical measurement based on rank orders and sequence matching techniques is being proposed instead. This measurement, when being applied to the human sample, also identifies clusters of subjects that use different strategies for the task. The acceptability of the fits achieved by the model is verified using permutation tests.

Keywords: Cognitive Modeling, Model Comparison, Sequence Matching, Cluster Analysis, Bootstrap Test

1. Introduction

Cognitive modeling allows the examination of psychological theory by means of empirical comparison between human subjects on the one hand and computer models that follow the theory very closely on the other hand. An emerging field in cognitive modeling is model comparison. Competing theories for a certain psychological finding are analyzed by comparing the goodness-of-fit of their computational counterparts to behavioral data.

A new model comparison effort is the Dynamic Stocks and Flows (DSF) modeling comparison challenge as held prior to the 9th International Conference on Cognitive Modeling (Lebiere, Gonzalez, and Warwick, 2009). Dynamic Stocks and Flows is a simulation of a simple dynamic system used for research in human decision making in every-day control tasks. The system consists of a single stock, i.e. water level in a tank, which is changing over time due to inflow and outflow of water from and to the environment (EI and EO). The task of the subject is to maintain the water level in the tank at a given target value by setting additional inflow and outflow rates. These rates are applied in the subsequent time step. The complexity of the task arises from dynamic changes in the environmental flows. During a session of 100 time steps, the subject has to build up a model of the underlying rules in order to predict the flows in the following time step.

In the DSF modeling comparison challenge, participants were asked to model four experimental conditions of the DSF task. These 'training' conditions are all characterized by the absence of environmental outflow ($EO = 0$) and varying environmental inflow. Depending on the condition, the inflow increases or decreases over time, and does so either linearly or non-linearly.[1] The submissions to the challenge were tested using data from five ‘transfer’ conditions that were unknown to the participants when they developed their models. Three of these are based on a short inflow sequence that is being repeated over and over. The remaining two conditions use the same inflow function as the ‘linear increase’ training condition, but delay the application of the user action by two or three time steps.

Taking part in competitions like the DSF challenge is fun. But do they contribute to science? According to Stone (2003), “research inspiration” and the creation of “common platform for exchanging ideas” belong to their main benefits. In the the context of the RoboCup robotic soccer championships, this has been proven to create not only qualitative, but quantitativly measurable progress (Gabel and Riedmiller, 2010). In the light of these findings, and trying to honor the goal of the DSF challenge (i.e. to “advance the state of the art in cognitive modeling”), the main focus of this paper is not the cognitive model itself, but the modeling methodology and statistical approaches that were used during the design of the model. The next section presents robust statistical methods for the assessment of model fit using sequence matching and bootstrapping techniques. These methods were applied during the development of the submission to the DSF challenge, which is described subsequently. The forth section is about the results of the model in the transfer conditions, followed by conclusions and general remarks.

2. Statistical methods for model evaluation

Modelers tend to use basic statistical measures like means, standard deviations and correlations only. The human data provided by the organizers of the challenge shows the pitfalls of this approach quite well.

The first and most general problem is the sequential nature of the task. In the training conditions, the human samples consisted of less than twenty subjects who completed 100 trials each. During the first ten to twenty trials, the subjects had to learn the task. This led to high variability in the data. After the learning phase, especially the relatively easy linear increase and decrease conditions were solved almost perfectly, leading to very small differences between subjects (see Figure 2).

Even the second half of the experiment, i.e. long after completion of the noisy initial learning phase, yielded extreme outliers, such as subject t33 who drove the water level far above 500 gallons (instead of the target value of 4 gallons) for about 50 of the 100 trials. Looking at the logs suggests that after having become acquainted with the task, some subjects could not maintain their motivation and followed the temptation of putting in large numbers to observe what would happen.

Given this data quality, computing averages and correlating the results with some model output, or summing squared residuals for the RMSE would lead to a goodness-of-fit index of questionable value.

2.1 Descriptive statistics and graphs

For this reason, only non-parametrical statistics were used during the development of the model presented in this paper. Graphical checks were done using a ‘boxplot over time’. This plot displays the three quartiles of the behavioral data at every time step as lines and fills the area between the

[1] timestep t in [1;100]. linear increase: $EI(t) = 2 + .08t$; linear decrease: $EI(t) = 10 - .08t$; non-linear increase: $EI(t) = 5 \lg t$; non-linear decrease: $EI(t) = 5 \lg(101 - t)$

first and third quartile with solid color, just like the box in an ordinary boxplot. The overlay of results of the cognitive model allows for a fast visual check of how the model performs compared to humans, especially whether it falls within the middle half of the subjects. Examples are provided in Sections 3 and 4.

2.2 Distance measures for sequential data

During the first stages of the development of a model, graphical checks are often of more value than numerical goodness-of-fit indices. In the finalization phase, numerical indices become predominant, as they provide stronger arguments that do not suffer from the inherent fuzziness of graphical methods. What kind of goodness-of-fit measure suits the DSF data best?

When the object of interest is not a single measure, but a sequence of measures, the statistics taught in psychology courses do not offer any suitable method for dealing with such data. In some artificial intelligence domains, namely speech recognition, sequences are compared using dynamic programming techniques (Bertsekas, 2000), but no inference testing is done. An approach that uses the AI methods for the comparison of behavioral data has been described in a previous paper by this author (Halbrügge, 2007a).

In the context of the DSF challenge, the most complex subtask of sequence matching – finding the sequence elements that correspond to each other – is not necessary. All sequences have a fixed length of $n = 100$ steps, and these steps have been generated by identical environmental situations. In this case, the sum of absolute differences is sufficient as a simple distance measure. Let a and b be the sequences to be compared (in the case of the DSF task: a_t is the amount of water in the tank at time step t), then the distance d between a and b can be computed using the city block distance (Bortz, 1999, p. 552)

$$d(a,b) = \sum_{t=1}^{n} |a_t - b_t| \tag{1}$$

Of course, this is not the only possible metric. Depending on the emphasis that should be laid on relatively large differences at a given timestep, other metrics could be more suitable. During the analysis of the human data provided by the organizers of the DSF challenge, metrics of the form $(\sum |a_t - b_t|^r)^{1/r}$ with $r = 1$ (city block, as given in Eq. 1) and $r = 2$ (euclidian) were tried. The application of a hierarchical cluster analysis on the data yielded the best results with regards to interpretability with $r = 1$, which can be explained by the presence of outliers in the data. For higher values of r, the influence of outliers increases.

As Eq. 1 allows the computation of differences between all objects from the data sample, the prerequisites for the application of hierarchical clustering techniques (Mirkin, 1996) and group comparisons using permutation tests (Lehmann and Romano, 2005) are met.

2.3 Significance testing

Modelers often face the problem of reduced variance: although they can easily generate hundreds of model runs, there is only little variation. In other words, the model runs cannot be viewed as independent observations. Therefore standard techniques like t-Test or ANOVA cannot be used. For this reason, model runs will be treated as a single observation on the subsequent pages.

Before constructing a significance test based on d from Eq. 1, we must answer the question of which properties of d should be used for such a test. The noisiness of the data suggests abandoning parametrical statistical models in favor of methods that use only ranking information. Which ranks can be used?

The rank order of two $d(a, b)$ and $d(a, c)$ seems meaningful (a, b, c form a triangle). Whether the comparison between $d(a, b)$ and $d(e, f)$ yields any usable result is questionable. Therefore, a test statistic that only uses ranks of the former type has been constructed.

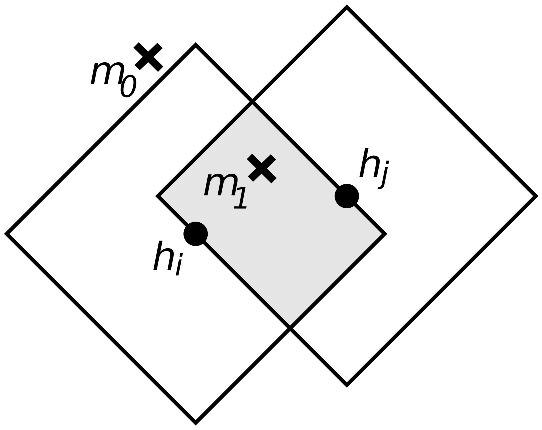


Figure 1: Visualization of Eq. 2: While m_0 is too far away from h_i and h_j, m_1 would be considered acceptable. The squares around h_i and h_j are 'circles' with radius $d(h_i, h_j)$ using the city block metric.

In the following, m denotes the sequence generated by the model, h_i the sequence generated by the human subject i. In order to assess the fit of m, the following steps are taken: For every pair i, j with $i \neq j$ of observations (i.e., subjects) compute whether the distance between m and the humans is higher or lower than the distance between h_i and h_j. If

$$\max(d(m, h_i), d(m, h_j)) < d(h_i, h_j) \tag{2}$$

then m has an acceptable distance to h_i and h_j. The rationale is visualized in Figure 1. Model m_0 lies outside the grey area for which Eq. 2 holds, therefore it is unacceptably off mark. Model m_1, however, would be considered acceptable.

The overall acceptability of a model sequence m can be computed by testing Eq. 2 for every pair i, j and counting the relative frequency of acceptable combinations:

$$\alpha(m) = P((i, j); \max(d(m, h_i), d(m, h_j)) < d(h_i, h_j)) \tag{3}$$

Although being an interesting value in itself when interpreted as a probability, $\alpha(m)$ should not be confused with the level of significance of a statistical test. It is rather a descriptive statistic for the position of the model sequence distances in the space of observed human sequence distances.

Confidence intervals for $\alpha(m)$ can be obtained by using bootstrap methods. One sequence from the human data is chosen to stand in for the model sequence, the control data is randomly sampled with replacement from the remaining human sequences (Davison and Hinkley, 1997, chapter 4.4).

3. The DSF model

The aim of the submission to the DSF challenge presented in this paper was to create the simplest model that is nondistinguishable from a human subject using the statistical methods described in the previous section. Of the behavior displayed by human subjects, only the most relevant aspects were modeled. That left many interesting topics, for example visual search, reaction times, or mental math unaddressed. Special emphasis was placed on the learning curves of human subjects.

3.1 Theoretical background

The model is based on two simple assumptions, namely that (1) humans cannot cope well with non-linear growth, and (2) the human concept of space does not allow an object to contain a second object that is bigger than the first one.

The first assumption is based on Dörner's (1989) work on human performance in complex situations. Dörner's analysis suggests that humans can well anticipate linear relationships, but have difficulty with non-linear ones. Wagenaar and Sagaria (1975) have described human shortcomings in the anticipation of exponential growth before. In the context of the four training conditions of the DSF task, these findings explain the poor performance in the 'non-linear' conditions when compared to the 'linear' conditions.

The second assumption is needed to explain the performance drop in the 'decrease' in contrast to the 'increase' conditions of the DSF task. As shown in Figure 2, the human subjects needed much much more time and displayed much more variance in the 'decrease' case.

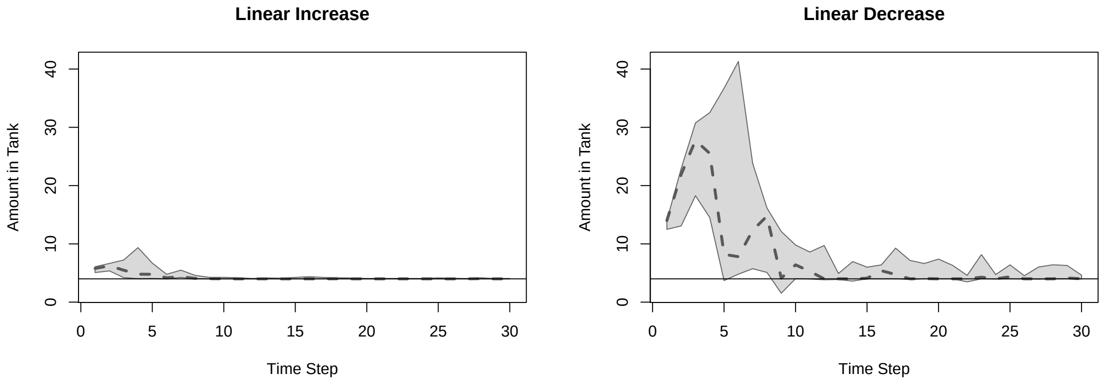


Figure 2: Human performance in the 'Linear Increase' versus 'Linear Decrease' conditions. Dashed line: median of the human sample. Grey area: second and third quartile of the human data. Horizontal line at 4 gallons: Target value.

In the case of the 'linear' conditions, the environmental flow functions differ only in a constant and the sign of a multiplier: The environmental inflow is $EI(t) = 2 + 0.08t$ in the 'increase' and $EI(t) = 10 - 0.08t$ in the 'decrease' condition. Despite this small difference, the human subjects needed more than twice the number of time steps in the 'decrease' condition before most of them found a strategy to cope with the situation.

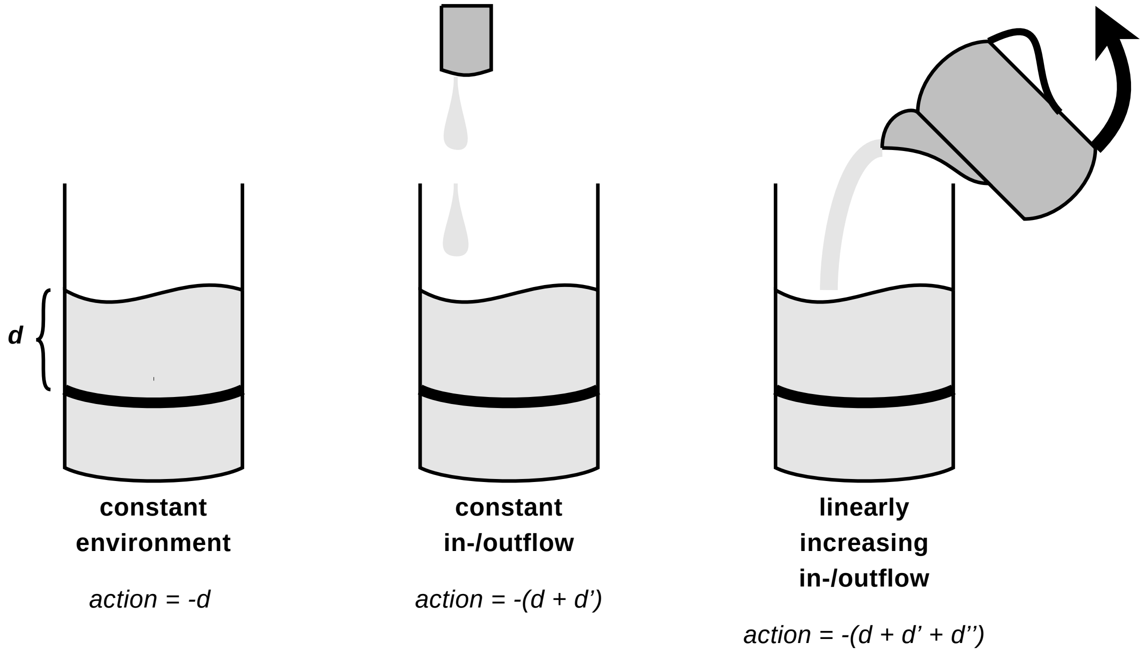
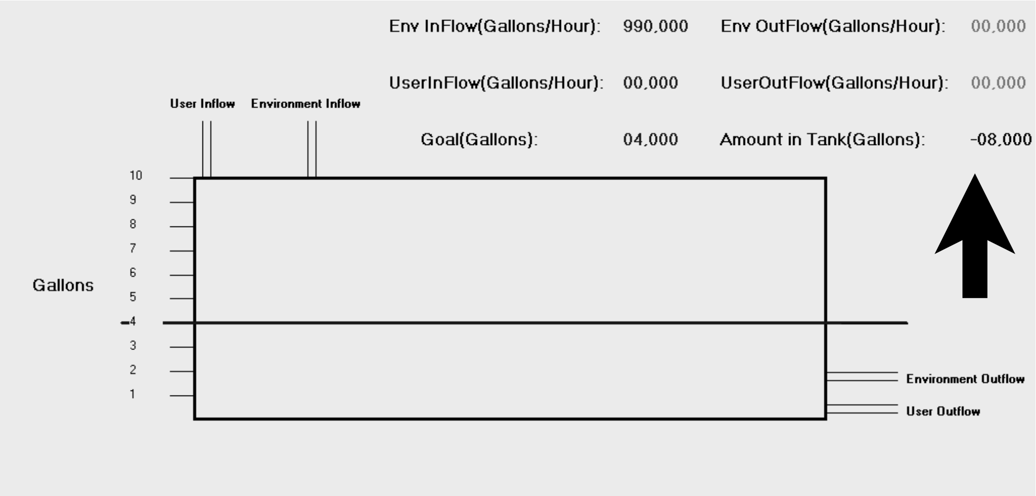


Figure 3: Postulated mental models of environmental flow with increasing complexity from left to right

Dutt and Gonzalez (2007) suggested that the bad performance in the 'decrease' condition is caused by the negative slope of $EI(t)$, i.e. the sign difference in the two functions. In contrast to this, the theory described here emphasises the different constants 2 and 10. The rationale goes as follows: The subjects use mental models acquired during their daily lives in order to solve the DSF task (Figure 3). Such models are for example a glass of water or a cup of tea (left model in Figure 3). A model of pouring tea into a cup (right model) matches the 'linear increase' condition of the DSF task very well.

But in the 'linear decrease' condition, the DSF task displays a property not present in daily life: It can be more than full and even less than empty. The graphical representation of the tank indicates that the tank ceiling, where the inflow pipes are mounted, is at ten gallons. But during the course of the DSF simulation, the amount of water in the tank can reach any value above ten and below zero. In addition to this violation of the human concept of space, the interface enters an inconsistent state in these situations. As shown in Figure 4, the display of the originally blue water area on the screen and the numerical display contradict each other, making the task even harder.

Along these lines, the bad performance in the decrease conditions is explained by the strong inflow (nearly ten gallons) at the early time steps. With an initial amount of two gallons in the tank, the early inflow forces the water level outside the given size of the tank. Why is this more difficult than the 'increase' condition? It is not solvable with intuitive thinking. No human can create a mental image of a glass that is more than full. No one has ever perceived such a glass. One needs to abstract from the every day concept of 'glass of water' to accept it to hold any amount of liquid.



Env InFlow(Gallons/Hour): 990.000 Env OutFlow(Gallons/Hour): 00.000

UserInFlow(Gallons/Hour): 00.000 UserOutFlow(Gallons/Hour): 00.000

Goal(Gallons): 04.000 Amount in Tank(Gallons): -08.000

Figure 4: Inconsistent user interface of the DSF task: While the graphical display in the center shows an empty tank, the ‘Amount in Tank’ readout displays a negative value. *(Arrow added by the author. Some parts of the original display have been removed for reasons of simplicity.)*

3.2 Implementation details

The model has been implemented using the ACT-R framework (Anderson et al., 2004). The following section is intended to give readers familiar with the architecture a detailed insight into the model that would allow a trained modeler to recreate the model from the text. Readers without previous knowledge in ACT-R might want to skip to the results section. The source code of the model can be downloaded from the author’s website.[2]

Following the first assumption from the previous section, three mental models of water tanks as displayed in Figure 3 have been implemented, one for constant environments (‘glass of water’) and one for constantly changing environments (‘steady drops of water into a glass’). The most complex model is an environment with linearly increasing or decreasing change (‘pouring tea into a cup’). Any of these three environment types are implemented as a ‘decision’ production in ACT-R that makes the cognitive model manipulate the water tank in the optimal way for this environmental flow situation. A mathematical representation of these optimal manipulations is given in Figure 3 in the bottom lines.

Two versions exist for any of the three ‘decision’ productions, one using the practical (or intuitive[3]) rule that a glass cannot be more than full or less than empty, and another adding more analytical reasoning that the amount of water in the tank can reach any value. The same model behavior would be obtained by introducing two modes of processing the DSF graphical interface,

[2] http://www.marc-halbruegge.de/act-r/

[3] The Kantian epistomological concept of ‘Anschauung’ provided some background to this rule. As the usual english translation of the term (i.e., intuition) lacks the notion of ‘knowledge being obtained through the senses’ in ‘Anschauung’, ‘practical’ was chosen as name for the strategy instead of ‘intuitive’.

one reading the amount of water from the height of the blue area only, while the other uses the numerical display on the right top only. But following assumption 2 from the previous section, the difference between the practical and analytical strategy occurs on a higher cognitive level, justifying the implementation with ACT-R's procedural system.

ACT-R's learning facility provides the means to try different strategies and stick with the one that is working best. The amount of reward that drives the switching between the six 'decision' productions is given in Table 1.

Name	d_t	d_{t-1}	Reward
perfect match	$= 0.0$		10
very good match	< 0.05		7
good match	< 1.0		3
bad match	≥ 1.0		0
made progress	≥ 1.0	$> d_t$	1
no progress	≥ 1.0	$< d_t$	-5
regression	≥ 1.0	< 1.0	-10

Table 1: Rewards for ACT-R's online learning facility. d_t is the absolute difference between the target and the current amount of water in the tank in gallons at timestep t.

The initial utilities of the decision productions in combination with the rewards determine how long the model sticks with an improper strategy until it chooses another one. The initial utilities have been set using the following heuristics: Rules are tried from simple to complex. Intuitive rules are tried before analytical ones. This is justified by the additional abstraction needed for the analytical rules.

4. Results

The goodness-of-fit in the four training conditions are displayed in Table 2. All p are very high, denoting that the model output did not differ significantly from the human subjects. This finding is not very surprising, as the same dataset was used during the development of this model.

4.1 Transfer conditions

As shown in Table 2, the goodness-of-fit is considerably lower in the transfer conditions, but still on an overall acceptable level. The three sequence conditions were not solved as well as in the delay conditions, which is consistent with a model that does not have any sense of time or history.

The most obvious deviation from the human sample occurs in the sequence-4 condition. In order to investigate this, a cluster analysis using the distance measure from Eq. 1 has been performed on the human data, yielding two clusters. Figure 5 shows the first 50 trials for each cluster with the model output overlaid. As can be seen in the 'boxplot over time', Cluster 0 is made up of subjects who do not seem to notice the periodicity in the environmental flow and therefore keep showing a recurring pattern of four water levels, never reaching the target value of six gallons. The

Condition	$\alpha(m)$	p
Training		
Linear increase	.827	.982
Linear decrease	.663	.815
Non-linear increase	.831	.988
Non-linear decrease	.816	.974
Transfer		
Sequence length 2	.213	.161
Sequence length 2 + noise	.425	.329
Sequence length 4	.231	.089
Feedback delay 2	.752	.843
Feedback delay 3	.465	.435

Table 2: Bootstrap test results in the training and transfer conditions. $\alpha(m)$ is defined in Eq. 3. p was obtained using a bootstrap procedure with 5000 runs as explained in Section 2.3.

cognitive model describes this group very well, a bootstrap test produced $\alpha(m) = .442$, which is not significant at $p = .536$.

For the bigger cluster 1 however, the median of the human data gradually converges to the target amount. The model prediction differs significantly from this group of people who managed to solve the sequence-4 condition ($\alpha(m) = .005, p < .001$).

In the delay conditions, the performance of the model is comparable to the 'average' human subject. Interestingly, a cyclical variation present in the human data is predicted by the model (see Figure 6). The Gluck et al. (2010, this issue) paper contains further analyses of this result.

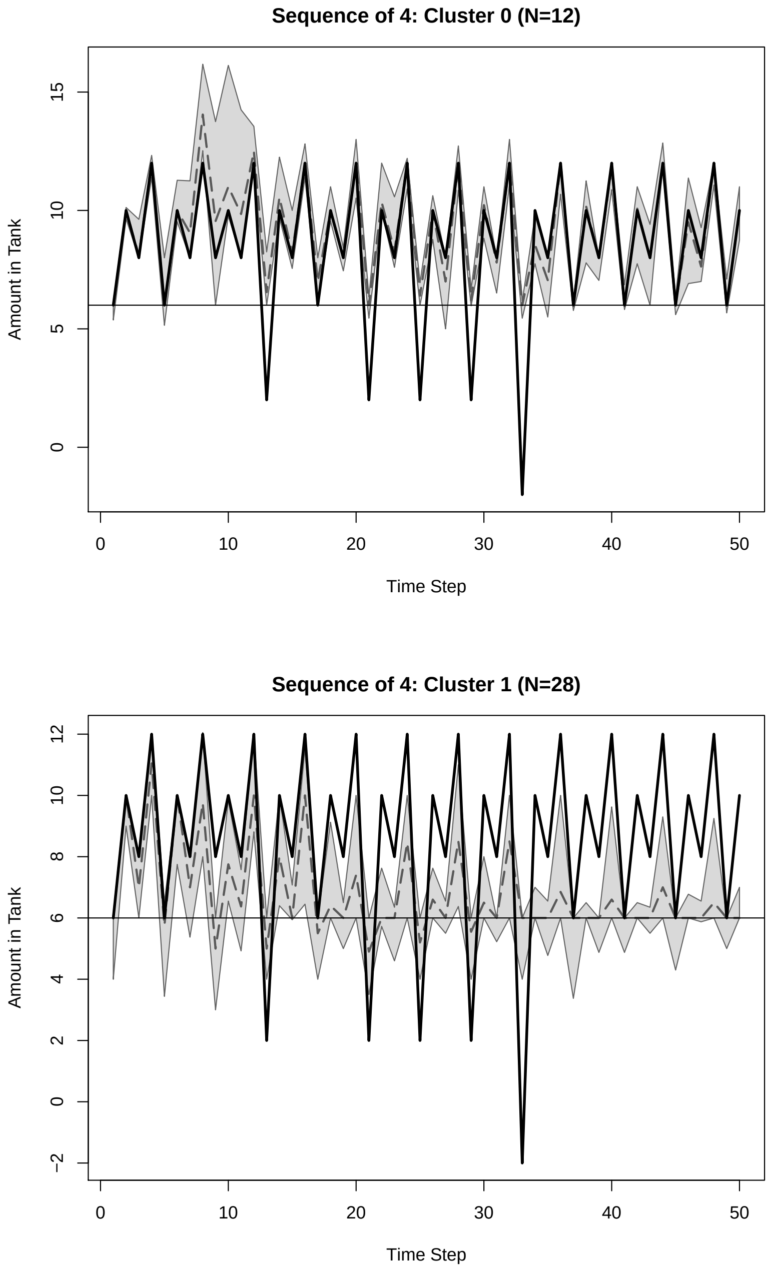
5. Discussion

From an empirical point of view, the results of the model are satisfying. The permutation tests do not show significant differences between the model and the human sample (Table 2). The qualitative difference between the 'linear increase' and 'decrease' conditions as discussed in Section 3.1 is present in the model output as well.

5.1 Goodness-of-fit

As always, statistical methods give hints, but cannot guarantee theoretical soundness. In the case of the DSF competition, another problem arises from the small size of the human sample. In simple words, the statistical approach used in this paper compares the number of anomalies of the model to the number of anomalies in the human data. As the sample size per condition was (at least in the training conditions) less than 20, every single person made up more than five percent of the empirical distribution used for the bootstrap test.

The presence of only one unfit person in the human sample would mean that being unfitting is not exceptional, allowing a model with an arbitrarily bad fit to pass the statistical test with a



Sequence of 4: Cluster 0 (N=12)

Sequence of 4: Cluster 1 (N=28)

Figure 5: Sequence of 4: Only two thirds of the human subjects realized the periodical nature of the task (Cluster 1). The model displays good fit to the subgroup of subjects who did not realize it (Cluster 0). Dashed line: median of the human sample. Grey area: second and third quartile of the human data. Bold line: model output. The horizontal line at 6 gallons displays the target value

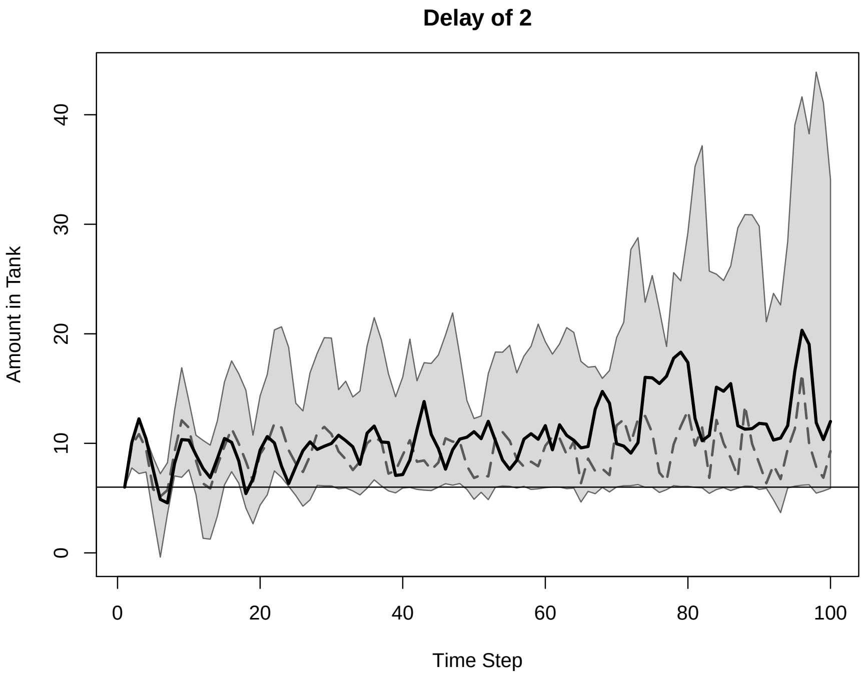


Figure 6: Comparison of human performance and model predictions in the 'Delay of 2' condition. Dashed line: median of the human sample. Grey area: second and third quartile of the human data. Bold line: model prediction. Horizontal line at 6 gallons: Target value.

significance level above five percent. In this light, the p values given in the previous sections should be interpreted with care.

Furthermore, the given p values are probabilites for alpha errors when beta errors (to miss an existing effect) would be of more interest.

5.2 Model complexity

When comparing model fit, one has to take model complexity into account (Halbrügge, 2007b). Overly complex models tend to overfit, that is they show extremely good fit on data that were used during model development, but do not generalize well to new data, even when the new data has been taken in the same experimental condition that was used at first.

The easiest way to assess its complexity is to count free parameters in a model (Baker, Corbett, and Koedinger, 2003). For the model described in this article, these are the initial utilities of the 'decision' productions, the rewards (see Table 1), and ACT-R's architectural parameters of the procedural learning system (learning rate and utility noise), adding up to 15 numerical values.

Despite the high number of free parameters, the overall complexity is rather low. If one thinks of Kolmogorov complexity (Li and Vitányi, 1997) as the best suitable measurement for cognitive models, the model presented here is really simple. The most complex analytical decision production can be expressed by a few lines of C:

```
diff = goal - amount; // current difference
diff1 = diff - lastDiff; // first derivation
diff2 = diff1 - lastDiff1; // second derivation

out = - (diff + diff1 + diff2);

lastDiff = diff; // save values for next run
lastDiff1 = diff1;
```

The comparatively good performance on the transfer data of the DSF challenge can be interpreted as a measurement of generalizability of the model. This view would be consistent with a rather low complexity of the model as is supposed here.

6. Conclusions

The cognitive model described in this article was developed with the aim to explain the different learning curves of the human subjects in the four training conditions. With regard to the goodness-of-fit statistics developed in Section 2.3, this goal was met.

Only two theoretical assumptions were made, first that humans can cope with linear relationships but not non-linear ones, and second that the idea of a container holding more than its original volume is hard to imagine.

As part of the DSF modeling comparison challenge, the model was exposed to previously unknown transfer conditions. In these situations, the model performed reasonably well, i.e. the goodness-of-fit was still acceptable. However, a closer inspection showed that at least in some conditions the cognitive model represented only a subset of the human sample, namely a group of subjects that did not notice the underlying structure of the task (e.g., periodicity in the ‘sequence’ conditions).

Nonetheless, the model scored third in the DSF challenge. This relative success can be seen as an approval of the modeling methodology chosen for the development, i.e. to stick to as few theoretical assumptions as possible and to stop the refinement of the model as soon as an acceptable p is reached and no substantial improvements are expectable. In short: Keep it simple!

Acknowledgments

The author would like to thank Hendrik Neumann for comments on an early draft.

References

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglass, S.; Lebiere, C.; and Qin, Y. 2004. An Integrated Theory of the Mind. *Psychological Review* 111(4):1036–1060.

Baker, R. S.; Corbett, A. T.; and Koedinger, K. R. 2003. Statistical Techniques For Comparing ACT-R Models of Cognitive Performance. In *Proceedings of the 10th Annual ACT-R Workshop*.

Bertsekas, D. P. 2000. *Dynamic Programming and Optimal Control*. Belmont, Mass.: Athena Scientific.

Bortz, J. 1999. *Statistik für Sozialwissenschaftler*. Berlin: Springer, 5. edition.

Davison, A. C., and Hinkley, D. V. 1997. *Bootstrap Methods and their Application*. New York: Cambridge University Press.

Dörner, D. 1989. *Die Logik des Misslingens*. Reinbek: Rohwohlt Verlag GmbH.

Dutt, V., and Gonzalez, C. 2007. Slope of Inflow Impacts Dynamic Decision Making. In *Proceedings of the 25th International Conference of the System Dynamics Society*, 79. Boston, MA: System Dynamics Society.

Gabel, T., and Riedmiller, M. 2010. On Progress in RoboCup: The Simulation League Showcase. In *RoboCup 2010: Robot Soccer World Cup XIV*, Lecture Notes in Artificial Intelligence. Berlin: Springer.

Gluck, K. A.; Stanley, C. T.; Moore, L. R.; Reitter, D.; and Halbrügge, M. 2010. Exploration for Understanding in Cognitive Modeling. *Journal of Artificial General Intelligence* 2(2):89–108.

Halbrügge, M. 2007a. Clusteranalysen für Blicksequenzen. In Röttling, M.; Wozny, G.; Klostermann, A.; and Huss, J., eds., *Gestaltung von Mensch-Technik-Interaktion – 7. Berliner Werkstatt Mensch-Maschine-Systeme*, number 25 in Fortschritt-Berichte VDI Reihe 22, 17–22. Düsseldorf: VDI Verlag GmbH.

Halbrügge, M. 2007b. Evaluating Cognitive Models and Architectures. In Kaminka, G. A., and Burghart, C. R., eds., *Evaluating Architectures for Intelligence. Papers from the 2007 AAAI Workshop*, 27–31. Menlo Park, California: AAAI Press.

Lebiere, C.; Gonzalez, C.; and Warwick, W. 2009. A comparative approach to understanding general intelligence: Predicting cognitive performance in an open-ended dynamic task. In Goertzel, B.; Hitzler, P.; and Hutter, M., eds., *Proceedings of the Second Conference on Artificial General Intelligence*, 103–107. Amsterdam-Paris: Atlantik Press.

Lehmann, E. L., and Romano, J. P. 2005. *Testing Statistical Hypotheses*. New York: Springer, 3. edition.

Li, M., and Vitányi, P. 1997. *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer, 2. edition.

Mirkin, B. 1996. *Mathematical classification and clustering*. Dordrecht: Kluwer Academic Press.

Stone, P. 2003. Multiagent Competitions and Research: Lessons from RoboCup and TAC. In Kaminka, G. A.; Lima, P. U.; and Rojas, R., eds., *RoboCup 2002: Robot Soccer World Cup VI*, Lecture Notes in Artificial Intelligence, 224–237. Berlin: Springer.

Wagenaar, W. A., and Sagaria, S. D. 1975. Misperception of exponential growth. *Perception & Psychophysics* 18(6):416–422.