JDDM
Journal of Dynamic Decision Making

Open Access

Original Research

A web-based feedback study on optimization-based training and analysis of human decision making

Michael Engelhart[1], Joachim Funke[2], and Sebastian Sager[1]

[1] Faculty of Mathematics, Otto-von-Guericke-Universität Magdeburg and [2] Ruprecht-Karls-Universität Heidelberg

The question "How can humans learn efficiently to make decisions in a complex, dynamic, and uncertain environment" is still a very open question. We investigate what effects arise when feedback is given in a computer-simulated microworld that is controlled by participants. This has a direct impact on training simulators that are already in standard use in many professions, e.g., flight simulators for pilots, and a potential impact on a better understanding of human decision making in general.
Our study is based on a benchmark microworld with an economic framing, the *IWR Tailorshop*. N=94 participants played four rounds of the microworld, each 10 months, via a web interface. We propose a new approach to quantify performance and learning, which is based on a mathematical model of the microworld and optimization. Six participant groups receive different kinds of feedback in a training phase, then results in a performance phase without feedback are analyzed. As a main result, feedback of optimal solutions in training rounds improved model knowledge, early learning, and performance, especially when this information is encoded in a graphical representation (arrows).

Keywords: Complex problem solving, training, dynamic decision making, feedback, mixed-integer nonlinear optimization, Tailorshop

Modern life imposes daily decision making, often with important consequences. Illustrative examples are politicians who decide on actions to overcome a financial crisis, medical doctors who decide on complementary chemotherapy drug delivery strategies, or entrepreneurs who decide on long-term strategies for their company.

The process of human decision making is the subject of research in the field of *Complex Problem Solving (CPS)*, which deals with *complex problems*. The complexity may result from one or several different characteristics, such as a coupling of subsystems, nonlinearities, dynamic changes, opaqueness, or others (Dörner, 1980). Such problems are considered to be similar to problems we encounter and solve in everyday life. Thus, investigation of CPS is claimed to yield more insight into real-world human decision making than *simple problems* with a well-defined problem space, like the *Tower of Hanoi*. Apparently, our introductory examples are complex problems and as such, they are ill-defined. More precisely, their problem space is open and a problem solver has to deal with lots of variables, dependencies and dynamics making them complex problems: Which information is relevant? How is the data connected? What is the exact aim?

The main intention in CPS research is to understand how certain *exogenous variables* influence a solution process. In general, *personal and situational variables* are differentiated. The most typical and frequently analyzed personal variable is *intelligence*. It is an ongoing debate how intelligence influences complex problem solving (Wittmann & Hattrup, 2004). Other interesting personal variables are *working memory* (Robbins et al., 1996), *amount of knowledge* (Kluwe, 1993), and *emotion regulation* (Otto & Lantermann, 2004). Situational variables like the impact of *goal specificity and observation* (Osman, 2008), *feedback* (Brehmer, 1995), and *time constraints* (Gonzalez, 2004) attracted less attention. In a recent work (Selten, Pittnauer, & Hohnisch, 2012), an abstract computer-simulated monopoly market is used to investigate dynamic decision making based on the choice of *goal systems*. For investigations in the field of CPS, computer-based simulations of small parts of the real world, *microworlds*, are frequently used. These simulations present users with situations similar to those encountered when attempting to solve real-world complex problems, but offer researchers the possibility to conduct studies under controlled conditions. In CPS, the performance of participants in a clearly defined microworld is investigated, evaluated and correlated to certain characteristics, such as the participant's capacity to regulate emotions.

Previous research with the microworld Tailorshop

One microworld that comprises a variety of properties such as dynamics, complexity and interdependence, discrete choices, lack of transparency, and polytely in an economical framing is the *Tailorshop*. Participants have to make economic decisions to maximize the overall balance of a small company, specialized in the production and sales of shirts. The *Tailorshop* sometimes is referred to as the *Drosophila* for CPS researchers (Funke, 2010) and thus is a prominent example for a computer-based microworld. It has

Corresponding author: Sebastian Sager, Otto-von-Guericke-Universität Magdeburg: sager@ovgu.de

been used in a large number of studies, e.g., Putz-Osterloh, Bott, and Köster (1990); Kluwe, Misiak, and Haider (1991); Kleinmann and Strauß (1998); Meyer and Scholl (2009); Barth (2010); Barth and Funke (2010). Comprehensive reviews on studies with *Tailorshop* have also been published, e.g., Frensch and Funke (1995); Funke (2003); Funke and Frensch (2007); Funke (2010).

The calculation of *indicator functions* to measure performance of CPS participants is by no means trivial. To measure performance within the *Tailorshop* microworld, different indicator functions have been proposed in the literature, see Danner, Hagemann, Schankin, Hager, and Funke (2011) for a recent review. Hörmann and Thomas (1989) proposed a comparison of the variable which the participants were requested to maximize. Such a performance criterion seems natural. However, it cannot yield insight into the temporal process and is not objective in the sense that the performance depends on what other participants achieved. Analyzing the temporal evolution of other variables of this microworld has also been proposed (see, e.g., Putz-Osterloh (1981); Süß, Oberauer, and Kersting (1993); Funke (1983); Barth and Funke (2010)). An obvious drawback of comparing the development of variables which were not the actual objective for the participants is that a monotonic development does not necessarily indicate good or even optimal decision making.

The lacking availability of an objective performance indicator is an obstacle for analysis and it has often been argued that inconsistent findings are due to the fact that an objective indicator function yielding detailed insight into the participants' performance is not available, e.g., in Wenke and Frensch (2003). To overcome this problem, we propose to use indicator functions based on optimal solutions. In Sager, Barth, Diedam, Engelhart, and Funke (2010) as well as ins Sager, Barth, Diedam, Engelhart, and Funke (2011) the question of how to get a *reliable performance indicator* for the *Tailorshop* microworld has been addressed. Because all previously used indicators have unknown reliability and validity, decisions are compared to mathematically optimal solutions. For the first time, a complex microworld such as *Tailorshop* has been described in terms of a mathematical model. Thus, the assumption that the *fruit fly of complex problem solving* is not mathematically accessible has been disproved. This novel methodological approach has also been combined with experimental studies (Barth, 2010; Barth & Funke, 2010; Sager et al., 2011) but beyond these works, has to our knowledge not yet received much attention.

Training and relation to optimization

With tasks for humans becoming more complex in the real-world, there is also an increasing need to train and assist persons performing complex tasks. In Hüfner, Tometzki, Kraja, and Engell (2011), a framework for training engineering students in designing controllers for complex systems like chemical reactors is presented. In this approach, students can learn from the results of simulations depending on their inputs. In the context of *CPS*, an interesting approach would be to determine optimal solutions and corresponding controls for a microworld to compute a feedback for participants to support and train them. However, as Cronin, Gonzalez, and Sterman (2009) show, the presentation of information in a dynamic context is crucial for the success of the participants. To the best of our knowledge, there have been no studies investigating the effects of an optimization-based feedback.

So far, CPS microworlds have been developed in a purely disciplinary trial-and-error approach. A systematic development of CPS microworlds based on a mathematical model, sensitivity analysis, and eventually optimization methods to choose parameters that lead to a wanted behavior of the complex system has not yet been applied. An example for this necessity is the fact that the mathematical modeling of the *Tailorshop* microworld in Sager et al. (2011) led to the discovery of unwanted and unrealistic winning strategies. Based on this experience with modeling oddities, bugs, and other undesirable properties, a new microworld has been built from scratch designed as a mathematical model for CPS by Engelhart, Funke, and Sager (2013), the *IWR Tailorshop*. The *IWR Tailorshop* is the first CPS test-scenario with functional relations and model parameters that have been formulated based on optimization results yielding desirable (mathematical) properties. Compared to the *Tailorshop*, the setting is slightly more general. For example, *machines* have been replaced by *production sites*, and *vans* by *distribution sites*.

The optimization problems that need to be solved in the context of the *IWR Tailorshop* scenario are *mixed-integer nonlinear programs (MINLP)* with non-convex continuous relaxations. Whenever optimization problems involve variables of continuous and discrete nature together, the term *mixed-integer* is used. In this case they can be interpreted as *discretized optimal control problems (dMIOCPs)*. We use the mathematical approaches presented in Engelhart et al. (2013) and Engelhart (2015) that are based on a tailored decomposition technique to determine ε–optimal solutions for *IWR Tailorshop* in (almost) real time.

About this study

In the interest of a compact presentation we focus on the most important results of a study which has been described in full detail in the PhD thesis of Engelhart (2015).

Method

We describe the Tailorshop microworld, the feedback study with the experimental groups, the hypotheses, a prestudy, details of the data collection, and the statistical methods.

IWR Tailorshop: A new complex microworld

We work with a systematically built new microworld with controlled properties, the *IWR Tailorshop*. It was first described in Engelhart et al. (2013) and Engelhart (2015) and is based on the economical framing of *Tailorshop*. Table 1 lists all states and controls (interventions for the participants) that the *IWR Tailorshop* contains together with corresponding units. The final mathematical model of the *IWR Tailorshop* consists of 14 state variables x (i.e., dependent variables) and 10 control variables u (i.e., independent variables) including 5 integer controls. All equations and constraints, the objective function, and the parameter and initial values are specified in the Appendix.

States	**Variable**	**Unit***
employees	x^{EM}	person(s)
production sites	x^{PS}	site(s)
distribution sites	x^{DS}	site(s)
shirts in stock	x^{SH}	shirt(s)
resources in stock	x^{RS}	shirt(s)
production	x^{PR}	shirt(s)
sales	x^{SA}	shirt(s)
demand	x^{DE}	shirt(s)
reputation	x^{RE}	—
shirts quality	x^{SQ}	—
machine quality	x^{MQ}	—
resources quality	x^{RQ}	—
motivation of empl.	x^{MO}	—
resources price**	x^{RP}	$^{\text{MU}}/_{\text{shirt}}$
capital	x^{CA}	MU
Controls	**Variable**	**Unit***
shirt price	u^{SP}	$^{\text{MU}}/_{\text{shirt}}$
advertising	u^{AD}	MU
wages	u^{WA}	$^{\text{MU}}/_{\text{person}}$
working conditions**	u^{WC}	MU
maintenance	u^{MA}	MU
buy resources**	u^{DRS}	shirt(s)
sell resources**	u^{dRS}	shirt(s)
resources quality	u^{RQ}	—
recruit/dismiss empl.	u^{dEM}/u^{DEM}	person(s)
create production site	u^{DPS}	site(s)
close production site	u^{dPS}	site(s)
create distribution site	u^{DDS}	site(s)
close distribution site	u^{dDS}	site(s)

Table 1. States and controls in the *IWR Tailorshop* microworld (* *MU* means monetary units, ** not part of the final model for the web-based study).

The equations describe how the different state and control variables are connected. Some of these equations may be trivial, as, for example, the number of production sites (x^{PS}) in Equation (A.1b) in the Appendix, where the numbers of newly created (u^{DPS}) or closed distribution sites (u^{dPS}) are added to or subtracted from the current value. They may also involve more variables and include nonlinear expressions as, e.g., in the demand which depends nonlinearly on shirt price, advertisement, reputation, and others, compare Equation (A.1d). These mathematical relations are intransparent to the study participants, as it is a part of the task to explore and understand the microworld.

The objective is the maximization of the capital at the end of the discrete time-scale in this work, see Equation (A.4) in the Appendix. The constraints are basically bounds on the controls or non-negativity of variables. The objective is communicated to participants, the constraints can be determined from admissible values in the web interface.

IWR Tailorshop has been implemented including different optimization-based feedback methods in a web-based interface, compare Figure 1. For the analysis of data collected with this interface, optimization-based analysis methods have been implemented in the analysis software *Antils*. Both the web front end and the analysis back end are available as open-source software under the *GPL (GNU General Public License)* and thus can easily be used for further investigations. Analysis and feedback based on optimal solutions enabled insights on human decision making which else would not have been possible.

A web-based feedback study

From November to December 2013, we conducted a feedback study with the described *IWR Tailorshop* microworld. We collected data from 148 participants ($N = 94$ after removal of incomplete datasets and outliers, see below) and applied our optimization-based analysis and feedback approach. The participants were asked to play four rounds with 10 "months" each of the economic simulation via its web interface. Different approaches for both feedback computation and feedback presentation have been applied in the first two rounds (so-called *training* or *feedback rounds*). In the last two rounds, however, no one received any feedback. These rounds will be referred to as *performance rounds*.

Task. Participants had to play four rounds of the *IWR Tailorshop* microworld of 10 months each via its web interface. They were allowed to interrupt the process at any time. For the four rounds, different initial values were used, see Table A.3 in the Appendix, but the same for all participants. Rounds 1 and 3 started with the same values, whereas in rounds 2 and 4 pairwise different values were used. Control values for recruitment and dismissal of employees and creation and closing of sites were always reset to 0 in order to avoid accidental execution.

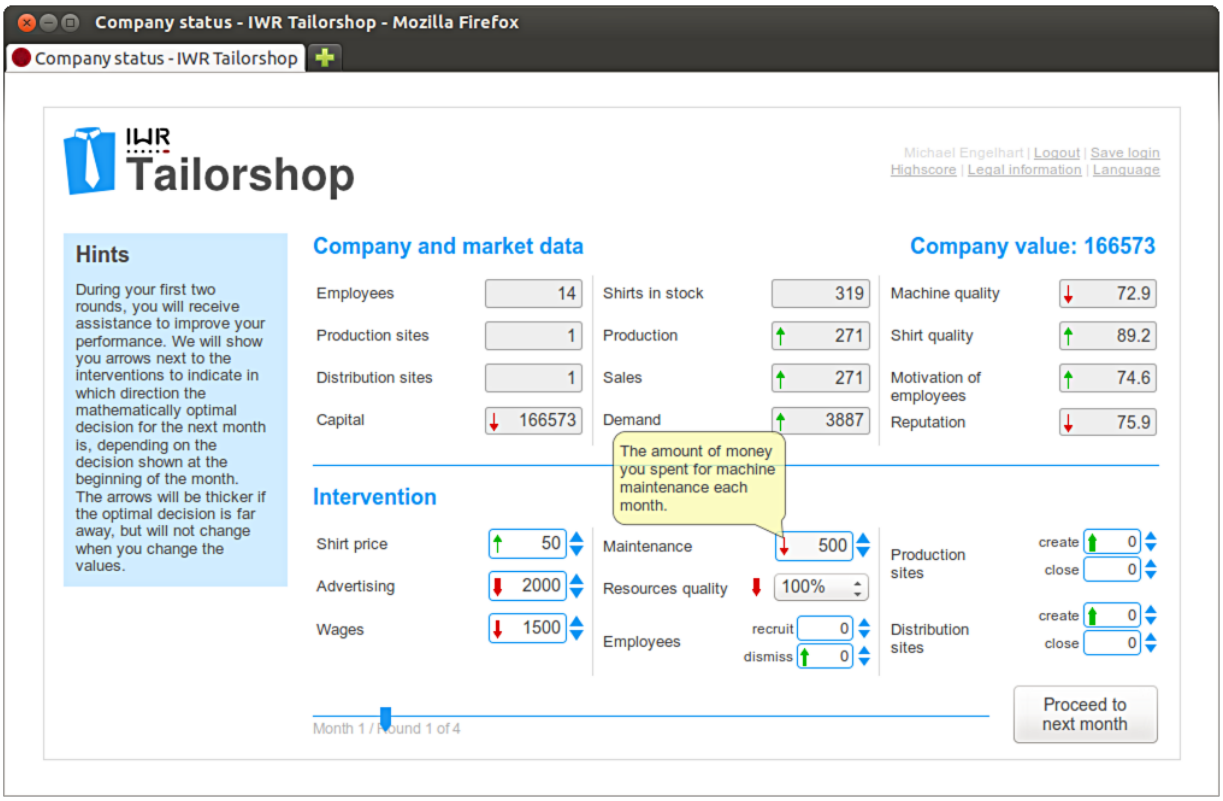


Figure 1. The *IWR Tailorshop* web interface with arrows as feedback for the *trend* group (compare Figure 2) and a hint for *maintenance* control.

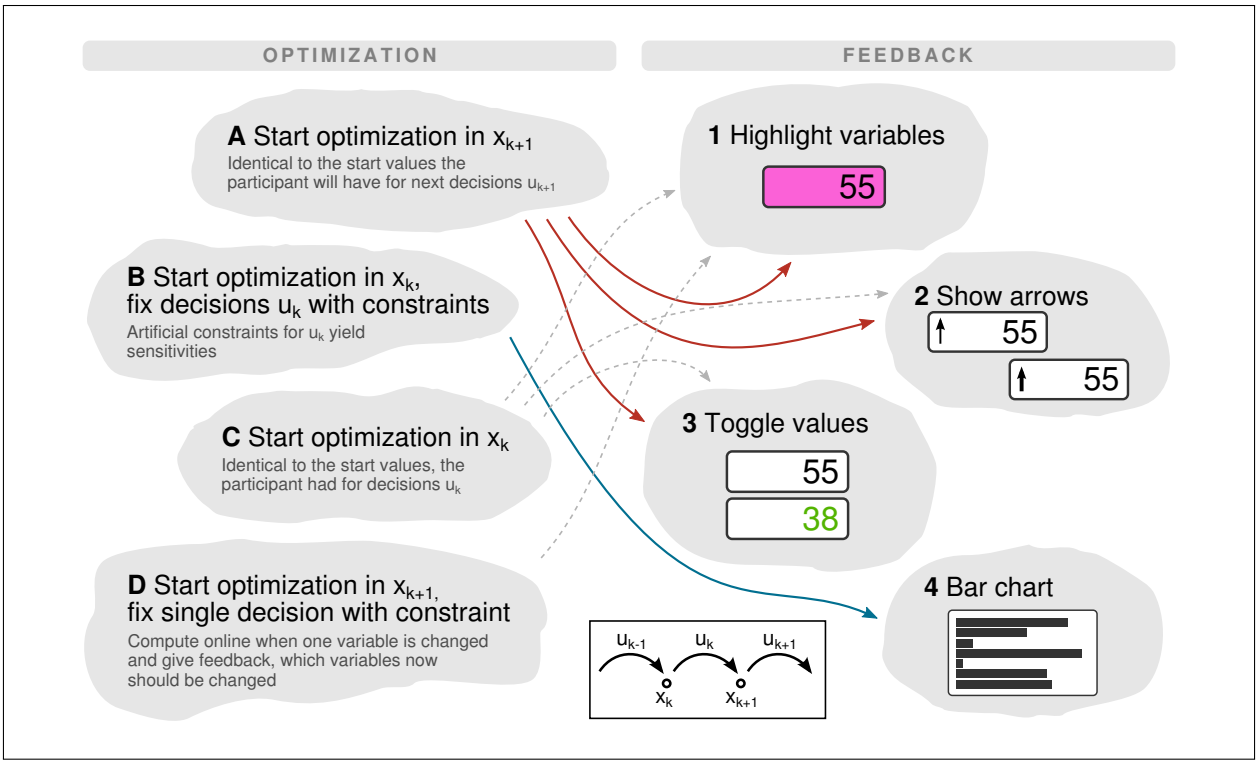


Figure 2. Optimization-based feedback at month $k+1$: on the left hand side, there are different methods to compute a feedback and on the right hand side there are different types of feedback presentation. Optimization method A is used with feedback presentations 1, 2, and 3 (corresponding to *indicate* group, *trend* group, and *value* group) and optimization method B is used with feedback presentation 4 (corresponding to *chart* group). Note that x_k refers to the state variables and u_k refers to the control variables of month k.

As an incentive, there was a competition with chances weighted according to success in which participants could win one of six 20 euro *Amazon* gift cards. For this, only the results of performance rounds were considered.

Procedure. For the main task, the control of the *IWR Tailorshop* microworld, the participants received guidance by the following introduction:

> *Thank you! Now you can start into the **IWR Tailorshop microworld**. Please note, that you need to finish **4 rounds of 10 "months" each** to participate in the competition.*
>
> *All in all it will take you about **30–45 minutes**. You ideally play all 4 rounds at a stretch, but you may interrupt after each "month" and continue at a later date. The first two rounds are **training rounds**, only your points (not your rank) in the last two rounds are considered for the drawing.*
>
> *Now, please imagine you are the **head of a company, which produces shirts**. Your aim is to **maximize** the **company's capital** at the end of each round, i.e., in month 10. For this there are several possibilities of **intervention** available, which will be located in the lower part. In the upper part you will find **important figures** of your company.*
>
> *However, your intervention possibilities are subject to certain **constraints**, e.g., you are not allowed to close all company sites. At the end of each round, you will find a **highscore table** and after the last round the table, which is important for the competition. In the blue **hint box** you can find assistance and useful hints during your game. Good luck!*

The *hint box* the introduction refers to was displayed at the left side and contained hints corresponding to the situation and the feedback group the participant was in, compare Figure 2, e.g.,

> *During your first two rounds, you will receive assistance to improve your performance. We will show you arrows next to the interventions to indicate in which direction the mathematically optimal decision for the next month is, depending on the decision shown at the beginning of the month. The arrows will be thicker if the optimal decision is far away, but will not change when you change the values.*

Hints on each state and control, e.g., "*the wages for each employee per month in money units*" for control *wages*, were available as a tooltip on mouse rollover. After each round, participants were shown an anonymized highscore list with the top 20 participants in their group.

Additional variables. Additional information on the participants was collected via three questionnaires. The first survey comprised gender, interest in economics, interest in computer games, age, and a self-assessment of systematic problem solving. This survey had to be answered before participants could start the main task, i.e., the four *IWR Tailorshop* rounds. The other two surveys were carried out after the main task. The second survey was targeted on participants' model knowledge. Participants were shown five claims about the *IWR Tailorshop* microworld and had to decide if they were right or wrong, compare Table A.8 in the Appendix. Final survey was the 10-item short version of the *Big Five Inventory* test proposed by Rammstedt and John (2007) to measure the *Big Five dimensions* of personality (Digman, 1990), i.e., *agreeableness*, *conscientiousness*, *extraversion*, *neuroticism*, and *openness*.

The experimental groups

Participants were divided randomly into six groups based on pseudorandom numbers generated by a *Mersenne twister* (Matsumoto & Nishimura, 1998). They differ in the way they received additional information in the first two (feedback) rounds. Compare Figure 2 for an illustration of the optimization-based feedback. The six groups were designed as follows.

The ***control* group** (co) did not receive any feedback.

The ***highscore* group** (hs) received a feedback based on the results of previous participants during training rounds, giving a ratio of participants who performed better and worse of the kind "Until now x% of participants performed better and y% performed worse than you."

The ***indicate* group** (in) received optimization-based feedback via highlighted control values. Variables are highlighted if they differ from the optimal value more than a given threshold, e.g., 30 % of the difference δ between lower and upper bound of a variable.

The ***trend* group** (tr) received optimization-based feedback via up and down *arrows* in different thickness. Arrow thickness is also determined by thresholds depending on δ. Arrows indicate the direction of the optimal control: if the optimal control value is larger, the arrow points up and vice versa.

The ***value* group** (va) received optimization-based feedback via toggled *values*, showing the optimal solution. Note that participants of this group could theoretically obtain a 100% performance (in the two feedback rounds) by simply copying all values.

The ***chart* group** (ch) received optimization-based feedback via bar charts. LAGRANGE multipliers are displayed scaled according to δ. These dual variables indicate the sensitivity of the objective function with respect to the current value.

Hypotheses

Before the beginning of the study, specific hypotheses were formulated. In the interest of a compact presen-

tation, we list a subset of them directly in the corresponding result sections in Tables 2, 3, 4, 5, 6, 8.

The full set of hypotheses that have been formulated and tested can be found in the PhD thesis of Engelhart (2015). They concern correlations with the additional variables mentioned above (computer games, economic interest, gender, age, Big Five) and a detailed analysis of processing times. No statistically significant effects were found (for age and gender possibly due to low numbers of old/female participants). Therefore in this paper we focus on the main result, namely the impact of optimization-based feedback on performance and learning.

Prestudy

In October 2013, 18 participants (recruited directedly via e-mail) took part in a prestudy. The aim was twofold: on the one hand, this was a test under realistic conditions for the main study and an opportunity to eliminate bugs in the interface. On the other hand, the data were used for *highscore* feedback in the main study. This was particularly necessary to avoid a feedback like "0% performed better and 0% worse than you" for the first participant in that group. However, the data were considered neither in our statistical nor in our optimization-based analysis.

Data collection

Starting from November 15, 2013, the study was announced in several first and third term lectures for mathematics, physics, computer science, engineering, and psychology students at *Heidelberg University* and *Otto von Guericke University Magdeburg* in Germany. These announcements were complemented by public announcements in the social networks *Google+* and *Facebook* as well as selective announcements via e-mail.

Potential participants were informed that they would have to play four rounds of the economic simulation *IWR Tailorshop* via a device of their choice with a web browser (e.g., PC, tablet, or smartphone) which in total would take approximately 30–45 minutes. It was advertised as an incentive that there will be a competition with chances weighted according to success where participants can win one of six 20 euro *Amazon* gift cards. The deadline for participation was December 15, 2013.

Participants had to create an account with an e-mail address, which they needed to confirm in order to avoid multiplicate participation. Creating multiple accounts was also prohibited by terms of participation leading to exclusion from the competition.

Until the end of data collection, 157 accounts were registered for participation. Two accounts have not been activated, maybe because of erroneous e-mail addresses or the like. Furthermore, seven participants did not answer the first survey and therefore could not start the main task, i.e., no data was recorded for them at all. Thus, we received data from 149 participants, of which 101 provided complete datasets, i.e., they played four full rounds and answered all three surveys. One account was identified as a duplicate participation and was excluded from the analysis. The first account of the corresponding participant was part of the analysis, but was not considered in the competition. This resulted in 100 complete datasets and 148 datasets in total for our statistical analysis.

Model knowledge

A true/false questionnaire, Table A.8 in the Appendix, was used at the end of the four rounds to determine the participants' knowledge about the *IWR Tailorshop* microworld. The overall ratio of correct answers varies a lot for the five claims. This shows that the questions had a varying difficulty, which was intended.

Correct answers were identified as *knowledge* about the model. Participants who chose *don't know* were considered to be *uncertain* about the corresponding claim.

Statistical methods

Statistical analysis of the data was done using the open source package *R Version 3.0.1* (R Development Core Team, 2008).

Statistical significance. We tested the statistical significance of differences between means of scores and other variables. To this end we applied Student's t-test and Welch's t-test. Usually all tests have also been confirmed qualitatively by Wilcoxon rank sum tests.

For all tests, p-values of < 0.05 were considered statistically significant (i.e., $\alpha = 0.05$). All such values are printed in bold face in tables.

Normality of distributions. Statistical tests like Student's t-test and Welch's t-test require normality of the population—although these two are known to be relatively robust against non-normality (e.g., Sawilowsky & Blair, 1992).

We applied the implementation of the Kolmogorov-Smirnov test for normality (Lilliefors, 1967) from the R package *nortest* to the score variables. For this test, the alternative hypothesis is that the data is *not* normally distributed.

Student's t-test—in contrast to Welch's t-test—also requires homogeneity of variances between the groups. This has been tested using Levene's test (Levene, 1960), Brown-Forsythe test (Brown & Forsythe, 1974; both as implemented in R package *lawstat*), and Bartlett's test (Bartlett, 1937).

For $\alpha = 0.05$, the hypothesis of the data being normally distributed cannot be rejected for most groups and rounds by a majority of the applied tests for normality.

However, we cannot assume homogeneous variances between feedback groups. Thus, for the sake of comparability, Welch's t-test will be used for comparison of score means for *all* rounds.

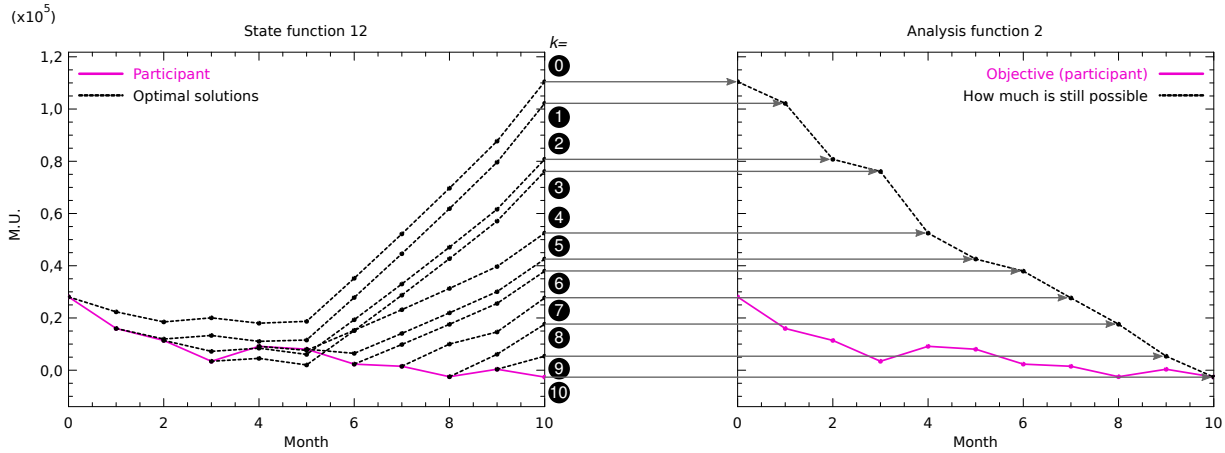


Figure 3. Relation between optimal scores and the *How Much is Still Possible*-function, illustrated for a specific participant. Left: development of score (capital) over time. An optimization starting at month k provides an optimal value that could have been achieved. The specific shape of the optimal solutions (approximately constant, then linear increase of capital) is due to an investment that pays off later. Therefore taking the score itself as an indicator is not a good performance measure. Right: The optimal objective values at the final month 10 are plotted for different starting months k, resulting in the *How Much is Still Possible*-function. Participant decisions are good (even optimal), whenever the values stay constant, and the worse, the more it decreases.

Dropouts and outliers. 148 datasets have been considered, 100 of which were complete. Our statistical analysis showed that incomplete datasets did not show any systematic differences compared to complete datasets. In particular, there were no significant effects on the dropout concerning feedback group, gender, or the performance until the dropout.

Grubbs' test is a statistical test proposed by Frank E. Grubbs (1950, 1969) which detects one outlier at a time in a normally distributed population. We used the implementation of Grubbs' test available in the R package *outliers*. Another approach are the *outer fences* for boxplots, as described by John W. Tukey (1977).

An analysis of the score variable with Grubbs' test and *outer fences* detected 6 severe outliers, which were excluded from further analysis. The analysis in the remainder, including the optimization-based analysis, is therefore based on N=94 datasets.

Optimization-based analysis

As discussed in the Introduction, measuring performance in a complex microworld is by no means trivial. In previous work we suggested a completely novel approach: to use mathematical optimization and the so-called *How Much is Still Possible*-function and the *Use of Potential*-function (Sager et al., 2011; Engelhart et al., 2013). We applied these techniques also in the current study as follows.

Optimization. We computed optimal solutions for each participant (1 to 94) and round (1 to 4) and month (1 to 10). As illustrated in Figure 2, the starting value is identical to the one of the participant in the specific round and month, and hence pairwise different. Alltogether, we solved $94 \cdot 4 \cdot 10 = 3760$ mixed-integer nonlinear optimization problems for our analysis, using a specifically developed optimization algorithm (Engelhart et al., 2013; Engelhart, 2015). Note that this approach is very similar to the computation of an optimization-based feedback, compare Figure 2. The main difference is whether this is done a priori (feedback for training) or a posteriori (analysis).

How much is still Possible-function. The optimal solution starts in the identical state as the participant in a specific round and month. Hence we know how much could have been achieved if all of the participant's future decisions would have been optimal. The optimal objective function values are interpreted as a monotonically decreasing function (because participants can't do better than the optimal solution) over rounds and months. An illustrating example is shown in Figure 3.

Use of Potential-function. The *Use of Potential*-function is derived from the *How Much is Still Possible*-function by taking the difference between two succeeding months. Doing this for each month one obtains a function that indicates how much of the potential of optimal decisions was used by a participant.

Learning

To enable conclusions on learning effects, we are going to analyze the *Use of Potential* function. As this function indicates how close to optimality the decisions of a participant (group) for each month were, the function can be seen as a learning curve. We experimented with different functional parameterizations, and decided eventually to use a piecewise *linear model* for our analysis.

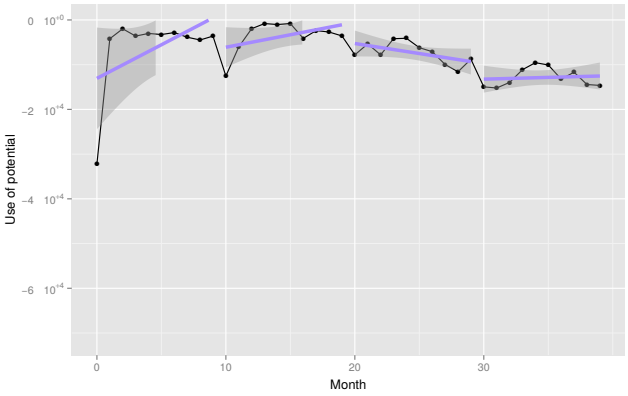
We used R's *lm* to fit the linear model for *Use of Potential* for each participant and each round,

$$y = m \cdot x + c, \tag{1}$$

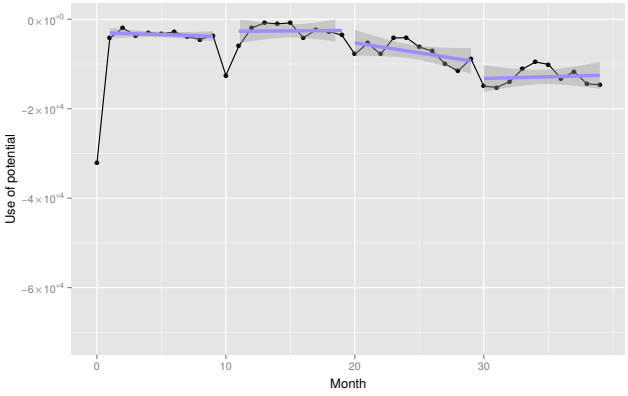
based on given values y_i of *Use of Potential* at months $x_i = i$. The regression parameters are m and c, and estimate the gradient and the intercept of *Use of Potential*.

The estimate m for the gradient characterizes how much more potential the participant was able to use over time, i.e., how much the participant *learned*. We use statistical tests on the values of m for different participant groups for our a priori hypotheses on learning.

The first months of the feedback rounds (i.e., months 1 and 11) were not considered for the linear regression. No feedback is given before the *first* decision and thus *Use of Potential* may change drastically from month 0 to month 1.



(a) based on all months



(b) without first month

Figure 4. Regression lines for *Use of Potential* for *value* group over all rounds (one round consists of 10 months). (a) shows a regression with all months of each round, for (b) the first month of feedback rounds has been excluded.

The importance of this is shown in Figure 4, where Figure 4a exhibits linear regressions based on all months, and Figure 4b the corrected approach. In performance rounds this effect does not occur, so all months are considered.

Technical implementation

For data collection, the *IWR Tailorshop* web interface was used, which is implemented using *XHTML* and *JavaScript* with *jQuery 1.10* and usage of *AJAX* client-side, complemented by a server-side *PHP* code. For the online optimization, *AMPL Version 20131012*

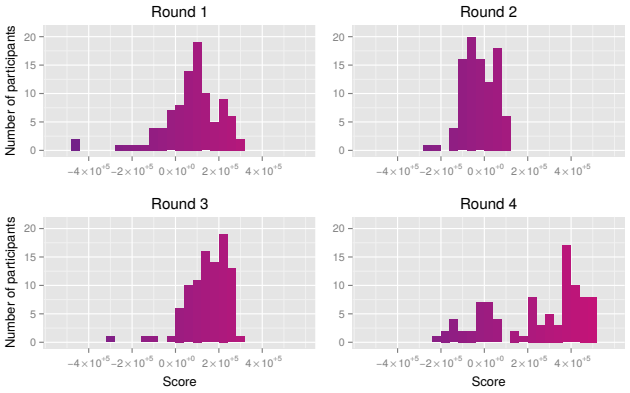


Figure 5. Score histogram for all four rounds for all complete datasets without 6 outliers ($N = 94$).

together with *Bonmin 1.5* and *Ipopt 3.10* was used via *IWR Tailorshop*'s *AMPL* interface. The web server for the study was an *Intel Core i7 920* machine with 12 GB RAM running *PHP 5.5* and *MySQL 5.5* with an *Apache 2.4 HTTP server* on *Ubuntu 13.10 64-bit*. The web interface implemented a so-called *responsive grid*, which allowed participants to use both mobile devices and desktop PCs conveniently. Usage statistics based on user logins show that approximately 20% of participants used mobile devices.

The methods for an optimization-based analysis are implemented in the open-source software package *Antils* (*An*alysis *T*ool for *I*WR Tailorshop Resu*l*ts and *S*olutions). All computations were carried out on an *Intel Core i7 920* machine with 12 GB RAM running *Ubuntu 14.04 64-bit*. For the solution of the arising optimization problems, *AMPL Version 20140331* together with *Bonmin 1.5* and *Ipopt 3.10* was used via *IWR Tailorshop*'s *AMPL* interface.

Results

We are going to test hypotheses related to the different participant groups. First we will focus on performance, second on learning, and third on model knowledge. We will close by an illustrating investigation of the strategies of exemplary participants.

We start with a look at the score and the *Use of Potential*–functions of the study participants. Figure 5 shows how the performance (score) is distributed over all participants in the four rounds. Note that rounds 1 and 3 had the same initial values, whereas rounds 2 and 4 had different initial values, and thus also different optimal solutions and scores. Rounds 1 and 2 are training rounds with feedback, rounds 3 and 4 performance rounds without feedback.

Obviously, it is only meaningful to investigate the impact of the different types of feedback, if the role of the participants' prerequisites is not a decisive factor (e.g., because one group simply consisted of better problem solvers at the beginning of the study). This would have biased the groups' performance and is relevant, given the low number of samples for some of the participant groups.

Hypothesis		Confirmed
(A)	initial performance is not important for final performance	✓

Table 2. Hypothesis about participant prerequisites.

The optimization-based analysis gives us the possibility to check this by comparing the first *Use of Potential* value. At this point, all participants had received the same information, as feedback only started after the first decision, so there should be no significant difference in the performance. Table A.4c contains mean values, Kolmogorov-Smirnov test results, and Welch's t-test results (in comparison to *control* group). The Kolmogorov-Smirnov test shows that the first *Use of Potential* values can be considered to be normally distributed for all groups. No significant differences to *control* group can be observed by the Welch's t-test for all groups, so we can suppose that there were no systematic differences among the participants of the six groups. Correlation between first *Use of Potential* and score in performance rounds is 0.067, confirming Hypothesis (A), see Table 2.

Effects of optimization-based feedback on performance

We investigate whether the optimization-based feedback in the first 2 training rounds had a significant effect on the performance in the rounds 3 and 4, where no feedback was given (Table 3). We start by looking at all optimization participants, i.e., the ones in groups indicate (in), trend (tr), value (va), and chart (ch). We assess Hypothesis (B) visually via Figure 6

Hypothesis		Confirmed
(B)	participants with **optimization-based feedback** perform better overall, better in feedback rounds, and better in performance rounds compared to control group	✓
(C)	**control group** performs worst overall and performs worse in performance rounds than groups with optimization-based feedback	—

Table 3. Hypotheses related to performance of participants who received optimization-based feedback (groups in, tr, va, ch) and to performance of the *control* group.

and statistically via Table A.4a.

Figure 6 shows a boxplot of the different participant groups' performance via the obtained score. The four groups which received optimization-based feedback (in, tr, va, and ch, rightmost in Figure 6) show different performance, which will be discussed later. Relevant for Hypothesis (B) is that the mean scores are above the ones of the *control* group (co). This is true for training rounds 1 and 2, for performance rounds 3 and 4, and thus also overall. The statistical significance based on a comparison between optimization-based feedback groups and the other two groups is shown in Table A.4a. Participants who received optimization-based feedback performed significantly better than those without feedback, in each round and in total, proving Hypothesis (B). Looking closer at Table A.4a one observes that this significance holds for both comparisons, the one to all participants without feedback (*highscore* group and *control* group) and only to those from *control* group. The value of the statistical test is larger in the training rounds by roughly one order of magnitude, which is not surprising given the direct benefit of the feedback on the performance.

The performance of the four optimization-based feedback groups is quite diverse, compare again Figure 6: *value* group was the best by far in all the rounds, *trend* group comes second. The two other feedback groups, *indicate* group and *chart* group, do not exhibit such a good performance. As a result, the performance of the *control* group is only significantly worse on average, but not compared to all of the single feedback groups as tested in Table A.4b. Consequently, Hypothesis (C) can be considered as disproved, both for the performance rounds as overall.

The results of the group–specific Welch's t-test in Table A.4b are also helpful for an assessment of the hypotheses of Table 4. For $\alpha = 0.05$, *value* group is sig-

Hypothesis		Confirmed
(D)	**trend group** performs best overall and best in performance rounds	—
(E)	**value group** performs best in training rounds and worst in performance rounds, compared to other feedback groups	(✓)
(F)	**indicate group** and **chart group** do not perform significantly better than control group in performance rounds	✓

Table 4. Hypotheses on specific feedback types (arrow feedback in *trend* group and toggled values in *value* group).

nificantly better than *control* group in all the rounds. *Trend* group misses significance only in round 3 by narrow margin, but exhibits significant differences in the other rounds. *Indicate* group is significantly better only in round 1. The remaining groups are not

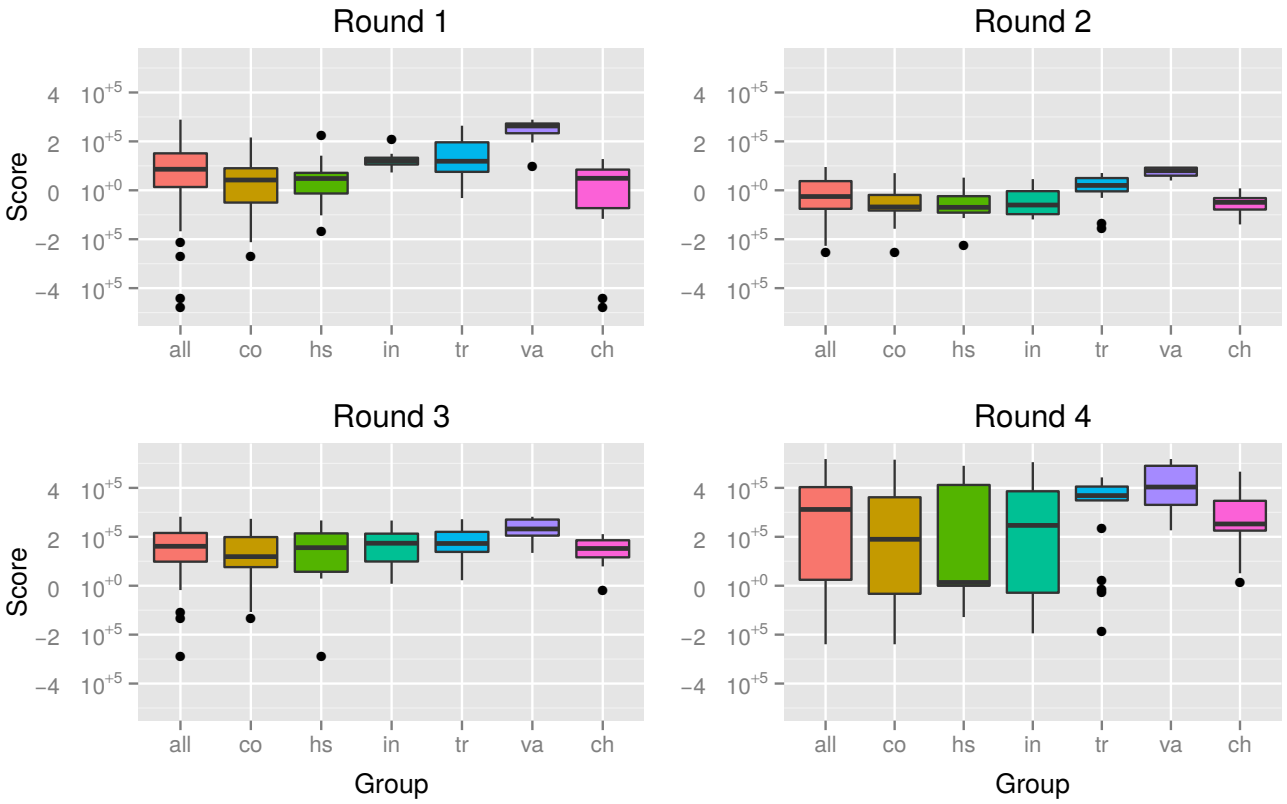


Figure 6. Score boxplot of all feedback groups (co: control, hs: highscore, in: indicate, tr: trend, va: value, ch: chart) for all rounds and all complete datasets without 6 outliers ($N = 94$). The boxplot indicates that *value* group and—except for round 3—*trend* group are better than the others.

significantly different than *control* group. The difference between *value* group and all other groups is also significant in all rounds for $\alpha = 0.05$ (not in the table).

As *value* group showed the best performance, and *trend* group only second–best, Hypothesis (D) can be considered as disproved.

It is true that *value* group performed best in training rounds, but it did not perform worst in performance rounds. So, the first statement of Hypothesis (E) is likely to be true, the second to be false.

The two other feedback groups, *indicate* group and *chart* group, do indeed not perform significantly better than *control* group, confirming Hypothesis (F).

Figure 7 contains the average *Use of Potential* for each feedback group over all rounds. This plot reveals much more detail on the performance of the different groups, as it contains also temporal information. This will be helpful in the next section. Looking at the average values (remember: *Use of Potential* is the better, the closer it is to 0), additional evidence is given for the results for Hypotheses (B–F).

Effects of optimization-based feedback on learning

As described in section"Learning", we use the gradient m obtained from a linear regression as an indicator for learning. As *Use of Potential* may hence be considered as the *learning curve*, it is worthwhile to have a look at Figure 7 to assess the first hypothesis on learning in Table 5. The visual impression is that on average the *Use of Potential* has a tendency to increase, at least

Hypothesis		Confirmed
(G)	participants learn how to solve the complex problem	✓
(H)	learning function is approximately logarithmic	—

Table 5. Hypotheses related to learning.

for rounds 1, 2, and 4. This is confirmed quantitatively by looking at the average values and the p values in Table A.6. On average, participants show significant learning effects in all rounds except for round 3. This supports the assumption that participants learn how to control the microworld, i.e., Hypothesis (G). Additional evidence for Hypothesis (G) comes from Figures 5 and 6. Comparing round 1 (with feedback) and 3 (without feedback) one can see that the distribution is shifted slightly to the right, i.e., to higher scores, hinting at an overall learning effect. That this learning effect is dependent on the feedback in the training rounds 1 and 2, can already be guessed by looking at round 4. Round 4, which is a performance round with initial values the participants have not seen before in the training rounds, exhibits a non–normal distribution of performance.

Trying to fit a logarithmic function to the *Use of Potential* was not successful. A closer inspection of Figure 7 indicates that although for certain partici-

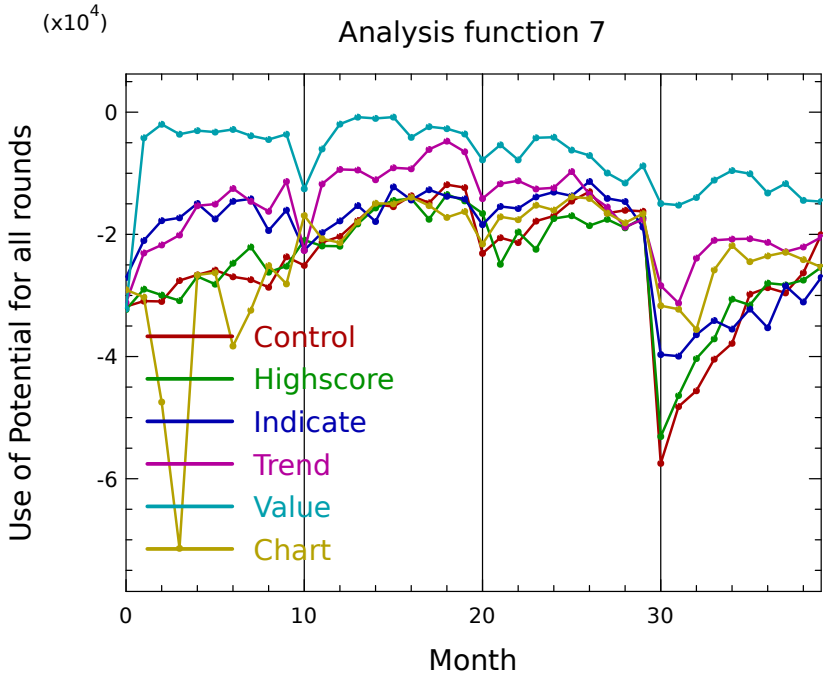


Figure 7. *Use of Potential* for all four complete datasets without 6 outliers ($N = 94$) over all rounds (one round consists of 10 months), but averaged for the six different participant groups (see section "experimental groups"). *value* group is always on top and almost constant in feedback rounds, but decreases slightly in performance rounds. All other groups show a more (*control* and *highscore* group) or less (*trend* group) severe decline at the beginning of round 4.

pant groups and rounds (e.g., *trend* group in rounds 1, 2, and 4) there is a stronger increase at the beginning that flattens toward the end of the round, the Hypothesis (H) cannot be confirmed based on our data. This is also the impression from investigating *Use of Potential* of single participants, compare Figure 9.

We now have a closer look at the effect of optimization-based feedback on learning. To test Hypothesis (I), see Table 6, we look at the regression parameters m for the four optimization-based feedback groups (of, consisting of in, tr, va, ch) and the two other groups (nof, consisting of co and hs) in Table 7. The mean for parameter m for of is higher in round 1 and lower in all other rounds. This suggests that, given the performance of these groups, optimization-based feedback groups learned faster, namely mainly in the first round. However, Welch's t-test only shows significance for rounds 2–4. We see this as an indication that (I) might be true, but it cannot be fully confirmed with our data.

Hypothesis		Confirmed
(I)	optimization-based feedback groups learn faster	(✓)
(J)	*trend* group learns fastest	✓

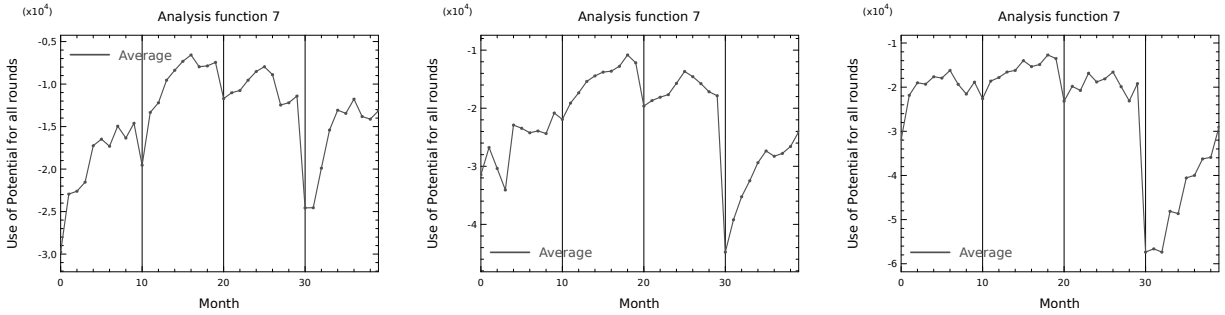
Table 6. Hypotheses related to learning, specific for participant groups.

To shed more light on the issue, we investigate the learning curves of the single participant groups. As above, Figure 6 hints at improved scores in round 3 compared to round 1 (with identical initial values) for all participant groups with the exception of *value* group. *Value* group remained static (-4%) at a higher level than the other groups. A reason for this may be that participants profited so strong from the *value* feedback during the feedback rounds that their performance without feedback slightly decreased. However, the group's mean is on a high level, so there was not much space for improvement anyhow. For the other five groups performance improved drastically (20% at least).

Rnd	nof	of	nof < of	of < nof
1	651.2	1063.1	0.2384	0.7616
2	1086.6	550.3	0.9642	**0.0358**
3	670.9	-263.4	0.9997	**0.0003**
4	3445.1	817.0	1.0000	**0.0000**

Table 7. Columns 2 and 3: mean regression parameters m for non-optimization based feedback groups (nof) and optimization-based feedback groups (of). Columns 4 and 5: corresponding significances from Welch's t-test. Rnd means Round. One observes that of learned more in round 1, however not significant, and co&hs learned significantly more in rounds 2–4.

Again, more insight comes from our novel analysis approach, the study of *Use of Potential* depicted in Figure 7. *Value* group is always on top as expected and almost constant in feedback rounds, but decreases slightly in performance rounds. This means that the performance of participants in this group is on a very high level from the beginning and hardly improves, in fact rather impairs. All other groups show a more or



(a) High **(b)** Mid **(c)** Low

Figure 8. *Use of Potential* according to *high*, *mid*, and *low* model knowledge for all complete datasets without 6 outliers ($N = 94$) over all rounds (one round consists of 10 months). Participants with low knowledge show a severe decline in their score at the beginning of round 4, whereas they stay on the same level in the rounds before. *High* and *mid* group show an increase in feedback rounds and *high* group also stays almost on the same level later in round 4. Note that the start of round 4 is challenging due to new initial values.

less severe decline at the beginning of round 4 with *control* and *highscore* group at the one end and *trend* group at the other. However, all groups except *value* group seem to improve their performance during the first three rounds.

To quantify this, Table A.5 contains the mean values for the regression parameters m of the different feedback groups. The Kolmogorov-Smirnov test results show that the mean values can be considered to be normally distributed in all rounds, except for *chart* group in round 1. The Welch's t-test results show whether the hypothesis that the mean value of m is positive, and hence a positive learning effect occured, is significant or not. *Trend* group is the only group with a significant learning effect in both rounds 1 and 2. Therefore we see Hypothesis (J) as confirmed.

For *control* group, the learning effects get significant from round 2 on, and for *highscore* group they are significant in rounds 2 and 4. The mean values in performance rounds for *control* and *highscore* group are drastically higher than for the optimization-based feedback groups. *Value* group is the only one with a significantly decreasing performance in round 3 and also the only one with an overall mean below 0. Note that *chart* group performs even worse at least in the feedback rounds. This changes in performance rounds, so one can suppose that the feedback consternated the participants. A possible reason could lie in a misinterpretation of the sensitivity information participants were given by this feedback. All other optimization-based feedback groups received direct information on the optimal solution.

Effects of model knowledge

The focus of this section are the two variables *knowledge* and *uncertainty*. We look at the hypotheses in Table 8.

To investigate Hypothesis (K), quartiles have been used to build groups of participants with *high* (best 25%), *mid* (those between first and third quartile), and *low* (worst 25%) score for each round. Means of correspondent model knowledge and uncertainty scores can be found in Tables A.9a and A.9b. *High* groups have

Hypothesis		Confirmed
(K)	well-performers know more about the model	✓
(L)	participants with high model knowledge perform well	✓
(M)	participants with high model knowledge learn more	(✓)
(N)	*trend* group has highest model knowledge and lowest uncertainty	✓

Table 8. Model knowledge related hypotheses.

the highest means which increase over the rounds. Except for round 1, *mid* groups are between *low* and *high* groups. In performance rounds, all differences are significant according to the Welch's t-test. Significance roughly increases over the rounds, which suggests that model knowledge is a crucial factor for successful control of the *IWR Tailorshop* microworld.

Concerning Hypotheses (L) and (M), participants have been merged in 3 (*low* (0/1), *mid* (2/3), and *high* (4/5)) and 2 (*low* (0/1) and *mid* (2/3)) groups respectively according to their *knowledge* and *uncertainty* score, which both are between 0 and 5. No participant achieved an *uncertainty* score of 4 or 5, thus there are only two groups for *uncertainty*. Tables A.10a and A.10b contain the mean score values of all four rounds for these groups.

For *knowledge*, the *high* group has the highest score means by far. Except for round 1, *mid* group lies between *low* and *high* group. Student's t-test in Table A.10c shows that *high* group was almost always significantly better than the two other groups. Significance increases over the rounds, which means that model knowledge becomes a better predictor for participants success the more rounds the participants played. Comparing round 1 and 3, participants with low model

knowledge could barely improve their performance, whereas the *high* group approximately doubled their score. Indeed, correlation between score and model knowledge increases from about 0.09 in round 1 to 0.48 in round 4. As a summary, we see Hypothesis (L) as confirmed.

For *uncertainty*, the *low* group has higher means in all rounds, but again the differences are much smaller than for *knowledge*. Hence, the differences between the groups are not significant. Correlation with score is about -0.2 for all rounds except the first.

Concerning Hypothesis (M), the average *Use of Potential* for the three model knowledge groups can be found in Figure 8. Participants with low knowledge show a severe decline at the beginning of round 4, whereas they stay on the same level in the rounds before. *High* and *mid* group show an increase in feedback rounds and *high* group also stays almost on the same level in round 4.

The values in Table A.12 reveal that participants with low model knowledge learned significantly less in round 1 than those with high knowledge. Again, the situation reverses in round 4. Hypothesis (M) could thus be confirmed with a restriction to round 1. However, it seems also likely that model knowledge changes from round to round and is an indicator of success in learning, rather than a predictor. Therefore a high use of potential in the training rounds could also be considered as a predictor for model knowledge at the end of the experiment. In summary, Hypothesis (M) can not be decided.

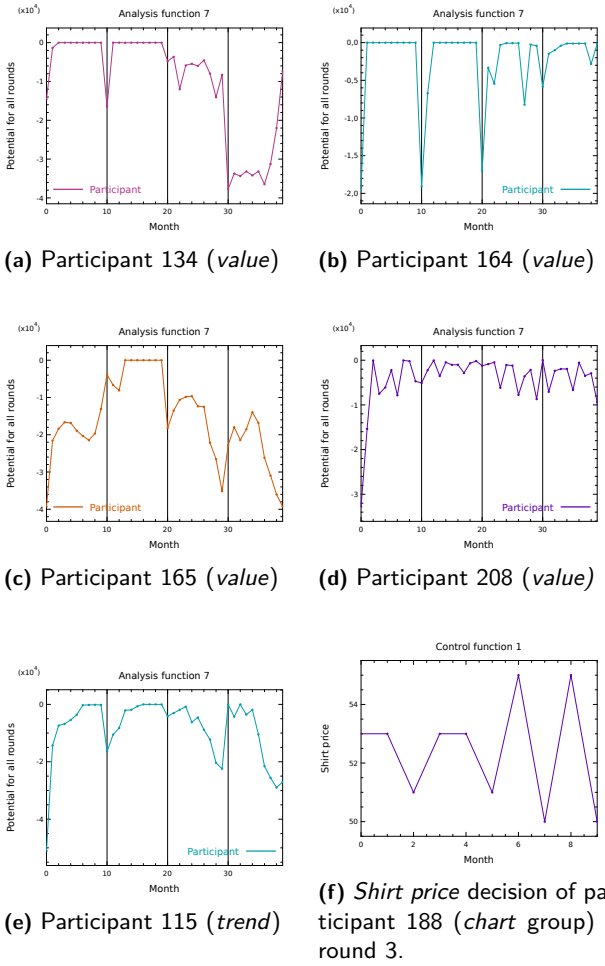
Concerning Hypothesis (N), for an analysis of differences between the groups, ratios of model knowledge and uncertainty levels and mean values are given in Table A.11. *Trend* and *value* group have the highest knowledge, but only *highscore* and *trend* group are significantly better than *control* group. *Indicate* and *chart* group have a much lower knowledge, which together with these groups' performance suggests that participants were rather confused by the optimization-based feedback.

Trend group has by far the lowest uncertainty among the groups and is the only one which has significantly lower uncertainty than *control* group. All other groups are on a similar level.

Exemplary participants

A more detailed look on single participants reveals different decision patterns. Figure 9 shows *Use of Potential* for participants 134, 164, 165, and 208 from *value* group and of participant 115 from *trend* group. Participants 134 and 164 seem to more or less copy the optimal solution in the feedback rounds. Remember that feedback for these participants consisted of the numeric values of the optimal solution. Participant 208, in contrast, seems to pursue a different strategy which is less solution-oriented.

The success in the performance rounds 3 and 4 also varies a lot: participant 164 seems to remember the solution, which is especially useful in round 3 as it started with the same value as round 1, but participant 134 does not and lacks knowledge how to control the model. Participant 165, who seems to change strategy during feedback rounds from exploration to solution-oriented, decreases in round 3, too. Participant 208, who possibly has found an own strategy, stays on the same level throughout all rounds.



(a) Participant 134 (*value*) **(b)** Participant 164 (*value*)

(c) Participant 165 (*value*) **(d)** Participant 208 (*value)*

(e) Participant 115 (*trend*) **(f)** *Shirt price* decision of participant 188 (*chart* group) in round 3.

Figure 9. *Use of Potential* for single participants from *value* group (a–d) and *trend* group (e), and exemplary shirt price decisions (f).

Participants 115 from *trend* group reaches a comparably high level of *Use of Potential* with monotonically increasing curves during the first two rounds converging to 0, i.e., coming close to optimality at the end of each round. Not surprisingly, a solution-oriented pattern like among the participants from *value* group in Figure 9 (a–d), cannot be observed due to the different type of feedback.

Figure 9f shows the *shirt price* decision of participant 188 from *chart* group. Although already in a performance round, the participant seems quite unsure about the right strategy and changes the control a lot. Such a pattern at that time point can particularly be found among the datasets from *chart* group.

Conclusion and outlook

In this work, optimization methods were used in the context of *Complex Problem Solving (CPS)* both as an

analysis tool and to provide feedback in real time for learning purposes. While first works on optimization-based analysis for *CPS* (Sager et al., 2010, 2011) had a focus on understanding how external factors influence thinking, in the work at hand, we also investigated learning effects. The use of optimization as an analysis *and* feedback tool for psychological studies is completely new to our knowledge.

We presented a variant of the *IWR Tailorshop*, a new microworld for CPS. This turn-based test-scenario yields a mixed-integer nonlinear program with non-convex relaxation and consists of functional relations based on optimization results. With the proof of feasibility for the *IWR Tailorshop* in this article, we intend to start a new era beyond *trial-and-error* in the definition of microworlds for analyzing human decision making.

In our web-based feedback study with 148 participants, we used the *IWR Tailorshop* microworld to investigate the effects of optimization-based feedback. Optimization-based feedback could significantly improve participants' performance in the *IWR Tailorshop* microworld if the presentation was chosen appropriately. In our study, *value* group performed significantly better than all other groups.

We could show that such a feedback can significantly improve participants' performance in a complex microworld and for some kinds of feedback, the difference to *control* group was huge. However, it also became apparent that the representation of feedback is important. Feedback based on a kind of sensitivity information seemed to rather confuse participants in this study, which was also suggested by our optimization-based analysis.

The best-performing group was the *value* group which received the most precise information about the optimal solution. Knowledge about the model was better amongst another well-performing group, the *trend* group. Since we could show that model knowledge is a predictor for performance, perhaps these participants would have outperformed the others on a longer timescale. More data is needed to verify this hypothesis, though.

Optimization-based analysis could show that participants learn to control the model over time by an analysis of *Use of Potential*. Different aspects of the analysis indicate that for a high performance, learning during the first round is crucial. It turned out that the best way to enforce learning at the beginning was by *trend* feedback. Through the optimization-based analysis, we were also able to show that there were no systematic differences between the groups at the beginning and that initial performance was not relevant for performance at the end of the time scale. For some of the hypotheses, however, significance could not or only partly be shown. In these cases, more data and investigation will be necessary.

The main intention of this paper is to present the optimization-based feedback and to show their usefulness in a feedback situation. The test of (learning) theories was not the focus. Our different hypotheses are not drawing on specific literature but are kind of "informed guesses" about what might happen. This is also due to the fact that there exist no reference studies with the Tailorshop in a feedback setting that could be used as a baseline for expected effects. However, coupling our approach to theoretically based hypotheses on learning seems a promising line of future research.

Another interesting research direction could be if the widely spread assumption that positive feedback increases performance is true. In Barth and Funke (2010) it has been shown that negative feedback impairs performance. However, it is unclear if this is also true in the long run. From former studies we know that positive and negative feedback lead to different processing styles. Therefore one could expect that a quotient of positive and negative feedback (*carrot and stick*) impairs performance the most. 40% positive feedback and 60% negative feedback might lead to the best performance, for instance.

Finally, the parameter set used for the computations of the *IWR Tailorshop* microworld in this work has been set up manually to achieve a reasonable model behavior. Here we still see high potential for improvement. One could use derivative-free optimization methods to optimize the *parameter values* such that two (or even more) previously defined strategies (e.g., a high and a low price strategy) yield a similar objective value. By that, participants could follow different strategies and perform quite well in all of them if decisions are made appropriate.

Acknowledgements: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 647573) and from the German BMBF under grant 05M2013 - GOSSIP. The authors gratefully acknowledge this.

Declaration of conflicting interests: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

Author contributions: The main contribution is due to the first author (ME) who performed the study as part of his PhD thesis (Engelhart, 2015). SS and JF helped in designing and analysing the study and did part of the writing.

Supplementary material: Supplementary material available online.

Handling editor: Andreas Fischer

Citation: Engelhart, M., Funke, J. ,& Sager, S. (2017). A web-based feedback study on optimization-based training and analysis of human decision making. *Journal of Dynamic Decision Making, 3*, 2. doi:10.11588/jddm.2017.1.34608

Received: 04 January 2017
Accepted: 17 April 2017
Published: 26 May 2017

References

Barth, C. M. (2010). *The impact of emotions on complex problem solving performance and ways of measuring this performance* (Unpublished doctoral dissertation). Ruprecht–Karls–Universität Heidelberg.

Barth, C. M., & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion*, *24*(7), 1259–1268. doi: 10.1080/02699930903223766

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society Series A*, *160*(901), 268–282. doi: 10.1098/rspa.1937.0109

Brehmer, B. (1995). Feedback delays in dynamic decision making. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 103–130). Hillsdale, NJ: Erlbaum.

Brown, M., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367. doi: 10.1080/01621459.1974.10482955

Cronin, M. A., Gonzalez, C., & Sterman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes*, *108*(1), 116–130. doi: 10.1016/j.obhdp.2008.03.003

Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (2011). Beyond IQ. a latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence*, *39*(5), 323–334. doi: 10.1016/j.intell.2011.06.004

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 471–440. doi: 10.1146/annurev.ps.41.020190.002221

Dörner, D. (1980). On the difficulties people have in dealing with complexity. *Simulation and Games*, *11*(1), 87–106. doi: 10.1177/104687818001100108

Engelhart, M. (2015). *Optimization-based analysis and training of human decision making* (Unpublished doctoral dissertation). Ruprecht-Karls-Universität Heidelberg.

Engelhart, M., Funke, J., & Sager, S. (2013). A decomposition approach for a new test-scenario in complex problem solving. *Journal of Computational Science*, *4*(4), 245–254. doi: 10.1016/j.jocs.2012.06.005

Frensch, P. A., & Funke, J. (1995). *Complex problem solving: The european perspective*. Taylor & Francis. doi: 10.4324/9781315806723

Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? *Diagnostica*, *29*(4), 283–302. doi: 10.11588/heidok.00008131

Funke, J. (2003). *Problemlösendes Denken*. Stuttgart, Germany: Kohlhammer.

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, *11*(2), 133–142. doi: 10.1007/s10339-009-0345-0

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective – 10 years after. In D. H. Jonassen (Eds.), *Learning to solve complex scientific problems* (pp. 25–47). New York: Erlbaum.

Gonzalez, C. (2004). Learning to make decisions in dynamic environments: Effects of time constraints and cognitive abilities. *Human Factors*, *46*(3), 449–460. doi: 10.1518/hfes.46.3.449.50395

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, *21*(1), 27–58. doi: 10.1214/aoms/1177729885

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1–21. doi: 10.1080/00401706.1969.10490657

Hörmann, H. J., & Thomas, M. (1989). Zum Zusammenhang zwischen Intelligenz und komplexem Problemlösen. *Sprache & Kognition*, *8*(1), 23–31.

Hüfner, M., Tometzki, T., Kraja, T., & Engell, S. (2011). Learn2Control: Eine webbasierte Lernumgebung im Bio- und Chemieingenieurwesen. *Journal Hochschuldidaktik*, *22*(1), 20–23.

Kleinmann, M., & Strauß, B. (1998). Validity and applications of computer simulated scenarios in personal assessment. *International Journal of Seclection and Assessment*, *6*(2), 97–106. doi: 10.1111/1468-2389.00078

Kluwe, R. H. (1993). Knowledge and performance in complex problem solving. *Advances in Psychology*, *Volume 101*, 401–423. Amsterdam, Netherland: Elsevier. doi: 10.1016/s0166-4115(08)62668-0

Kluwe, R. H., Misiak, C., & Haider, H. (1991). The control of complex systems and performance in intelligence tests. In H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 227–244). Hillsdale, NJ: Erlbaum.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292). Stanford, CA: Stanford University Press.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402. doi: 10.1080/01621459.1967.10482916

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 3–30. doi: 10.1145/272991.272995

Meyer, B., & Scholl, W. (2009). Complex problem solving after unstructured discussion. Effects of information distribution and experiece. *Group Process and Intergroup Relations*, *12*(4), 495–515. doi: 10.1177/1368430209105045

Osman, M. (2008). Observation can be as effective as action in problem solving. *Cognitive Science*, *32*(1), 162–183. doi: 10.1080/03640210701703683

Otto, J. H., & Lantermann, E.-D. (2004). Wahrgenommene Beeinflussbarkeit von negativen Emotionen, Stimmung und komplexes Problemlösen. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *25*(1), 31–46. doi: 10.1024/0170-1789.25.1.31

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg. *Zeitschrift für Psychologie*, *189*(1), 79–100.

Putz-Osterloh, W., Bott, B., & Köster, K. (1990). Models of learning in problem solving – are they transferable to tutorial systems? *Computers in Human Behavior*, *6*(1), 83–96. doi: 10.1016/0747-5632(90)90032-c

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. doi: 10.1016/j.jrp.2006.02.001

Robbins, T. W., Anderson, E. J., Barker, D. R., Bradley, A. C., Fearnyhough, C., Henson, R., Hudson, S. R., & Baddeley, A. D. (1996). Working memory in chess. *Memory & Cognition*, *24*(1), 83–93. doi: 10.3758/bf03197274

Sager, S., Barth, C. M., Diedam, H., Engelhart, M., & Funke, J. (2010). Optimization to measure performance in the Tailorshop test scenario — structured MINLPs and beyond. In *Proceedings EWMINLP10* (pp. 261–269). CIRM, Marseille, France.

Sager, S., Barth, C. M., Diedam, H., Engelhart, M., & Funke, J. (2011). Optimization as an analysis tool for human complex problem solving. *SIAM Journal on Optimization*, *21*(3), 936–959. doi: 10.1137/11082018x

Sawilowsky, S. S., & Blair, C. R. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin*, *111*(2), 352–360. doi: 10.1037/0033-2909.111.2.352

Selten, R., Pittnauer, S., & Hohnisch, M. (2012). Dealing with dynamic decision problems when knowledge of the environment is limited: An approach based on goal systems. *Journal of Behavioral Decision Making*, *25*, 443–457. doi: 10.1002/bdm.738

Süß, H.-M., Oberauer, K., & Kersting, M. (1993). Intellektuelle Fähigkeiten und die Steuerung komplexer Systeme. *Sprache & Kognition*, *12*, 83–97.

Tukey, J. W. (1977). *Exploratory data analysis*. Boston, MA: Addison-Wesley.

Wenke, D., & Frensch, P. A. (2003). Is success or failure at solving complex problems related to intellectual ability? In J. Davidson & R. Sternberg (Eds.), *The psychology of problem solving* (pp. 87–126). Cambridge, England: Cambridge University Press. doi: 10.1017/cbo9780511615771.004

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science*, *21*(4), 393–409. doi: 10.1002/sres.653

Appendix

The mathematical model for the IWR Tailorshop consists of the following set of equations, for $k = t_0, \ldots, t_f$, shown in Equations (A.1a) to (A.1l).

$$x_{k+1}^{EM} = x_k^{EM} + u_k^{EM} \tag{A.1a}$$

$$x_{k+1}^{PS} = x_k^{PS} - u_k^{dPS} + u_k^{DPS} \tag{A.1b}$$

$$x_{k+1}^{DS} = x_k^{DS} - u_k^{dDS} + u_k^{DDS} \tag{A.1c}$$

$$\begin{aligned} x_{k+1}^{DE} = {} & p^{DE,0} \cdot \exp\left(-p^{DE,1} \cdot u_k^{SP}\right) \\ & \cdot \log\left(p^{DE,2} \cdot u_k^{AD} + 1\right) \\ & \cdot \left(x_k^{RE} + p^{DE,3}\right) \end{aligned} \tag{A.1d}$$

$$\begin{aligned} x_{k+1}^{RE} = {} & p^{RE,0} \cdot x_k^{RE} \\ & + p^{RE,1} \log\Big(\left(p^{RE,2} \cdot u_k^{AD}\right. \\ & + p^{RE,3} \cdot u_k^{SP} \cdot (x_k^{SQ})^2 \\ & \left. + p^{RE,4} \cdot u_k^{WA}\right) \cdot p^{RE,5}\Big) \end{aligned} \tag{A.1e}$$

$$\begin{aligned} x_{k+1}^{PR} = {} & p^{PR,0} \cdot x_{k+1}^{PS} \\ & \cdot \log\left(\frac{p^{PR,1} \cdot x_{k+1}^{EM}}{x_{k+1}^{PS} + x_{k+1}^{DS} + p^{PR,2}} + 1\right) \end{aligned} \tag{A.1f}$$

$$\begin{aligned} x_{k+1}^{SA} = \min\Bigg\{ & p^{SA,0} \cdot x_{k+1}^{DS} \\ & \cdot \log\left(\frac{p^{SA,1} \cdot x_{k+1}^{EM}}{x_{k+1}^{PS} + x_{k+1}^{DS} + p^{SA,2}} + 1\right); \\ & x_k^{SH} + x_{k+1}^{PR}; p^{SA,3} \cdot x_{k+1}^{DE} \Bigg\} \end{aligned} \tag{A.1g}$$

$$x_{k+1}^{SH} = x_k^{SH} - x_{k+1}^{SA} + x_{k+1}^{PR} \tag{A.1h}$$

$$\begin{aligned} x_{k+1}^{SQ} = {} & p^{SQ,0} \cdot x_k^{MO} + p^{SQ,1} \cdot x_k^{MQ} + \\ & p^{SQ,2} \cdot u_k^{RQ} \end{aligned} \tag{A.1i}$$

$$\begin{aligned} x_{k+1}^{MQ} = {} & x_k^{MQ} \cdot p^{MQ,0} \\ & \cdot \exp\left(-p^{MQ,1} \frac{x_k^{PR}}{x_k^{PS} + p^{MQ,2}}\right) \\ & + p^{MQ,3} \cdot \log\left(u_k^{MA} \cdot p^{MQ,4} + 1\right) \end{aligned} \tag{A.1j}$$

$$\begin{aligned} x_{k+1}^{MO} = {} & \left(1 - p^{MO,0}\right) \cdot x_k^{MO} + p^{MO,0} \\ & \cdot \log\Big(p^{MO,1} \cdot (u_k^{EM} + p^{dEM}) \\ & + p^{MO,2} \cdot u_k^{DPS} + p^{MO,3} \cdot u_k^{DDS} \\ & + p^{MO,4} \cdot u_k^{WA} + p^{MO,5} \cdot x_k^{RE} \\ & + p^{MO,6}\Big) \cdot \exp\Big(-\left(p^{MO,7} \cdot u_k^{dPS}\right. \\ & \left. + p^{MO,8} \cdot u_k^{dDS}\right) + p^{MO,9}\Big) \\ & \cdot p^{MO,10} \end{aligned} \tag{A.1k}$$

$$\begin{aligned} x_{k+1}^{CA} = {} & p^{CA,0} \cdot \Bigg(x_k^{CA} + \left(x_{k+1}^{SA} \cdot u_k^{SP}\right) \\ & + \left(u_k^{dPS} \cdot p^{CA,1}\right) + \left(u_k^{dDS} \cdot p^{CA,2}\right) \\ & - \left(x_{k+1}^{EM} \cdot u_k^{WA}\right) \\ & - \left(x_{k+1}^{PR} \cdot u_k^{RQ} \cdot p^{CA,3}\right) \\ & - \left(x_k^{PS} \cdot p^{CA,4}\right) - \left(x_k^{DS} \cdot p^{CA,5}\right) \\ & - u_k^{MA} - u_k^{AD} - \left(x_{k+1}^{SH} \cdot p^{CA,6}\right) \\ & - \left(u_k^{DPS} \cdot p^{CA,7}\right) \\ & - \left(u_k^{DDS} \cdot p^{CA,8}\right)\Bigg) \end{aligned} \tag{A.1l}$$

Additional constraints are given by the inequalities shown in equations (A.2a) to (A.2e),

$$u_k^{dPS} + u_{k-1}^{dPS} \leq p^{dPS}, \tag{A.2a}$$

$$p^{DEM,0} \cdot x_k^{PS} + p^{DEM,1} \cdot x_k^{DS} \geq u_k^{EM}, \tag{A.2b}$$

$$x_k^{EM}, x_k^{PS}, x_k^{DS} \geq 1, \tag{A.2c}$$

$$x_k^{SH}, x_k^{PR}, x_k^{SA}, x_k^{DE} \geq 0, \tag{A.2d}$$

$$x_k^{RE}, x_k^{SQ}, x_k^{MQ}, x_k^{MO} \geq 0, \tag{A.2e}$$

and the simple bounds on the controls (A.3a) to (A.3j),

$$u_k^{SP} \in [35\,\text{M.U.}, 55\,\text{M.U.}], \tag{A.3a}$$

$$u_k^{AD} \in [1000\,\text{M.U.}, 2000\,\text{M.U.}], \tag{A.3b}$$

$$u_k^{WA} \in [1000\,\text{M.U.}, 2000\,\text{M.U.}], \tag{A.3c}$$

$$u_k^{MA} \in [10\,\text{M.U.}, 5000\,\text{M.U.}], \tag{A.3d}$$

$$u_k^{RQ} \in \left\{p^{RQ,1}, p^{RQ,2}\right\}, \tag{A.3e}$$

$$u_k^{EM} \in \left[-p^{dEM}, \infty\right] \cap \mathbb{Z}_+, \tag{A.3f}$$

$$u_k^{DPS} \in \left[0, p^{DPS}\right] \cap \mathbb{Z}_+, \tag{A.3g}$$

$$u_k^{dPS} \in \left[0, \infty\right] \cap \mathbb{Z}_+, \tag{A.3h}$$

$$u_k^{DDS} \in \left[0, p^{DDS}\right] \cap \mathbb{Z}_+, \tag{A.3i}$$

$$u_k^{dDS} \in \left[0, p^{dDS}\right] \cap \mathbb{Z}_+. \tag{A.3j}$$

We use the objective function

$$\max_{x,u,p} \quad x_{t_f}^{CA}, \tag{A.4}$$

i.e., maximizing the capital at the end.

Of course, the set of parameters has a significant influence on the model behavior. One could, e.g., think of applying derivative-free optimization methods with a subset of the parameters to determine an appropriate parameter set for a microworld like *IWR Tailorshop*. For this work, however, we set up a parameter set manually such that the model fulfills a certain desired behavior. The chosen parameters also yield a model behavior that makes sense for the optimization, i.e., there are feasible solutions and the optimization problem is not unbounded. The parameter values used throughout this work unless otherwise stated are listed in Tables A.1 and A.2.

Parameter	Value
$p^{DE,0}$	$2200.0\,\text{shirts}$
$p^{DE,1}$	$2 \cdot 10^{-2}\,\text{shirts}/\text{MU}$
$p^{DE,2}$	$2 \cdot 10^{-2}\,1/\text{MU}$
$p^{DE,3}$	0.5
$p^{RE,0}$	0.5
$p^{RE,1}$	0.672
$p^{RE,2}$	$2.5 \cdot 10^{-5}\,1/\text{MU}$
$p^{RE,3}$	$10^{-4}\,\text{shirts}/\text{MU}$
$p^{RE,4}$	$6 \cdot 10^{-5}\,\text{persons}/\text{MU}$
$p^{RE,5}$	12.0
$p^{PR,0}$	$99.9\,\text{shirts}/\text{sites}$
$p^{PR,1}$	$2.0\,\text{sites}/\text{persons}$
$p^{PR,2}$	$10^{-6}\,\text{sites}$
$p^{SA,0}$	$99.9\,\text{shirts}/\text{sites}$
$p^{SA,1}$	$2.0\,\text{sites}/\text{persons}$
$p^{SA,2}$	$10^{-6}\,\text{sites}$
$p^{SA,3}$	1.0
$p^{SQ,0}$	0.2
$p^{SQ,1}$	0.3
$p^{SQ,2}$	0.5
$p^{MQ,0}$	0.8
$p^{MQ,1}$	$6 \cdot 10^{-3}\,\text{sites}/\text{shirts}$
$p^{MQ,2}$	$10^{-6}\,\text{sites}$
$p^{MQ,3}$	0.13
$p^{MQ,4}$	$0.2\,\text{MU}^{-1}$
$p^{MO,0}$	0.5
$p^{MO,1}$	$2 \cdot 10^{-2}\,\text{persons}^{-1}$
$p^{MO,2}$	$0.5\,\text{sites}^{-1}$
$p^{MO,3}$	$0.25\,\text{sites}^{-1}$
$p^{MO,4}$	$2.0 \cdot 10^{-4}\,\text{persons}/\text{MU}$
$p^{MO,5}$	0.3
$p^{MO,6}$	1.0
$p^{MO,7}$	$2.5\,\text{sites}^{-1}$
$p^{MO,8}$	$2.0\,\text{sites}^{-1}$
$p^{MO,9}$	1.0
$p^{MO,10}$	0.5
$p^{CA,0}$	1.03
$p^{CA,1}$	$5000\,\text{MU}/\text{site}$
$p^{CA,2}$	$3500\,\text{MU}/\text{site}$
$p^{CA,3}$	$5.0\,\text{MU}/\text{shirt}$
$p^{CA,4}$	$1000\,\text{MU}/\text{site}$
$p^{CA,5}$	$700\,\text{MU}/\text{site}$
$p^{CA,6}$	$1.5\,\text{MU}/\text{shirt}$
$p^{CA,7}$	$10000\,\text{MU}/\text{site}$
$p^{CA,8}$	$7000\,\text{MU}/\text{site}$

Table A.1. Parameter set for states used with *IWR Tailorshop*. *MU* means monetary units.

Parameter	Value
n^{RQ}	2
$p^{RQ,1}$	0.5
$p^{RQ,2}$	1.0
$p^{DEM,0}$	$5\,\text{persons}/\text{site}$
$p^{DEM,1}$	$10\,\text{persons}/\text{site}$
p^{dEM}	$10\,\text{persons}$
p^{DPS}	$1\,\text{site}$
p^{dPS}	$1\,\text{site}$
p^{DDS}	$2\,\text{sites}$
p^{dDS}	$1\,\text{site}$

Table A.2. Parameter set for controls used with *IWR Tailorshop*.

Variable		Round 1	Round 2	Round 3	Round 4
Employees	x_0^{EM}	14	3	14	42
Production sites	x_0^{PS}	1	1	1	2
Distribution sites	x_0^{DS}	1	5	1	7
Shirts in stock	x_0^{SH}	319	0	319	0
Production	x_0^{PR}	270	69	270	467
Sales	x_0^{SA}	270	69	270	467
Demand	x_0^{DE}	3877	2399	3877	3065
Reputation	x_0^{RE}	0.7934	0.1805	0.7934	0.4711
Shirts quality	x_0^{SQ}	0.7500	0.6558	0.7500	0.8136
Machine quality	x_0^{MQ}	0.8125	0.9998	0.8125	0.7712
Motivation of employees	x_0^{MO}	0.7403	0.4032	0.7403	0.5108
Capital	x_0^{CA}	175226	28075	175226	323907
Shirt price	u_0^{SP}	50	39	50	42
Advertising	u_0^{AD}	2000	1599	2000	1337
Wages	u_0^{WA}	1500	1750	1500	1451
Maintenance	u_0^{MA}	500	3000	500	267
Resources quality	u_0^{RQ}	2	1	2	2
Recruit employees	u_0^{DEM}	0	0	0	0
Dismiss employees	u_0^{dEM}	0	0	0	0
Create production site	u_0^{DPS}	0	0	0	0
Close production site	u_0^{dPS}	0	0	0	0
Create distribution site	u_0^{DDS}	0	0	0	0
Close distribution site	u_0^{dDS}	0	0	0	0

Table A.3. Initial values for each round used in *IWR Tailorshop* feedback study. Note that values for controls (lower part) were only preset values and could still be changed by the participant. The last six controls, starting from *recruit employees*, were always set to the value in the table after each month to avoid accidental recruitment and dismissal as well as site creation and closing. Round 1 and 3 had the same initial values.

Round	Means			t test	
	co&hs	control	of	co&hs < of	control < of
1	25869.2	24274.1	112042.8	**0.0009**	**0.0020**
2	-58869.2	-57289.0	-1174.4	**0.0000**	**0.0003**
3	124185.3	128502.7	172860.8	**0.0091**	**0.0182**
4	170923.2	166039.1	293403.4	**0.0029**	**0.0059**
Sum	262108.6	261526.8	577132.7	**0.0000**	**0.0002**

(a) Welch's t-test p-values of comparison of score means for each round between *control* and *highscore* groups (co, hs) on the one side and groups with optimization-based feedback (of) on the other side with all complete datasets without 6 outliers ($N = 94$). With $\alpha = 0.05$, optimization-based feedback groups were significantly better than those without (co&hs as well as co alone).

Round	Highscore	Indicate	Trend	Value	Chart
1	0.4429	**0.0001**	**0.0005**	**0.0000**	0.8531
2	0.5891	0.3804	**0.0002**	**0.0000**	0.5414
3	0.6216	0.2168	0.0507	**0.0000**	0.3622
4	0.4200	0.4037	**0.0133**	**0.0000**	0.0577
Sum	0.4947	0.1539	**0.0007**	**0.0000**	0.3935

(b) Welch's t-test p-values of comparison of score means for each round to *control* group with all complete datasets without 6 outliers ($N = 94$). Alternative hypothesis was that mean of *control* group is lower. With $\alpha = 0.05$, only *value* group is significantly better than *control* group in all rounds. However, *trend* group misses significance only in round 3 by narrow margin.

	control	highscore	indicate	trend	value	chart
Mean	-31807.3	-32308.6	-27065.5	-31202.2	-32194.4	-29073.8
KS test	0.2192	0.6468	0.5051	1.0000	0.6880	0.9652
t-test	—	0.8988	0.1455	0.8231	0.9335	0.4110

(c) Comparison of *Use of Potential* by feedback groups in first month for all complete datasets without 6 outliers ($N = 94$): no significant differences between groups. Values can be considered to be normally distributed.

Table A.4. Different statistical tests. Bold valus of the test statistics indicate significance ($\alpha = 0.05$).

Round	control	highscore	indicate	trend	value	chart
1	599.1	767.5	359.9	1286.5	-91.3	2366.5
2	1140.9	965.4	714.2	725.0	22.2	610.7
3	814.4	350.8	104.6	-616.5	-448.5	294.7
4	3717.4	2837.5	1304.5	847.9	78.0	1097.9
Feedback rounds sum	1740.0	1732.9	1074.1	2011.5	-69.1	2977.2
Performance rounds sum	4531.8	3188.3	1409.2	231.4	-370.5	1392.6
Total sum	6271.8	4921.2	2483.2	2242.9	-439.6	4369.9

(a) Means

Round	control	high-score	indicate	trend	value	chart
1	0.1551	0.2901	0.7662	0.4528	0.0748	**0.0493**
2	0.5016	0.9603	0.9348	0.4203	0.6070	0.6826
3	0.8186	0.9434	0.7300	0.7786	0.4601	0.9627
4	0.9961	0.8713	0.8615	0.9498	0.9832	0.6299

(b) Kolmogorov-Smirnov test

Round	control	high-score	indicate	trend	value	chart
1	0.1051	0.1820	0.2194	**0.0036**	0.6708	0.0960
2	**0.0002**	**0.0248**	0.1263	**0.0045**	0.4718	0.0787
3	**0.0002**	0.1528	0.3853	0.9399	0.9646	0.2284
4	**0.0000**	**0.0016**	**0.0435**	0.1053	0.4542	0.0858

(c) Welch's t-test for $\mu>0$

Round	control	high-score	indicate	trend	value	chart
1	0.8949	0.8180	0.7806	0.9999	0.3292	0.9040
2	0.9998	0.9752	0.8737	1.0000	0.5282	0.9213
3	0.9998	0.8472	0.6147	0.0601	**0.0354**	0.7716
4	1.0000	0.9984	0.9565	0.8947	0.5458	0.9142

(d) Welch's t-test for $\mu<0$

Table A.5. Parameter m by feedback groups for all complete datasets without 6 outliers ($N = 94$): means, Welch's t-test, and Kolmogorov-Smirnov test. The values of all groups can be considered to be normally distributed in all rounds except for *chart* group in round 1. *trend* group is the only group with a significant learning effect in the first two rounds, *value* group the only one with a significantly decreasing performance in round 3. Bold valus of the test statistics indicate significance ($\alpha = 0.05$).

Round	Mean	t-Test $\mu > 0$
1	879.1	**0.0016**
2	789.9	**0.0000**
3	154.0	0.1365
4	1991.2	**0.0000**

Table A.6. Regression m for all complete datasets without 6 outliers ($N = 94$): means and Welch's t-test results ($\alpha = 0.05$). Participants show significant learning effects in all rounds except for round 3, in which especially *value* group is significantly < 0.

Round	Low	Mid	High
1	305.2	924.5	1365.9
2	1439.3	637.2	433.2
3	549.4	148.2	-230.2
4	4409.5	1888.7	-230.4
Feedback	1744.4	1561.7	1799.2
Performance	4958.9	2036.9	-460.7
Sum	6703.3	3598.6	1338.5

Table A.7. Means for regression m according to performance in performance rounds (*low*: below lower quartile, *mid*: between lower and higher quartile, *high*: above higher quartile) for all complete datasets without 6 outliers ($N = 94$): high performers have the highest mean for m in the first round and the lowest in all other rounds.

Claim	Answer	Correct	Wrong	Don't know
Motivation of employees plays an important role.	false	56%	28%	16%
Maintenance is an important intervention possibility.	false	55%	26%	19%
The higher the shirt price is, the lower is the demand.	true	41%	45%	14%
Opening and Closing sites are important intervention possibilities.	true	90%	3%	6%
It is wise to dismiss employees at the end.	true	31%	33%	36%

Table A.8. Survey on model properties at the end of task. The participants were told that *"We would like to ask you a few questions once again. Your answers will help us very much and it only takes two minutes. [...] Please decide if the following propositions are correct or wrong according to your experience from all four rounds."* Participants could always choose between *true*, *false*, and *don't know*. The content of the five items can be found in the *claim* column, the correct *answer* is shown in the corresponding column. The remaining columns show the ratio of *correct*, *wrong* and *don't know* answers among all participants. Differences to 100% are due to rounding.

Round	High Score	Mid Score	Low Score	High > Low	High > Mid	Mid > Low
1	3.17	2.50	2.79	0.1477	**0.0205**	0.8417
2	3.42	2.65	2.25	**0.0004**	**0.0063**	0.0770
3	3.46	2.74	2.04	**0.0000**	**0.0061**	**0.0068**
4	3.50	2.70	2.08	**0.0000**	**0.0023**	**0.0142**
Sum	3.33	2.80	2.04	**0.0001**	**0.0384**	**0.0035**

(a) Means of model knowledge for participants with high (i.e., best 25%), mid (between 1st and 3rd quartile), and low (i.e., worst 25%) score in the corresponding round with all complete datasets without 6 outliers ($N = 94$). Pairwise comparison of means by Welch's t-test with $\alpha = 0.05$ shows, that high scorers know significantly more about the model than mid or low scorers.

Round	High Score	Mid Score	Low Score	High > Low	High > Mid	Mid > Low
1	0.75	1.07	0.79	0.4376	0.0909	0.8716
2	0.58	1.07	0.96	0.0711	**0.0181**	0.6733
3	0.67	0.87	1.25	**0.0097**	0.1667	0.0633
4	0.71	0.93	1.08	0.0820	0.1612	0.2740
Sum	0.67	0.98	1.04	0.0696	0.0930	0.3924

(b) Means of model uncertainty for participants with high (i.e., best 25%), mid (between 1st and 3rd quartile), and low (i.e., worst 25%) score in the corresponding round with all complete datasets without 6 outliers ($N = 94$). Uncertainty means are lower for high scorers. Pairwise comparison of means by Welch's t-test with $\alpha = 0.05$ barely shows significance, however.

Table A.9. Different tests for model uncertainty and model knowledge. Bold valus of the test statistics indicate significance ($\alpha = 0.05$).

Round	Low (0/1)	Mid (2/3)	High (4/5)
1	101085.2	42407.9	110944.2
2	-51748.1	-40915.9	10319.9
3	108448.1	135943.2	200281.3
4	80163.6	214925.1	366269.3

(a) Mean score values for different levels of model knowledge

Round	Low (0/1)	Mid (2/3)
1	69516.5	86706.5
2	-22096.4	-42847.0
3	159269.1	124416.9
4	259346.6	171036.4

(b) Mean score values for different levels of model uncertainty

R	Low < High	Low < Mid	Mid < High
1	0.3740	0.9743	**0.0188**
2	**0.0005**	0.2657	**0.0010**
3	**0.0004**	0.1447	**0.0007**
4	**0.0001**	**0.0223**	**0.0001**

(c) Student's t-test p-values for model knowledge

Round	Mid < Low
1	0.7335
2	0.1221
3	0.1020
4	0.0626

(d) Student's t-test p-values for model uncertainty

Table A.10. Scores for different model knowledge and uncertainty levels (R: round) with all complete datasets without 6 outliers ($N = 94$). With $\alpha = 0.05$, participants with high model knowledge have achieved a significantly better score in almost all rounds. For model uncertainty, no significant score differences have been observed.

Property		co	hi	in	tr	va	ch	All
Knowledge	low	24%	8%	44%	5%	18%	9%	17%
	mid	59%	54%	33%	48%	36%	73%	52%
	high	17%	38%	22%	48%	45%	18%	31%
	mean	2.38	3.00	2.22	3.19	3.09	2.64	2.74
	t-test	—	**0.0451**	0.6241	**0.0113**	0.0824	0.2377	—
Uncertainty	low	72%	69%	67%	95%	82%	64%	77%
	high	28%	31%	33%	5%	18%	36%	23%
	mean	1.03	1.15	1.22	0.38	0.91	1.09	0.91
	t-test	—	0.6525	0.6630	**0.0017**	0.3545	0.5545	—

Table A.11. Ratio of model knowledge and uncertainty levels for all feedback groups (co: control, hs: highscore, in: indicate, tr: trend, va: value, ch: chart) with all complete datasets without 6 outliers ($N = 94$). Mean refers to mean uncertainty and knowledge per group. Alternative hypothesis for Welch's t-test was that mean of control group is lower (knowledge) or higher (uncertainty) respectively. For $\alpha = 0.05$, only *trend* group is significantly better in both knowledge and uncertainty. Differences to 100% are due to rounding.

Round	Low	Mid	High
1	93.3	1013.0	1086.4
2	666.2	888.5	691.6
3	111.0	301.0	-70.5
4	3257.4	1979.4	1312.6

(a) Means for Regression m

Low < High	Mid < High	Low < Mid
0.0207	0.4518	0.0686
0.4788	0.7564	0.3262
0.6712	0.8680	0.3020
0.9828	0.8480	0.9291

(b) Welch's t-test

Table A.12. Regression m according to model knowledge (*low*: below lower quartile, *mid*: between lower and higher quartile, *high*: above higher quartile) for all complete datasets without 6 outliers ($N = 94$): those with low model knowledge learned less in the first round, and more in the last round. In comparison with *high* group, this is significant.