Autonomous E-Coaching in the Wild:

Empirical Validation of a Model-Based Reasoning System

Bart A. Kamphorst
Utrecht University
Dept. of Philosophy and Religious Studies
Janskerkhof 13A, 3512 BL
Utrecht, the Netherlands
b.a.kamphorst@uu.nl

Michel C. A. Klein
VU University Amsterdam
Dept. of Computer Science
De Boelelaan 1085, 1081 HV
Amsterdam, the Netherlands
michel.klein@cs.vu.nl

Arlette van Wissen
VU University Amsterdam
Dept. of Computer Science
De Boelelaan 1085, 1081 HV
Amsterdam, the Netherlands
a.van.wissen@vu.nl

ABSTRACT

Autonomous e-coaching systems have the potential to improve people's health behaviors on a large scale. The intelligent behavior change support system eMate exploits a model of the human agent to support individuals in adopting a healthy lifestyle. The system attempts to identify the causes of a person's non-adherence by reasoning over a computational model (COMBI) that is based on established psychological theories of behavior change. The present work presents an extensive, monthlong empirical validation study (N=82) of eMate in which participants were coached in their everyday life — using a mobile app and a website — towards taking the stairs more often. The eMate reasoning mechanism is evaluated on its accuracy and its ability to promote behavior change. Results show that eMate (i) identifies and accurately targets the problematic constructs for an individual and (ii) positively affects aspects of behavior change through tailored interventions.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Decision support

Keywords

e-coaching; decision support system; HCI; model-based diagnostics; eHealth; behavior change

1. INTRODUCTION

Advances in pervasive computing and agent systems are opening up new possibilities for intelligent decision support. Owing to their ability to constantly and unobtrusively monitor behavior and provide support, ambient agents are increasingly being incorporated in decision support systems for aiding self-improvement in aspects of people's daily life. Systems that can help negotiate a good price for a new house [8], offer support for in-store purchases [27], or coach one towards more efficient energy consumption [17] are only a few examples of how intelligent decision support systems (IDSS) can contribute to people's self-improvement. Healthy living is one of the domains in which the potential of IDSS is especially large. Using modern, persuasive technologies to support and coach people to increase their level of self-monitoring and to adhere to a healthy lifestyle may aid in decreasing the cost of health-care services as well as the work load of medical professionals [7, 20]. At the same time, these technologies are able to provide information that is closely tailored to the needs of an individual [18, 13]. Although such technologies are receiving more and more attention (e.g., [2]), implementing and especially evaluating decision support systems remains challenging. Models and guidelines concerned with behavior change support are often not evaluated beyond being tested in restricted contexts and with prototypes. Studies examining the effects of fully functional IDSS on behavior change in daily settings are still limited in number. Considering the potential impact of IDSS on people's decision making and behavior, validation of these systems — especially when they are concerned with sensitive domains such as health or safety — is essential and deserves attention.

The eMate system is a versatile e-coaching system that is designed to coach people towards lasting behavior change in health domains such as maintaining a healthy diet, regulating medicine intake and therapy adherence. It relies on a model of behavior change, called the COMBI model [11], which consists of constructs that represent the cognitive and emotional states of a user that are related to different stages of behavior change. In this paper, an extensive, monthlong validation study of the eMate system is presented, in which 82 Dutch students received support to take the stairs more often. The performance of the system is evaluated with regard to a) the method of determining the content of the support and b) the effect of the support. With respect to these measures, three hypotheses were formulated and tested:

- **H1** eMate identifies and accurately targets the problematic constructs for an individual;
- **H2** eMate's interventions do not have a negative effect on the constructs for coachees;
- **H3** targeting by eMate improves construct values and promotes stage progression.

Section 4 elaborates on these hypotheses. Results show that $H1$ and $H3$ are confirmed, but that $H2$ has to be rejected on account of a small but significant decrease for the construct social norms (but see Section 6.2 for discussion).

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

This paper is structured as follows. Section 3 introduces the eMate system and describes the COMBI model of behavior change. In Section 4 the details of the empirical validation study are presented. Section 5 presents the results, which are further discussed in Section 6. Finally, in Section 7 we conclude that the eMate system can be successfully used to target causes of non-adherence and that eMate is able to contribute to behavior change.

2. EVALUATING IDSS

To provide effective and accurate support that is consistent with the current state of the user, IDSS need to have knowledge of the domain as well as knowledge of the user. Such knowledge can be obtained by exploiting models that draw on insights from psychology, sociology, neurology and other disciplines. Computational models have shown their merit in mapping the interplay between human emotion, cognition and behavior, and enable reasoning about transitions in these over time (e.g., [23, 4]).

When insights from other disciplines are translated to formal models that form the core of decision support systems, the questions are a) whether the models provide a valid representation of human states (e.g., attitude, behavior, beliefs), and b) whether the systems that use such models are able to achieve change in those states. Although these questions are crucial for the scientific validity of agent-based support systems, they are often only partly addressed in research. Promising models are regularly presented that have not yet been implemented or only as a prototype (e.g., [1, 14]). Many support systems validations focus on 'face validity', expert evaluations or exploring hypothetical scenarios (e.g., [21, 28]). Although these approaches can provide valuable observations for further development and eventual use, they do not provide the insights that come from analyzing human-agent interaction between the system and target users in everyday life.

Alternatively, simulations can provide an answer to validation questions by comparing model outputs to data from particular domains. Yet there are many cases in which data is not readily available (for example when modeling emergency or security scenarios), or when it is difficult to establish the degree to which the used data will reflect future behavior (for example when modeling stock markets).

Another approach is to implement the model in a simplified context, and explore performance in a lab setting or a gaming context ([15, 6]). Although human behavior in experiments in these settings can add supportive evidence about establishing behavior change, the complexity and dynamics of human-agent interaction are not taken in to account. Using virtual worlds of training or serious gaming to test support systems can address this issue [3, 25], yet the effect of support on behavior in everyday life, i.e., 'in the wild', is still not a part of the equation.

Conducting experiments with people is timely and costly[1], and consequently decision support systems are often not thoroughly tested to see if the desired outcomes were established. The question of whether a decision support system can establish the desired change is the motivation for the present work.

[1]It has even led researchers to create a method for designing computer agents specifically for the purpose of evaluating other agent systems! See [16].

Table 1: The constructs of the model

construct	description
susceptibility	likeliness of being affected by consequences of the behavior
severity	severity of the consequences of the behavior
pros/cons	beliefs about the importance of the behavior change
emotions	feelings concerning the behavior change
social norms	influence of culture and environment of a person
barriers	practical obstacles that prevent behavior change
skills	experience and capabilities to overcome the barriers
cues	environmental or physical stimuli
threat	perceived risk of continuing to perform behavior
attitude	beliefs, emotions and dispositions towards behavior
self-efficacy	perceived behavioral control
coping strategies	ability to deal with tempting situations and cues
mood	temporary state of mind defined by feelings and dispositions
high-risk situations	contexts/environments that influence a person's behavior
awareness	conscious knowledge of one's health condition, the health threat and the influence of current behavior
motivation	incentives to perform goal-directed actions
commitment	(intellectual or emotional) binding to a course of action

3. EMATE AND THE COMBI MODEL

The COMBI model — depicted in Figure 1 — is the result of careful formalization and integration of several well-established psychological theories of behavior change [11]. It uses the idea of *stages of change* from the Transtheoretical Model [22] to describe phases of the behavior change process. The five consecutive stages are *precontemplation* (PC), *contemplation* (C), *preparation* (P), *action* (A), and *maintenance* (M). COMBI considers sixteen different constructs that influence each other and these stages. Table 1 lists the different constructs.[2]

The eMate system is an intelligent, autonomous e-coaching system designed to promote behavior change. An e-coach is an IDSS with the objectives to *reinforce* current attitudes, making them more resistant to change, to *change* a person's response (behavior), and to *shape* a pattern of behavior where such one did not exist beforehand, in line with those of persuasive systems [19]. eMate monitors the behavior of coachees to determine adherence to health goals and uses a mobile phone app (Figure 2) and a website (Figure 3) to interact with the coachees [12].

eMate uses rule-based reasoning with the COMBI model to hypothesize about the *stage of change* of a coachee and the constructs that cause non-adherence. Once the current stage of change of the coachee is determined (from the intake survey), it is checked for all constructs related to the *consecutive* stage of change whether their values are up to date, because those are the constructs that can contribute to change. If a value v is out of date, a question is sent to the coachee's smartphone to obtain a new value for this construct (see Figure 2b). When it turns out that the new v is lower than threshold τ (in this case if $v \leq 5$, indicating that it is a *problematic* construct), the constructs that relate to that construct are investigated. Pseudocode for this algorithm can be found in Algorithm 1. After the constructs are investigated, a *target* is randomly chosen from all constructs that have a value below 7.[3] This target is the construct that will be the subject of the intervention mes-

[2]More information on these constructs can be found in [11].
[3]A threshold of 7 was chosen to make sure that there was sufficient variety in messages.

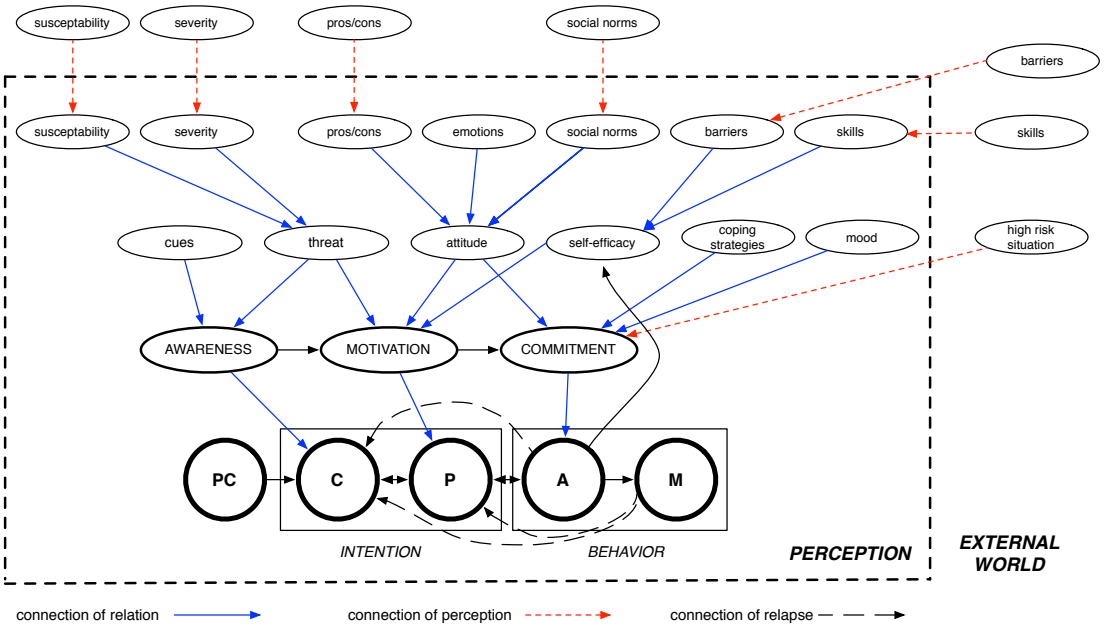
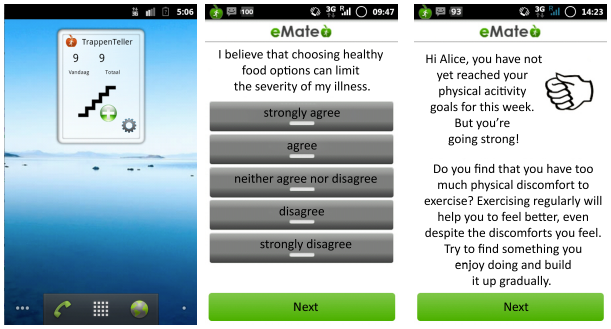


Figure 1: The integrated model of behavior change COMBI

sage sent to the user (see for an example Figure 2c). The messages are assembled automatically from 3 components: a summary of the coachee's behavior in the last few days with respect to his/her goal, construct-related motivational text (for each target, eMate chose from 3 alternative texts), and a concluding remark (similar to the message structure in [26]).

For this particular study, domain information about taking the stairs was added to eMate to use in the messages.[4] Furthermore, an additional monitoring instrument (i.e., an Android widget, see Figure 2a) was developed and added to the regular eMate app. The interface of the widget showed coachees how many stairs they had taken today and in total from the start of the experiment. With each tap on the widget one flight of stairs was counted (where one flight is defined as all the steps between two floors).



(a) The widget. (b) eMate questions (c) eMate messages

Figure 2: The Stairs Counter smartphone apps.

4. EVALUATION METHOD

The evaluative study was concerned with improving people's stairs taking behavior. All participants (N=82) were asked to monitor the number of stairs they climbed for four weeks. After one week, a new goal was set, based on individual behavior in that first week (the new goal was a 10% increase). Participants then received three weeks of remote coaching to reach that goal. In the study, half of the messages were automatically generated by eMate, and half were written by human (non-expert) coaches.[5] It was ensured that all messages shared the same format and informational content with respect to the constructs (this was

Figure 3: The eMate website showing a picture of a building with an equivalent number of stairs as the participant has taken (here: the Rocky Steps).

[4]This information was checked by several health psychologists and communication experts.

[5]This is because the experimental setup was also used to answer a different research question about how coachees perceived their coach. These conditions are however not relevant for the purposes of this evaluation and are therefore ignored here.

Algorithm 1 Finding the problematic constructs for behavior change

```
C ←the set of all constructs in the model graph
S ←the ordered set of all stages of change:{PC < C < P < A < M}
s_i ←the current stage of change s_i of the user, s_i ∈ S
s_j ←the stage that directly succeeds s_i, s_j ∈ S
τ_i ← the threshold for construct i
l_i ← the lifetime for a value of construct i
problem ←list of problematic constructs, initially empty
for all c_k ∈ C do ▷ cycle through all constructs linked to the stage
    if connected(c_k, s_j) then
        INVESTIGATE(c_k)
    end if
end for

function INVESTIGATE(construct c_i)
    if value(c_i) = ∅ OR age(c_i) > l_i then
        update c_i ▷ ask user questions about this construct
        if value(c_i) < τ_i then ▷ up-to-date value is indeed below threshold
            problem ← bottleneck + c_i
            for all c_j≠i ∈ C do
                if connected(c_j, c_i) AND age(c_j) < l_j then ▷ recursively investigate constructs on this path
                    INVESTIGATE(c_j)
                end if
            end for
        end if
    end if
end function
```

enforced by the way the human coaches could write the messages). The participants received questions and messages on the smartphone app and could also view these on the eMate website. On that same website, participants could monitor their progress and see a motivational picture relating to the stairs they had already climbed (see Figure 3).

Participants were asked fill out an intake questionnaire at the beginning of the study and an post questionnaire at the end. Both questionnaires were filled out online and consisted of several validated surveys as well as some single items, all pertaining to the model constructs. The intake questionnaire also included questions about demographics; the post questionnaire had additional questions about user experience (not discussed in this work).[6] At the end, participants were thanked for their participation and were payed €10.

The answers from the intake questionnaire served as initial data points for the model. Answers pertaining to the same construct were aggregated and scaled to a value between 0–10 using a conversion script. The conversion ensured that for all constructs a higher score is better (so, a higher score for barriers means that there were *less* barriers).

With regard to the evaluation of eMate, the study had three main objectives. The first was to evaluate whether eMate's model-based diagnosis — the identification of specific, problematic constructs that are preventing behavior change — was accurate and that eMate correctly targeted those low constructs. The second was that we would find no adverse effects of using the eMate system. The third was whether the targeted interventions positively affected the problematic constructs and progress through stages. In line with the objectives, we formulated and tested the three hypotheses listed in Section 1. For $H1$ we examined whether low constructs as identified by the coachees at intake were identified as problematic constructs by the eMate system. In addition, we studied the differences between the intake survey values for constructs and the (mathematically) derived construct values. For $H2$ we analyzed the values of the constructs that were targeted by intervention messages and of the constructs that were identified as problematic. Finally, for $H3$ we compared the construct values derived from the evaluation questionnaire with the construct values from the intake questionnaire. In addition, we tested whether the coachees had progressed through the stages and whether any changes in construct values were related with this progression.

[6]All used surveys (in Dutch) can be found at `http://bit.ly/stairs_surveys`.

5. RESULTS

For the analyses, two data sets were used. Of the 82 participants who started with the study, 8 (9.8%) quit early and did not complete the post questionnaire. Data set A (N=74) contains all people who participated in the study and filled out both the intake and the evaluation questionnaires. Data set $A\prime$ (N=65) excludes *non-active coachees*, which we defined as coachees for whom more than 40% of the widget data was missing. The following three subsections discuss the results pertaining to the three hypotheses.

5.1 Model-based diagnostics ($H1$)

During the study, a total of 448 messages were generated, of which 415 were received (from the others no explicit acknowledgement of receipt was returned by the phone, possibly because of network problems). For testing whether eMate accurately targeted the problematic constructs, we (re)generated the list of problematic constructs on the basis of 1) the intake questionnaire and 2) the results of the reasoning process. We compared this list to the targets of the messages that were sent to the coachees. We expected that each targeted construct was an element of the set of problematic constructs. Indeed, we found that this was the case for all 415 messages that were sent.

We also expected the mean initial value of the targeted constructs to be lower than the initial value of the constructs that were not targeted. This would be another indication that the system accurately chooses the right constructs to target. We found that this was the case: at the start of the intervention, the mean value of constructs that were targeted for an individual was 4.70, compared to 7.59 for constructs that were not targeted (t-test, p-value = 2.2e-16).

Another aspect of the evaluation of the model-based diagnoses is the analysis of the combination functions that capture the influences between constructs in the model. The COMBI model is developed as a computational model with the aim of eventually predicting user states in order to proactively target problematic constructs. For the present study, the reasoning mechanism did not make predictions for the constructs, but instead used user input in the form of question answers and widget data. A comparison of several simple mathematical combination functions to predict construct values was done in [12]. It was concluded that an algorithm of taking the maximum of inputs performed reasonably well (average error of 1.74 on a scale of 0-10). As inputs the values at the top level of the model were used; the output values were compared with the values that were derived from

Table 2: Error for calculated construct values using max and min strategies.

	PRE error		POST error	
Construct	Max	Min	Max	Min
threat	**1.99**	2.42	2.22	**2.00**
attitude	**1.20**	4.70	**1.05**	5.04
self-efficacy	**1.95**	2.55	**1.59**	3.24
awareness	**0.89**	5.59	**2.69**	5.38
motivation	6.19	**2.59**	5.47	**2.59**
commitment	**0.55**	6.88	**1.15**	6.86

the questionnaires. Similar analyses using a max and a min strategy were done for data set A. Results can be found in Table 2.

From Table 2 it follows that there are quite consistent performances of *max* and *min* with respect to the pre and post construct values. There are however some notable differences: for threat, *min* performs slightly better, and although *max* performs best for commitment and awareness, the mean error for those construct values vary greatly. Overall, it seems that *max* is a reasonably good strategy for all constructs except motivation, where *min* is clearly the better strategy. This combined max-min strategy results results in an overall mean error (for pre and post together) of 1.70.

5.2 No negative effects ($H2$)

We expected that the total intervention (i.e. monitoring and coaching) would have no clear negative effects. Table 3 shows whether people did worse, stayed neutral, or improved with respect to the COMBI constructs. Most coachees stayed the same with regard to awareness and commitment, but there are promising improvements in barriers (58), coping (49), motivation (37), and self-efficacy (37) for data set A.

Table 4 shows the means of all COMBI model constructs before and after the intervention. Three constructs were improved significantly by the coaching: barriers, coping and motivation. For the construct of social norms there was a significant decrease. These findings are further discussed in Section 6.

Finally, an analysis was performed that focused on the participants' problematic constructs. For all participants it was established whether their problematic constructs had increased at the end of the experiment (which differs from Table 4, where all the start and end values were compared independent of whether the constructs were problematic). The constructs that were identified as problematic for more than 25% of the participants, as well as the p-value for improvement (1-tailed paired t-test), can be found in Table 5. In line with the analysis shown in Table 4, motivation, barriers and coping all significantly improved. Additionally, threat, emotion and self-efficacy showed significant improvement. Of all constructs that were problematic for at least 25% of the participants, only severity did not increase significantly.

5.3 Effects of targeting ($H3$)

The effects of the intervention were measured in a number of ways, from recorded stairs use to reported elevator use, and from stage improvement to construct improvement.

Table 3: Differences in construct values before (a) and after (b), largest group in bold.

Construct	b-a < 0	b-a $= 0$	b-a > 0	b-a ≥ 0
susceptibility	23	25	**26**	51
severity	10	**53**	11	64
skills	25	16	**33**	49
cues	6	**53**	15	68
threat	15	**48**	11	59
mood	25	14	**35**	49
pros cons	14	**41**	19	60
emotions	12	**49**	13	62
barriers	6	10	**58**	68
social norms	**42**	15	17	32
attitude	13	**54**	7	61
coping	20	5	**49**	54
awareness	5	**68**	1	69
motivation	12	25	**37**	62
commitment	7	**66**	1	67
self-efficacy	21	8	**45**	53

Table 4: Construct μ before (a) and after (b).
* Significant at <0.05, ** significant at <0.05 after False Discovery Rate (FDR) correction.

Construct	μ_a	σ_a	μ_b	σ_b	p
susceptibility	6.87	1.37	6.85	1.40	0.94
severity	4.60	2.45	4.66	2.24	0.83
skills	7.51	0.98	7.69	1.74	0.46
cues	9.05	1.47	9.34	1.36	0.11
threat	6.26	1.65	6.10	1.54	0.44
mood	6.93	1.27	6.91	2.26	0.92
proscons	8.10	1.97	8.19	2.05	0.75
emotions	7.50	2.90	7.50	2.90	1.00
barriers	4.37	2.66	8.20	2.40	**1.48E-14****
social norms	4.62	2.14	3.78	3.08	**0.0042****
attitude	8.78	2.46	8.38	2.36	0.18
coping	6.50	1.75	7.26	2.32	**0.0086****
awareness	9.89	0.93	9.38	2.34	0.087
motivation	3.16	2.71	4.26	2.92	**0.0048****
commitment	9.73	1.63	8.92	3.13	**0.033***
selfefficacy	6.05	1.37	6.65	2.47	0.059

Results from the former measures are discussed elsewhere. Here, we focus on the measures pertaining to the COMBI model. We examine the change in the stages of change and the effect of the messages on the targeted constructs.

5.3.1 Stages of Change

In data set A we found no significant differences between begin and end stage (Wilcoxon signed rank test, p-value = 0.25). However, when the inactive participants were removed ($A\prime$, N=65), a significant improvement (p<0.05 level) was visible (p-value = 0.046). An interpretation of this result is given in Section 6.3. Since most participants started out in a stage that corresponded to good performance (A or M, N=56), another analysis was performed that focused on the participants that were in a stage in which they could improve, i.e. in stages PC, P, or C.[7] Participants who started

[7]In M they could not improve and participants could only progress from A to M when they performed the behavior for

Table 5: Increase in values of problematic constructs.
** Significant at < 0.01.

problematic construct	% of participants	improved? (p)
severity	92 %	0.07
motivation	86 %	**6.826E-07****
barriers	77 %	**2.2E-16****
threat	55 %	**0.001****
emotion	46 %	**3.768E-05****
self-efficacy	32 %	**3.872e-05****
coping	27 %	**0.004****

in stage PC, P or C (N=18) improved significantly over the course of the study (Wilcoxon signed rank test, p<0.01). On average they improved one stage. Those who started in A and M did not change significantly.

We also tested whether any of the constructs was correlated to end stage. We found that only one was: self-efficacy at intake was strongly correlated with the final stage (Pearson's r=0.32, p-value = 0.0048).

5.3.2 *Change in targeted constructs*

Of the 415 motivational messages sent, 122 messages targeted a construct for an individual coachee for the second or third time. Thus, 293 constructs–person combinations were targeted. We calculated the change in the mean value for the constructs between the start of the study and the end. The mean value of the 293 targeted constructs increased with 1.23 on average (on a 10-point scale), while the mean value of 891 non-targeted constructs decreased with 0.056 (t-test, p-value = 1.124e-10).

On the level of the individual constructs, the change in values between targeted and non-targeted constructs is less clear. Table 6 lists the constructs, the number of coachees for which this construct was targeted (of the 74 in data set A), and the change in mean value for the targeted and non-targeted set. There is a significant increase for susceptibility, skills and emotions, with a moderate to large effect size (Cohen's d). The constructs susceptibility, skills and emotions are significant even after applying the FDR (BH) correction to control for the large number of constructs.

Table 6: Change in values for (targeted) constructs (ordered by frequency).
* Significant at <0.05, ** significant at <0.05 after FDR.

Construct	#	δ target	δ ¬target	p-value	Cohen's d
motivation	34	1.7941	0.5000	0.0822	0.399
barriers	31	4.7742	3.2326	0.0563	0.443
severity	30	0.000	0.1136	0.8524	-0.042
social norms	29	-1.2759	-0.5556	0.2293	-0.295
coping	26	1.3462	0.4375	0.1299	0.377
self-efficacy	25	0.8000	0.4898	0.6108	0.116
susceptibility	25	0.7200	-0.3877	**0.0020****	0.728
skills	24	1.4167	-0.4200	**0.0001****	0.896
emotions	19	1.5789	-0.5455	**0.0038****	0.686
mood	19	0.3684	-0.1636	0.3900	0.222
threat	16	0.3750	-0.3103	0.0628	0.384
attitude	6	0.8333	-0.5147	0.1772	0.521
pros cons	4	5.2500	-0.2000	0.0547	2.170
commitment	2	0.0000	-0.8333	0.9471	0.260
cues	2	1.5000	0.2500	0.5568	0.821
awareness	1	–	–	–	–

6 consecutive months, longer than the duration of the study.

6. DISCUSSION & LIMITATIONS

In this section the results from Section 5 are discussed in relation to the hypotheses $H1$-$H3$. We will start with some general remarks about the functioning of the system, then discuss the hypotheses in order, and finish with some limitations with respect to the study. To begin, there were no major problems during the experiment. There were no crash reports filed for the eMate app at any time. For 10 of the 82 people (12%) who started the study, the interface of the newly developed widget was slightly malformed, but still fully functional. There was a technical difficulty with parsing some of the answers options for constructs at the second and third level of the model (e.g., social norms or attitude, but not self-efficacy). As a result, some answers resulted automatically in a value lower than 5 (regardless of whether it actually was low). Luckily, it turned out to have only affected constructs that were above a first-level construct that was already identified as a hypothesis (for only then were they used in the reasoning process, see Algorithm 1). Moreover, because of the way the targets were chosen (see Section 3), participants would not be targeted solely on that construct.

6.1 Hypothesis 1

We expected eMate to identify and accurately target the problematic constructs for an individual. We found that all 415 messages that were received by coachees correctly targeted a problematic construct for each individual. Moreover, the mean value of constructs that were targeted were significantly lower compared to the mean of constructs that were not targeted. In light of this evidence, we confirm $H1$.

Although preliminary, the results on calculation and prediction of the construct values are encouraging. For all constructs, the straightforward strategy *max* seems a reasonably good fit to calculate the value, except for motivation. This could indicate that people's motivation has a negative bias: motivation is more likely to be influenced by negative input than positive input. However, more work is needed to confirm this.

6.2 Hypothesis 2

$H2$ stated that eMate's interventions do not have a negative effect on the constructs for coachees. Contrary to $H2$, the construct social norms significantly decreased. We suspect that this difference is caused by initial overly optimistic answers by the coachees. It is quite plausible that before the experiment, people had never thought about social support regarding taking the stairs.[8] As a result, they could have overestimated the social support because of *attribute substitution* [10]: other examples of social support easily come to mind, so people judge that the social support for this activity is likely to be high. However, after a period of four weeks of monitoring their stairs use, they would have had a much more realistic picture of the social support they received.

In addition, the construct commitment also shows a slight decrease (though not significant after the FDR correction). The table shows that commitment was extremely high. This is not surprising, seeing how people had just committed themselves to a month of monitoring their stairs use. Overall, it is encouraging to see that only 7 people (less than

[8]In fact, several participants informally mentioned during the intake procedure that they found questions about social support in relation to stairs use strange.

10% of the total sample) were less committed after the experiment. The decrease can be explained by the fact that we used a Y/N question to measure commitment. Even people who were only slightly tired of monitoring their stairs use had no other choice but to answer that they were no longer committed. In future work, commitment will be measured with a more fine-grained measure.

In all, $H2$ has to be rejected. However, it should be stressed that that are no indications that the decreases in the constructs discussed above were caused by adverse effects of the interventions.

Lastly, most coachees stayed the same with regard to awareness and commitment. This is explained by the fact that all but one coachee gave the highest possible ranking for awareness, and all coachees gave the highest ranking for commitment at the start of the experiment (see the means of these constructs in Table 4). The result for severity and threat can be explained by considering the problem domain: people consistently (and accurately) judged not taking the stairs as only a small risk to their health. The fact that severity was valued low was corroborated by the analysis that examined only the set of problematic constructs with regard to improvement. It was shown that almost all participants scored low on severity and that it was not significantly improved, while all other problematic constructs improved significantly after use of the system. We suspect that this finding is not related to eMate but inherent to the domain.

6.3 Hypothesis 3

$H3$ posited that targeting improves model constructs (compared to those who are not targeted) and the stage of change. We found that taking into account all participants, there were no significant improvements of the stages. However, there were significant improvements when considering 1) only those participants who had used the widget regularly, or 2) only those participants who could improve their stage. The first result may be an indication that using the widget to monitor stairs use was an important part of the intervention. This idea is supported by the fact that coachees reported the use of the widget to be the most motivating aspect of the system (median answer: 'quite'). The second result shows that for the group for which improvement was feasible, their interventions were successful.

With respect to the change in targeted constructs, several observations can be made. On average, the targeted constructs significantly increased compared to the non-targeted constructs. This change was particularly large for the constructs susceptibility, skills or emotions. Furthermore, the constructs that were targeted most often (motivation, barriers, severity, social norms) showed no significant improvement, which suggests that eMate indeed chose to target those constructs that were indeed problematic. Future work could examine this more closely and adapt the reasoning mechanism in order to benefit from this knowledge, for example by introducing construct parameters of *changeability*. As for $H3$, we find that the results on the improvement in stage and of the targeted constructs support the hypothesis.

There are a few limitations to the present study and the underlying system. First, to be able to draw conclusions about an enduring change, behavior should be measured over a longer period of time (e.g., 6 months). Thus, it should be stressed that the aim of this study is not to evaluate the long-term effectiveness of a behavior change intervention. Instead, we evaluated the reasoning mechanism of eMate in a domain in which behavior can be influenced relatively quickly, so that the shorter time frame is appropriate. Secondly, and related to the first point, there was no control group in the experiment. Given this design, we cannot draw conclusions about the general effectiveness of the intervention. Thirdly, it should be noted that several studies have concluded that there is limited or no evidence for the effectiveness of stage-based interventions based on the Transtheoretical model [9, 24]. The COMBI model explicates which determinants contribute to the stages, in an attempt to improve the effectiveness of targeted interventions. Also, some studies have shown that stage transitions are common even without interventions, and that these transitions can occur in short time intervals [5]. These results underline the importance of regularly updating the stage classification (which eMate does when analyzing the telephone answers) to prevent the interventions to be tailored to the wrong stage.

7. CONCLUSIONS AND FUTURE WORK

This paper has presented an extensive, monthlong evaluation study of the eMate system and the underlying COMBI model of behavior change. We found that eMate accurately identifies and targets the problematic constructs in the model and that targeting positively influences the problematic constructs for non-adherence. Contrary to initial expectation, one construct, social norms, had decreased after the intervention. Overall, however, there were no indications that the coaching had adverse effects. In fact, we found that targeting by eMate improves construct values and promotes stage progression. We conclude that the eMate system can be successfully used to target causes of non-adherence and that deploying eMate can contribute to behavior change.

Many challenges for future work remain open; we shall name a few. First, eMate will have to be evaluated in other domains to further establish its worth as a versatile system for behavior change. Second, the computational model will be used to predict the construct values, in line with the method that was discussed in Section 5.1. More advanced mathematical combination functions will be used to derive construct values. An example could be a dynamic approach where state transitions will be derived by $V_{t+1} = V_t + \eta[f(\sum \omega_n V_n) - V_t] \cdot \Delta t$, where η is a learning parameter and f a combination function that specifies how thresholds are handled. Parameter tuning methods will be deployed to find optimal parameter values for prediction. Third, the system can be extended to learn to which of the persuasive techniques that are incorporated in the messages (e.g., authority, familiarity, social comparison) the user responds best. The system can then adaptively pick those that provide optimal support for the user.

Acknowledgments
This research was supported by Philips and Technology Foundation STW, Nationaal Initiatief Hersenen en Cognitie NIHC under the Partnership programme Healthy Lifestyle Solutions. We thank Inge Wolsky for her contributions to the content of the messages.

8. REFERENCES

[1] V. Aleven, B. Mclaren, I. Roll, and K. Koedinger. Toward meta-cognitive tutoring: A model of help

seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 16(2):101–128, 2006.

[2] J. H. Anderson and B. A. Kamphorst. Ethics of e-coaching: Implications of employing pervasive computing to promote healthy and sustainable lifestyles. In *Proc. of IEEE SIPC Workshop 2014, in conjunction with PerCom 2014 (to appear)*. IEEE Computer Society Press.

[3] T. Baranowski, R. Buday, D. I. Thompson, and J. Baranowski. Playing for real: Video games and stories for health-related behavior change. *American journal of preventive medicine*, 34(1):74–82, 2008.

[4] T. Bosse, M. Hoogendoorn, M. Klein, J. Treur, C. N. van der Wal, and A. van Wissen. Modelling collective decision making in groups and crowds: Integrating social contagion and interacting emotions, beliefs and intentions. *JAAMAS*, 27(1):52–84, 2013.

[5] J. de Nooijer, P. van Assema, E. de Vet, and J. Brug. How stable are stages of change for nutrition behaviors in the netherlands? *Health Promotion International*, 20(1):27–32, 2005.

[6] J. Dias and A. Paiva. I want to be your friend: Establishing relations with emotionally intelligent agents. In *Proc. of AAMAS*, pages 777–784, 2013.

[7] R. E. Glasgow, L. Chance, D. J. Toobert, J. Brown, S. E. Hampson, and M. C. Riddle. Long term effects and costs of brief behavioural dietary intervention for patients with diabetes delivered from the medical office. *Patient education and counseling*, 32(3):175–184, 1997.

[8] K. V. Hindriks and C. M. Jonker. Creating human-machine synergy in negotiation support systems: Towards the pocket negotiator. In *Proc. of the 1st International Working Conference on Human Factors and Computational Models in Negotiation*, HuCom'08, pages 47–54, New York, 2009. ACM.

[9] S. M. Horowitz. Applying the transtheoretical model to pregnancy and std prevention: A review of the literature. *American Journal of Health Promotion*, 17(5):304–328, 2003.

[10] D. Kahneman and S. Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman, editors, *Heuristics and Biases: The Psychology of Intuitive Judgment*, pages 49–81. Cambridge University Press, Cambridge, UK, 2002.

[11] M. Klein, N. Mogles, and A. Van Wissen. Why won't you do what's good for you? Using intelligent support for behavior change. In *International Workshop on Human Behavior Understanding (HBU'11)*, volume 7065 of *LNCS*, pages 104–116. Springer Verlag, 2011.

[12] M. Klein, N. Mogles, and A. Van Wissen. Intelligent mobile support for therapy adherence and behavior change (to appear). *Journal of Biomedical Informatics*, 2013.

[13] P. Krebs, J. O. Prochaska, and J. S. Rossi. A meta-analysis of computer-tailored interventions for health behavior change. *Preventive medicine*, 51(3):214–221, 2010.

[14] S. Li. Agentstra: an internet-based multi-agent intelligent system for strategic decision-making. *Expert Systems with Applications*, 33(3):565–571, 2007.

[15] R. Lin and S. Kraus. Can automated agents proficiently negotiate with humans? *Communications of the ACM*, 53(1):78–88, 2010.

[16] R. Lin, Y. Oshrat, and S. Kraus. Automated agents that proficiently negotiate with people: Can we keep people out of the evaluation loop. In *New Trends in Agent-Based Complex Automated Negotiations*, pages 57–80. Springer, 2012.

[17] C. Midden and J. Ham. Using negative and positive social feedback from a robotic agent to save energy. In *Proc. of Persuasive '09*, pages 12:1–12:6, New York, NY, USA, 2009. ACM.

[18] S. Noar, C. Benac, and M. Harris. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin*, 133(4):673–693, 2007.

[19] H. Oinas-Kukkonen and M. Harjumaa. A systematic framework for designing and evaluating persuasive systems. In H. Oinas-Kukkonen, editor, *LNCS*, volume 5033 of *PERSUASIVE'08*, pages 164–176. Springer-Verlag Berlin Heidelberg, 2008.

[20] R. L. Ownby, D. Waldrop-Valverde, R. J. Jacobs, A. Acevedo, and J. Caballero. Cost effectiveness of a computer-delivered intervention to improve HIV medication adherence. *BMC medical informatics and decision making*, 13(1):29, 2013.

[21] A. Pommeranz, P. Wiggers, W.-P. Brinkman, and C. M. Jonker. Social acceptance of negotiation support systems: Scenario-based exploration with focus groups and online survey. *Cognition, Technology & Work*, 14(4):299–317, 2012.

[22] J. Prochaska and C. DiClemente. *The transtheoretical approach: Crossing traditional boundaries of therapy*. Dow Jones-Irwin, Homewood, Ill., 1984.

[23] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60, 2005.

[24] S. Salmela, M. Poskiparta, K. Kasila, K. Vähsärja, and M. Vanhala. Transtheoretical model-based dietary interventions in primary care: A review of the evidence in diabetes. *Health Education Research*, 24(2):237–252, 2009.

[25] B. Silverman, D. Pietrocola, B. Nye, N. Weyer, O. Osin, D. Johnson, and R. Weaver. Rich socio-cognitive agents for immersive training environments: Case of nonkin village. *JAAMAS*, 24(2):312–343, 2012.

[26] M. J. Sorbi, S. B. Mak, J. H. Houtveen, A. M. Kleiboer, and L. J. van Doornen. Mobile web-based monitoring and coaching: Feasibility in chronic migraine. *Journal of medical Internet research*, 9(5), 2007.

[27] H. van der Heijden. Mobile decision support for in-store purchase decisions. *Decision Support Systems*, 42(2):656–663, 2006.

[28] J. M. v. d. Zwaan, E. Geraerts, V. Dignum, and C. M. Jonker. User validation of an empathic virtual buddy against cyberbullying. *Annual Review of Cybertherapy and Telemedicine 2012: Advanced Technologies in the Behavioral, Social and Neurosciences*, 181:243, 2012.