Agent-based support for behavior change

van Wissen, A.

2014

document version
Publisher's PDF, also known as Version of record

10

Human Involvement in e-Coaching: Effects on Effectiveness, Perceived Influence and Trust

Bart Kamphorst, Michel Klein and Arlette van Wissen[1]

Abstract Coaching practices are rapidly changing due to technological advances in the area of pervasive computing: the behavior of coachees can be monitored real-time, and coaching can be remote or even fully automated. The movement towards autonomous e-coaching systems holds promise for improving self-management, but also gives rise to questions about the importance of human involvement in the e-coaching process. This paper addresses this issue and describes an empirical 'Wizard of Oz' study in which coachees (N=82) were coached to regularly climb the stairs by either a human coach (N=20) or the autonomous e-coaching system eMate. However, some coachees were deceived into thinking that they received one type of coaching (human or computerized), while in reality they received the other. Results show that the coaching was equally effective in all groups, but suggest that people who believed to be coached by a human coach judged the coaching to be more influential. Finally, no difference was found in how trustworthy coachees found their coach for the different types of coaches.

10.1 Introduction

Technological advances are the cause of major changes in current coaching practices. Among such advances are decision aids and support systems, which belong to a growing area in pervasive computing. Due to their characteristics of being mobile, adaptive and context-aware, ambient systems show increasing promise for supporting humans in their everyday life. In a recent technology assessment of e-coaching systems that was carried out by the Dutch Rathenau Institute, Kool, Timmer, and van Est describe the current coaching practice and the roadmap to full digitalization of that practice, which involves the digitalization of both coachees and coaches (Kool et al., 2013). The internet has created opportunities for remote coaching, and more recently, smart sensor systems have enabled both coaches and coachees to gain fine-grained insight into the coachee's health behavior and physical states (e.g., by monitoring heart rate, body temperature, or sleep patterns).

[1]The authors are mentioned in alphabetical order and have made a comparable contribution to the article.

Autonomous e-coaching systems are a class of decision support systems designed to assist people with self-improvement in a variety of areas (Warner, 2012). In the foreseeable future, autonomous e-coaching systems are likely to increase in prevalence, due to low costs and a level of anonymity that many people favor (see Fogg (2003)). However, because coach-coachee relationships are of crucial importance for effective coaching (Gyllensten and Palmer, 2007), critics of the trend towards automated coaching might object that autonomous e-coaching systems will miss 'that particular human quality' that makes for good relationships. This, then, raises the question: What influence does human involvement in e-coaching actually have?

This paper describes an empirical study on computer-mediated coaching that involves a deception about the type of coaching that participants received.[1] This experiment tests whether people have different coaching experiences when being coached by a computer rather than by a human coach. Half of the coachees received coaching from the fully autonomous e-coaching system eMate (Klein, Mogles, and van Wissen, 2011), while the others received remote coaching from human coaches. However, some coachees were made to believe that they received one type of coaching, while in reality they received the other. By using this 'Wizard of Oz' type of experiment, it was possible to test the effects of the coachees' belief about human involvement. The application domain was healthy lifestyle improvement by stimulating people to take the stairs more often. This domain was chosen because it has a reasonably clean outcome measure (the number of stairs taken) and because it allowed healthy people to partake in the experiment. This work focuses on three outcome measures: the effectiveness of the coaching, the perceived influence of the coaching (as judged by coachees), and the trust that coachees developed in their coaches. With regard to these measures we formulated several hypotheses (the motivations for which are listed in Section 10.2):

1. The belief that coachees have about the identity of their coach (human or computer) will have no effect on effectiveness. That is, coaching by eMate and coaching by a human coach will lead to an equal degree of behavior change.
2. The belief that coachees have about the identity of their coach (human or computer) will have no effect on how influential coachees consider the intervention.
3. The belief that coachees have about the identity of their coach (human or computer) will have no effect on trust.

Results show that all groups improved in comparable measure (although there was no significant improvement overall, see Section 10.5.4), which confirms hypothesis 1. However, contrary to hypothesis 2, coachees did judge the intervention to be more influential if they believed to be coached by a human coach. Two possible explanations of this difference are proposed, namely that humans feel a larger degree of similarity towards humans compared to computers, and that the coachees have more experience with human coaches than with computer coaches. Finally, with regard to trust, no differences were found between groups, confirming hypothesis 3.

The structure of this paper is as follows. Section 10.2 describes related work on how humans generally perceive computers. Section 10.3 describes the autonomous e-coaching system eMate and the underlying COMBI model of behavior change. Section 10.4 gives an overview of the research methodology. Section 10.5 presents the results of the experiment, which are subsequently discussed in Section 10.6. Finally, we offer ideas for future work in Section 10.7.

[1] Because the human coaches were also participants, we will use 'coachees' and 'coaches' from now on to distinguish between these types of participants.

10.2 Related Work

The largest body of work on how people perceive computers comes from Nass (e.g., Fogg and Nass, 1997; Nass and Moon, 2000; Nass, Moon, and Carney, 1999; Reeves and Nass, 1996). His work on the "media equation" provides evidence that people treat computers as social actors, no other than they would treat other human beings. Reeves and Nass found that people unconsciously and automatically apply social rules when they interact with computers and other media (Reeves and Nass, 1996). This has been shown to hold for instance with regard to human factors such as flattery (Fogg and Nass, 1997) and politeness (Nass et al., 1999). Corroborating these findings is work by van Wissen, van Diggelen, and Dignum that shows that people exhibit similar levels of reciprocity towards humans and computer agents (van Wissen et al., 2009). Additionally, in an experiment on team formation in dynamic environments with uncertainty, van Wissen, Gal, and B. A. Kamphorst showed that people are as loyal to agent-led teams as they are to human-led teams (van Wissen et al., 2012). It was also shown that people prefer to create teams with others with whom they have had positive interactions (humans and agents alike), rather than those who offer them large rewards (van Wissen et al., 2012). Work by Bickmore and Picard further demonstrates that factors such as trust play an important role when designing for long-term interaction — which is particularly relevant for personal coaching — and that increased trust leads to a greater desire to continue the interactions (Bickmore and Picard, 2005). With regard to trust, it has also been shown that people generally trust computers in a similar way as they trust humans (Lee and Nass, 2010; Pruitt and Carnevale, 1993; van Wissen et al., 2012).

These findings together suggest that people would interact in a similar way with computer coaches as with human coaches. Other work, however, suggests some nuances to this claim for specific domains. For example, Blount showed that in an ultimatum game people are more likely to accept lower offers from computers than from human proposers (Blount, 1995). In addition, the work by van Wissen et al. shows that people treat computers differently in team settings with regard to fairness considerations (van Wissen et al., 2012). Their experiment used a 'Wizard of Oz' setup — similar to the one described in Section 10.4 — in which subjects were led to believe that they were interacting with autonomous computer systems, when in reality they were interacting with other human participants. This research differs however from the present work in that people could choose to stop collaborating with actors if the interaction had proven unsatisfactory. In contrast, in the experiments described here, people were paired with their coaches (eMate or human) for the full duration of the experiment. This means that continued interactions were required regardless of prior negative experiences. Another difference is that in van Wissen et al. (2012) computers fulfilled the role of teammates rather than decision aids. According to van Dongen and van Maanen, there is an asymmetry in how people attribute trust to themselves and to decision aids (van Dongen and van Maanen, 2006). Their work shows that although a lack of trust in one's own reliability disappears after practice, lack of trust in the reliability of the aid persists.

Although the mentioned literature shows some differences for particular social factors between human-human interaction and human-computer interaction, the present work does not involve these particular factors (negotiation, fairness, and choosing teammates). Hypotheses 1 and 2 are therefore based on the findings that showed similar social relations between humans and computers and between humans and humans. Furthermore, the computer system in the current experiment is not a mere decision aid (giving advice based on probabilities, algorithms and calculations), but also relies on persuasion strategies and repeated interaction to support

behavior change. Hypothesis 3 therefore assumes that the social role of the system will trigger similar trust responses in interactions with the eMate coach to interactions with a human coach.

10.3 eMate

The eMate system is an intelligent coaching system that uses both a mobile phone and a website to interact with the user Klein, Mogles, and van Wissen (2013); Klein et al. (2011). It is designed to support patients with Diabetes Mellitus type 2, HIV or cardiovascular disease in adhering to their therapy, which consists of lifestyle advice and/or precise instructions for medication intake. The system performs the following tasks: (1) it determines the *stage of change* of a user, (2) it *monitors* the behavior of the user to determine the level of adherence, (3) it *reasons* about the changes required for improvement, (4) it tries to *change the user's perception* of specific psychological aspects, and (5) it *updates* the beliefs about the user. The monitoring is performed via a combination of a web-based system, a smartphone app (for presenting messages and asking questions) and an optional electronic pillbox (not used in the present study).

The *stage of change* is a phase in the characterization of a behavior change process according to the Transtheoretical Model (Prochaska and DiClemente, 1984). This model assumes that behavior change passes through five stages: *precontemplation*, *contemplation*, *preparation*, *action*, and *maintenance*. The content of interventions depends on the stage in which a person is.

The core of the eMate system is formed by the COMBI model, which stands for COMputerized Behavior Intervention (Klein et al., 2011). This model integrates different theories of behavior change into a formal representation of a causal model. The eMate system uses the COMBI model to reason about underlying causes for non-adherence. It will determine what prevents the user from moving to the next stage of change by investigating whether the constructs that influence the consecutive stage (i.e., the factors that have a causal relation with this stage) are a *bottleneck*. In order to keep the information about the COMBI constructs up to date, eMate app sends the user questions. The answers are used to update the COMBI values. The intervention then targets the bottlenecks as determined by the reasoning process. That is, the user will receive motivational and informative messages related to these factors. Messages consist of three parts: a summary of the user's behavior based on the results of the monitoring, a motivational message targeting the bottleneck, and a concluding remark. The user thus gets an update of the relation between his goals and his actual behavior, in addition to personalized persuasive messages.

In this study, an adapted version of the eMate system was used to coach people on regularly taking the stairs. For this purpose, the messages and questions related to physical activity were refined and the coaching on the domains of medication intake and healthy diet was switched off.

10.4 Methods

10.4.1 Participants

The coachees were students of the VU University Amsterdam in the Netherlands, aged between 15 and 30 (average: 21 years). There were 41 male coachees and 41 female coachees. Their average Body Mass Index was 22; there were 3 underweight coachees and 10 overweight coachees, of which 1 was obese. They were recruited by handing out flyers and upon agreeing to participate in the experiment they were assigned to a condition randomly (see Section 10.4.2). The criterion for participation as a coachee was having a smartphone running Android, version 2.3.3 or higher.

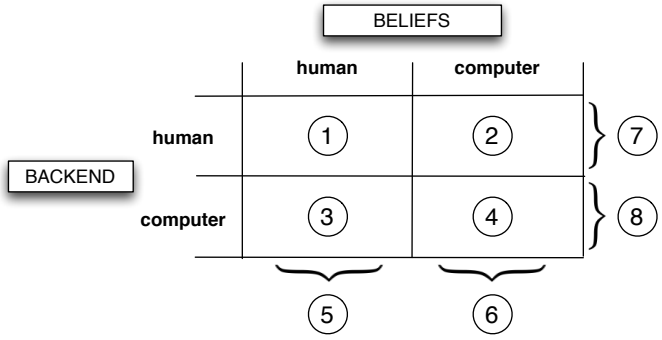

Figure 10.1: Overview of the experiment conditions.

The coaches were recruited at Utrecht University and were between 19 and 34 years old (average: 29 years). There were 9 male coaches and 11 female coaches. Overall the coaches had zero to some previous experience with coaching (65% and 35%, respectively). Both coaches and coachees were rewarded €10 for their efforts at the end of the study.

10.4.2 Design

The experiment consisted of a 'Wizard of Oz' setup: coachees received coaching from either eMate or a human coach, but in some conditions the coachees were deceived about the identity of their coach. There were four conditions in total. Each condition can be characterized by a tuple $\langle BACKEND, BELIEFS \rangle$, where BACKEND is the actual identity of the coach ($\{computer(c), human(h)\}$) and BELIEFS is the coachee's belief about the identity of the coach ($\{computer(c), human(h)\}$). The computer in this case is the eMate system. The conditions are:

1. In condition 1 $\langle h, h \rangle$ coachees were coached by a human coach and knew that this was the case. This condition represents standard e-coaching practices, where people stand in coach-coachee relations mediated by technology.
2. In condition 2 $\langle h, c \rangle$ coachees were coached by a human coach but made to believe that they were being coached by the eMate system.
3. In condition 3 $\langle c, h \rangle$ coachees were coached by eMate but made to believe they were being coached by a human coach.
4. In condition 4 $\langle c, c \rangle$ coachees were coached by eMate and (correctly) believed this to be the case. This condition represents regular interactions with eMate.

In addition to the four conditions, four other (partly overlapping) groups can be distinguished. Groups 5 and 6 are subgroups of the total sample based on the coachees' beliefs. That is, group 5 consists of everyone who believed that their coach was human (comprising conditions 1 and 3), and group 6 consists of everyone who believed that their coach was eMate (comprising conditions 2 and 4). Groups 7 and 8 are grouped by the actual coaching backend (human or eMate). See Figure 10.1 for an overview of all the groups.

10.4.2.1 Gender neutral names for coaches

In the experiment, all coachees received messages through the same Android app. To successfully keep up the deception, these messages were signed either with 'your eMate coach' or with a human name. To prevent the introduction of a gender-related dimension into the experiment, names were chosen that could be both for a male or female coach. These names were collected

Table 10.1: Used validated surveys to determine COMBI determinants

COMBI determinant	**Survey**	**Reference**
skills	Utrecht Proactive Coping Competence List (shortened) (UPCC)	Bode, Thoolen, and de Ridder (2008)
self-efficacy	Self-Efficacy Scale	Bandura (2006)
mood	Subjective Exercise Experiences Scale (SEES)	McAuley and Courneya (1994)
coping	Coping Inventory for Stressful Situation (CISS-21)	Endler and Parker (1999)
motivation	Exercise Self-Regulation Questionnaire (SRQ-E)	Ryan and Connell (1989)

in a small, online survey (N=17) that asked participants to judge whether a particular name was more likely to fit a male or a female person, or both. The names that received more than 7 neutral votes were used.[1]

10.4.3 Materials

The setup included a website (which provided coaches with a web form for writing messages and coachees with information about their stair climbing efforts), an Android widget for counting stairs (coachees only), an Android app for receiving questions and messages (coachees only), the eMate coaching engine, a database backend, and an email script for notifying coaches about their coaching tasks.

Coachees as well as coaches were presented with surveys both at the start and at the end of the experiment.[2] The surveys are described below.

10.4.3.1 Coachee intake survey

The survey at the start of the experiment took about 15 minutes to fill out and covered the following areas: demographics (age, gender, education level, length, weight); stages of change in regard to taking the stairs; knowledge, attitudes and habits of stairs use; and enabling and disabling factors for taking the stairs.

The majority of the questions were designed to establish the current stage of change of the coachee and to measure the values of the concepts in the COMBI model. Coachees were first asked a general question about exercise: whether they thought that they met the Dutch standard of exercise (30 minutes a day for at least 5 days a week). In order to determine the stage of change, questions were asked using a yes-no format in an algorithm adapted from Reed, Velicer, Prochaska, Rossi, and Marcus (1997), which is considered the most reliable and valid way to determine stage of change in the exercise domain. The questions were modified to focus on taking the stairs rather than on general exercise behavior.

To establish the values for the determinants of the COMBI model, several validated questionnaires were used. An overview can be found in Table 10.1. For those determinants that are not in Table 10.1, either no validated questionnaire was available, or it was judged too lengthly (and therefore costly). Instead, questions for the remaining determinants were mostly taken from the original COMBI intake survey (Klein et al., 2013).

[1] These names were: Dominique, Robin, Mattie, Rene, Gabriele, Sasha, Henny, Jo, Ilja, Dani, Charlie, Marijn, Jamy, Kris, Jip, Jorin, Beau.

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

10.4.3.2 Coachee evaluation survey

At the end of the experiment, after 4 weeks of monitoring stairs use, the coachees were admitted an evaluation survey. The objectives of this survey were a) to observe if there was any change in the coachee's knowledge regarding stair climbing and physical activity; b) to observe if there was any change in the coachee's attitudes towards the health benefits associated with stair climbing; and c) to monitor if there was any change in the coachee's behavior with regard to taking the stairs.

The evaluation survey addressed again the stage of change and the COMBI determinants. For most determinants, shortened versions of the surveys mentioned in Table 10.1 were used. This was necessary to restrict the length of the survey, as some additional surveys were incorporated with the purpose of gaining insights into matters of user experience. Also, some questions were added that addressed the deception component of the experiment, which is discussed in Section 10.5.

The first of the two incorporated validated surveys is the measure of human-computer trust (HCT; Madsen and Gregor (2000)). In Madsen and Gregor (2000), human-computer trust is defined as "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid." In this study, a psychometric instrument for measuring human-computer trust was developed and tested ($\alpha = 0.94$), that focuses on both cognitive and affective aspects of trust. The measure was specifically designed for intelligent systems that aid decision making and consists of five subscales: reliability, understandability, technical competence and faith. The second added survey is the Sport Climate Questionnaire (SCQ; Mageau and Vallerand (2003)), which measures perceived autonomy support.[1]

10.4.3.3 Coach surveys

Coaches were admitted short intake and evaluation surveys. The main objective of these surveys was to obtain knowledge of the coach's (i) demographics, (ii) previous coaching experience, and (iii) coaching style.

10.4.4 Procedures

10.4.4.1 Procedure for coaches

The coaches received a flyer with information about the experiment and their tasks. Together with an experimenter, each coach registered an account at the experiment website and was subsequently guided through the steps of performing their coaching task (they had one practice task). It was explained that for the duration of 4 weeks, they would be coaching other people by writing short, motivational messages in an online form. This form would guide the coaches in making suitable messages, by giving the coaches a relevant piece of information about taking the stairs (e.g., "Taking the stairs activates about 200 muscles" or "When taking the stairs at a slow pace you burn 3 times as many calories as you would when walking regularly"). The coaches were asked to use this information to compose their messages. Because of this, and because the length of their messages was restricted to the size of a computerized message, it was ensured

[1]There is evidence that autonomy-supportive coach behavior is associated with positive cognitive, affective and behavioral outcomes Mageau and Vallerand (2003). Due to the limited scope of this paper, the results of this survey will not be discussed here.

that the information value of the messages of both computerized coaches and human coaches was similar.

Every coach was paired with one coachee from condition 1 and one from condition 2. Coachees received two messages per week for three weeks (after the initial week of monitoring), so each coach was asked to write a total of four messages per week. The coaches received automated requests by email to write their coaching messages in the online web form. Coach compliance was very high: of the 240 messages in total, only 3 were missed. Coaches were instructed to write their messages on Monday and Wednesday, so that coachees from all conditions received their coaching messages on Tuesday and on Thursday.

10.4.4.2 Procedure for coachees

Coachees were divided into the different experimental conditions (1,2,3,4) in order of admission. Coachees were given a different flyer containing information about their coach, the setup of the experiment and the usage of the widget and the eMate app. This information was also repeated verbally. It was explained that after one week a goal would be set for them that would be based on the data from that first week. They would then receive coaching for a period of three weeks to reach that goal. Together with the experimenter, the coachees downloaded the app and the widget from Google's Market Place, and registered on the eMate website (this account was needed for using the apps).

The coachees were then asked to fill out the intake questionnaire. The coachees were also notified that at the end of the experiment, they would be admitted a final (online) questionnaire and that their reward depended on filling out that final questionnaire.

Coachees were asked to tap the widget every time they climbed a flight of stairs. In case no data was registered for a full day, the system automatically posed a question via the app about the number of stairs taken during the previous day. They received motivational messages twice a week.

10.5 Results

10.5.1 Three data sets

For the analyses, three sets of data were used. Data set *A* (N=74) groups coachees on the basis of the experiment condition that they were assigned to (1 to 4). It ignores coachees who failed to fill out the final questionnaire. Because of the method of assignment, the group sizes are fairly similar (17 to 21). Data set *A* is used for all analyses concerned with differences in the entire group before and after the coaching interventions. Data set *B* (N=64) was prepared by taking the coachees' reported belief about their coach into account. Recall the groups from Figure 10.1. Someone who was assigned to condition 1 but reported a belief that his or her coach was a computer, was for all practical purposes in a similar state as coachees in condition 2. Because the study focuses on the effects of the belief about one's coach, we regrouped a number of people based on their reported beliefs (see Section 10.5.2). In addition, we removed a number of coachees who reported to have no idea whether they were coached by a person or a computer. Data set *B* is used for analyses about differences between coachees in the different conditions. Finally, data set *C* (N=70) comprises all coachees for whom data was collected from the stair climbing widget. People who stopped before the end of the experiment (i.e., people who didn't use the widget during the last 7 days of the experiment: 5 people) and those who didn't regularly use the widget (i.e., people who missed more than 40% of the days: another 7 people)

Table 10.2: Beliefs about the coach's identity

Group	B(coach == eMate)	B(coach == human)	Don't know
1	4	11	2
2	16	0	3
3	3	16	2
4	14	0	3

were excluded. The grouping of coachees from data set *B* is also used in data set *C* (i.e., the condition is determined by the reported belief). Data set *C* is used for analyzing the widget data.

10.5.2 Beliefs about coaches: the success of the deception

For the experiment to yield valid results, it was crucial to establish that the deception was successful. To check this, coachees were asked in the evaluation survey whether they thought they were being coached by eMate or by a human coach. The results are in Table 10.2. It shows that in the deception conditions (2 and 3), only 3 people were not fooled (14.29%; in condition 3).[1] Table 10.2 also shows a number of people who were unsure or just did not remember. These were distributed equally among the 4 conditions. Given that the number of skeptical people in condition 1 was equal to that of condition 3, there are no indications that coachees in one of the deception conditions were aware of the deception. We therefore conclude that the deception was successful. If anything, the data shows a shared skepticism across conditions about the participation of human coaches in the experiment; the only people who doubted the identity of their coach thought it was a computerized one instead of a human one.

10.5.3 Descriptives

The coachees were asked in the evaluation survey to give their opinion about the different aspects of the system. When asked to describe the messages, coachees (from data set *A*) found the messages motivating (28%), clear (41%), and informative (41%). The main negative comment about the messages was that they could be too childish (14%). There was no significant difference between groups in data set A or B. Another set of items asked coachees to what extent they found the different aspects of the system motivating (using a 5-point Likert scale going from 'very much so' to 'not at all', later recoded as values from 1 to 5 (3 = 'somewhat')). As can be seen in Table 10.7, the use of the widget was clearly perceived as the most motivating aspect of the system (median answer: 'quite'), followed by the coaching messages from the app (median answer: 'somewhat'). There was no significant difference between groups in data set A or B.

10.5.4 Effectiveness of the intervention

An important outcome measure of the study was the effect of the intervention. Our hypothesis was that the belief coachees have about the identity of their coach (human or computer) will have no effect on effectiveness. Moreover, we had an additional hypothesis that the intervention would have a positive effect on the number of stairs that coachees climbed.

[1] To construct data set *B*, these 3 people from condition 3 — who rightly believed they were being coached by eMate — were placed in condition 4. In addition, the 4 coachees who believed that they had a computerized coach in condition 1 were placed in condition 2. People who did not report a belief about their coach's identity were excluded.

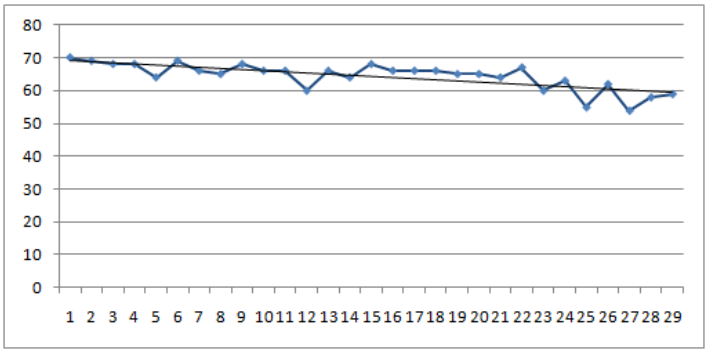

Figure 10.2: Number of users using the widget per day.

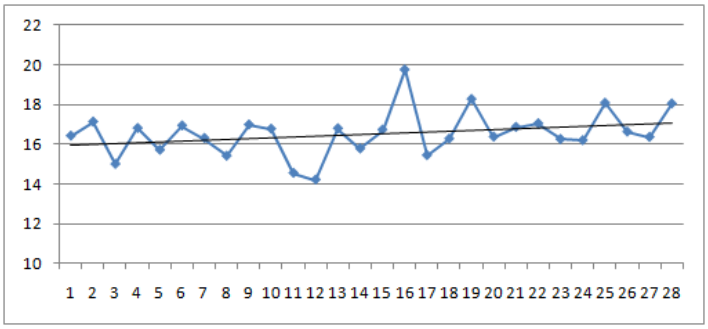

Figure 10.3: Average number of stairs climbed per day for the total sample.

10.5.4.1 Completeness of the data

To analyze the effect of the intervention, we used the data that was collected by the widget (data set *C*). The widget data combined with reminders resulted in a fairly complete set of data of the number of taken stairs each day. The number of active users per day (those who registered at least 1 stairs event) only slightly decreased during the course of the experiment (from 69 to 59) as can be seen in Figure 10.2.

10.5.4.2 Total sample

To be able to compare the effect sizes of the different groups, we examined the effect on the total sample. A graphical representation of the average numbers of stairs climbed per day is presented in Figure 10.3. A slight increase during the intervention is visible.

We define the effectiveness as the increase in the number of stairs climbed in the last week compared with the first week. The first week starts on the day after the inclusion. The mean number of stairs taken for the total sample is 16.29 (SD=8.31) in the first week and 16.80 (SD=8.87) for the last week. This difference is not significant (paired t-test, p-value = 0.69). It has to be noted that there are large differences between the average numbers of stairs climbed for the individual coachees. For example, for the first week, the person with the highest average climbed 58 stairs per day, while the person with the lowest average only climbed 4.2 stairs per day.

10.5.4.3 Analyses

The mean number of stairs climbed in the first and last week of the experiment for all different groups is listed in Table 10.3. Except for group 3 and 5, all groups showed an increase. The

Table 10.3: Mean of stairs climbed

Group	Mean first week	Mean last week	Difference	N
1	19.05	**19.07**	+0.02	12
2	12.18	**13.71**	+1.53	16
3	18.50	17.53	-0.97	16
4	17.19	**18.87**	+1.68	16
5	18.73	18.19	-0.54	28
6	14.68	**16.29**	+1.61	32
7	15.12	**16.01**	+0.89	28
8	17.84	**18.20**	+0.36	32

difference between group 5 (belief human) and group 6 (belief computer) looks interesting, but is not significant (t-test, p-value = 0.36)

The relatively low number of stairs climbed by coachees in group 2 draws attention: there is a significant difference with the first week data of group 1 (t-test, p-value = 0.045) and with group 4 (p-value = 0.014). This finding is further discussed in Section 10.6.

10.5.4.4 Self-reports on taking the stairs

Another way to assess the effectiveness of the intervention is by examining the answer given to the question in the surveys about the number of stairs taken per week. The answers were recoded to values from 1-5, where a higher score represents more stairs taken (1 point is a 10 stairs minimum increase). A paired t-test was used to compare means between before and after surveys for the total sample as well as for each group. We found a consistent, but non-significant increase in the mean for the total sample (of approx. 0.2 points). Furthermore, there were no significant differences for any of the groups (in either data set).

10.5.5 Perceived influence of the intervention

Another measure was defined as the extent to which people found that the intervention contributed positively to their behavior. Our hypothesis was that there will be no difference in how influential people perceived the intervention to be.

10.5.5.1 Total sample

Of the 74 coachees who filled out the evaluation survey (data set *A*), 37.8% thought the intervention did not have a positive impact on their behavior. 32.4% were ambivalent, and 29.7% thought they were influenced positively. For the analysis of this question, the Likert-like item was recoded to numerical values from 1-5, where a higher value represents a more positive answer. The mean answer for all coachees was 2.85 (SD=1.06). Table 10.4 shows the means for all groups in *A* and Table 10.5 shows the means of all groups in *B*.

At face value, both tables suggest that the people who believed that they were being coached by a human, gave slightly more positive answers (see groups 1, 3 and 5).

10.5.5.2 Analyses

A Wilcoxon rank sum test was performed on data set *B* for groups 5 (N=27) and 6 (N=37). This showed a difference that was significant at the $p<0.05$ level ($W = 648$, p-value = 0.036). This finding suggests that people who believe that their coach is human indeed report slightly

Table 10.4: Means Set A

Group	Mean	SD	N
1	3	0.87	17
2	2.68	1.06	19
3	3.05	1.20	21
4	2.65	1.06	17
5	3.03	1.05	38
6	2.67	1.04	36
7	2.83	0.97	36
8	2.87	1.14	38

Table 10.5: Means Set B

Group	Mean	SD	N
1	3.18	0.60	11
2	2.5	0.89	20
3	3.12	1.20	16
4	2.65	1.17	17
5	3.15	0.99	27
6	2.57	1.01	37
7	2.74	0.86	31
8	2.88	1.19	33

Table 10.6: Transitions in stages of change

stage of change	**start number**	**end number**	**change**
Precontemplation (PC)	6	7	+1
Contemplation (C)	10	5	-5
Planning (P)	2	0	-2
Action (A)	3	4	+1
Maintenance (M)	53	58	+5

more positively about the influence of coaching. To rule out that this difference was related to a difference in backend, Wilcoxon rank sum tests were also performed on the subgroups of 5 and 6, namely 1 and 3, and 2 and 4, respectively. These tests did not yield any significant differences, which is evidence for homogeneity within groups 5 and 6.

10.5.6 Effect Measure: Stage of Change

We expected that in all experimental conditions there would be an overall increase in stage of change. To measure this, we examined data set A with respect to the stage of change questions from the survey. The results are in Table 10.6. Two observations can be made from Table 10.6. First, the vast majority of the coachees started out and ended up in the Maintenance stage (72% and 78%, respectively). As such, only 21 people could improve their stage. Second, the overall trend is that people move from stages of non-compliance (PC,C and P) to stages of compliance (A and M). After recoding the stages to numerical values, where PC $= 1$ and M $= 5$, the average stage at the start of the experiment was 4.18 and at the end 4.37 (a non-significant increase of 0.19). There was no significant difference in the pattern of change between the experimental groups (1-4 and 5-8).

There were 14 coachees who changed their answer to the question 'Do you often choose to take the stairs instead of the elevator/escalator?' (answer: 'yes' or 'no') from the intake to the evaluation questionnaire: 4 of 56 coachees changed for the worse (from 'yes' to 'no'), and 10 of 18 changed for the better (from 'no' to 'yes'). Again, these results were not significant and there was no significant difference between the experimental groups.

10.5.7 Trust

We also examined the influence of the identity of the coach on trust. Our hypothesis was that the belief coachees have about the identity of their coach (human or computer) will have no effect on trust. To test this hypothesis we analyzed the survey answers from the HCT scale. For this analysis, the separate subscales of reliability, competence, understanding, faith and attachment

Table 10.7: Motivational aspects

Scale item	Mean	SD
coach messages on app	2.11	0.96
questions on app	1.99	0.92
use of widget	3.32	0.97
use of website	1.73	0.87
participation of others	1.97	1.13
total motivating aspects	2.22	0.57

Table 10.8: HCT scale means

Scale	Mean	SD
reliability	3.05	0.67
competence	3.03	0.66
understanding	3.69	0.69
faith	2.73	0.65
attachment	2.60	0.77
total HCT	3.02	0.48

Table 10.9: Highest and lowest rated items on the subscales (data set A)

rating	**subscale**	**item**	**mean**
highest	reliability	The system responds the same way under the same conditions at different times.	3.46
	competence	The system correctly uses the information I enter.	3.24
	understanding	I know what will happen the next time I use the system because I understand how it behaves.	3.97
	faith	If I am not sure about a decision, I have faith that the system will provide the best solution.	2.97
	attachment	I would feel a sense of loss if the system was unavailable and I could no longer use it. I find the system suitable to my style of decision making.	2.65
lowest	reliability	The system always provides the advice I require to make my decision.	2.28
	competence	The advice the system produces is as good as that which a highly competent person could produce.	2.54
	understanding	I recognize what I should do to get the advice I need from the system the next time I use it.	3.47
	faith	When I am uncertain about a decision I believe the system rather than myself.	2.51
	attachment	I have a personal preference for making decisions with the system.	2.46

were analyzed, as well as the individual items (5 questions per subscale). The answers to the questions were on a 5-point Likert scale from strongly agree to strongly disagree, which were recoded to values from 1 to 5 (3 = 'don't agree, don't disagree'). The group means are shown in Table 10.8.

In data set *A* we found that the overall trust value for the system is average (3.02). The values for the subscales faith and attachment have the lowest means (2.73 and 2.60, respectively), while the subscale understanding had the highest average mean (3.69). The items that were rated highest and lowest within the subscales were identified and can be found in Table 10.9. For checking for differences between groups, data set *B* was used. None of the subscales of trust showed a statistical difference between the groups (t-test).

10.6 Discussion

In this section the experimental results are discussed, together with additional reflections on the experiment as a whole.

Hypothesis 1 was addressed in Sections 10.5.4 and 10.5.6. Both the analyses of the widget data and of the survey data concerning the stage of change showed that there was no effect from the coachee's belief about coach identity on the effectiveness of the interventions. This confirms hypothesis 1. Although overall the coachees improved — both with respect to the widget data and their stage of change — these improvements were non-significant. We found a consistent, but non-significant increase in the mean for the total sample (of approx. 0.2 points) and no differences between groups. This is consistent with what was expected and with the widget data. However, it should be mentioned that the questions in the survery might be too coarse-grained to report small differences, as the scale increments by 10 stairs. So, someone who would have taken 9 stairs more than prior to the experiment would have given the same answer.

We have no ready explanation for the significantly lower number of stairs climbed in the first week by coachees in group 2 compared to groups 1 and 4. Our hypothesis is that this difference is caused by large differences between individual coachees.

Contrary to hypothesis 2, coachees did judge the intervention to have more positive influence if they believed to be coached by a human coach. According to O'Keefe, the extent to which a recommendation influences its receiver depends on four main characteristics: source characteristics, message characteristics, target characteristics, and context characteristics (O'Keefe, 2002). As effort was made to keep the message and context characteristics very similar, and the target characteristics were evened out by the random assignment to the different conditions, source characteristics are the most likely cause of this discrepancy. As defined in Yoo, Gretzel, and Zanker (2013), source characteristics in communication can be shaped by several cues, among which similarity, authority, and familiarity. One possible explanation of the found discrepancy is that people feel more similar to another human than to a computer and therefore attribute greater competence or trust to them. However, past empirical studies show contradicting results with respect to similarity and effects on perceived trustworthiness (see Yoo et al. (2013)). Moreover, our analysis of trust in Section 10.5.7 showed no difference in trust between groups. A more plausible explanation is that people are more familiar with humans than computers in a coaching role, and that this shapes their evaluation of the system. Additional work is needed to support this hypothesis.

Hypothesis 3 was confirmed, since no difference was found between the groups with regard to trust. Concerning the use of the eMate system, it can be concluded that overall the coachees felt that the system was averagely trustworthy. Their cognition-based trust was rated higher than their affect-based trust: coachees felt they had a good understanding of the system, but their attachment to the system and faith in its capabilities was somewhat lacking. The coachees were able to form a good mental model of the system's behavior and were able to predict future behavior, but they had only a moderate preference for the system and were not entirely convinced that the system would be able to perform in untried situations.

We will conclude this section with some additional reflections. First, we would like to note that some researchers have suggested that when evaluating persuasive and pervasive systems, not just trust but 'credibility' should be examined. Credibility consists of a combination of expertise and trustworthiness (e.g., Cialdini and Rhoads, 2002; Fogg, 2003; O'Keefe, 2002). The dimension of expertise captures the perceived knowledge and skill of the source Mayer, Davis, and Schoorman (1995); O'Keefe (2002) while trustworthiness of a source refers to aspects such as character or personal integrity (O'Keefe, 2002). Although we use the term 'trust' in this paper, we would like to argue that in this particular case the findings also relate to expertise (and therefore credibility), as one subscale of the HCT measures competence and examines whether the system is perceived to perform its tasks accurately and correctly based on the input information.

Interestingly, most coachees were assigned the stage 'maintanance' based on their self-reports, which made it hard to improve their stage of change. One explanation is that all coachees always chose stairs over elevators, and needed little encouragement to take the stairs more often. Another possible explanation is that the questions used to determine the stage of change did not completely match the spectrum of the different stages. This explanation is corroborated by the fact that analysis showed that (i) some coachees changed from the Contemplation to the Maintanance stage (which is strictly not possible since the Maintanance stage requires one to perform compliant behavior for at least the past six months), and (ii) some coachees relapsed to the Precontemplation stage (which also seems unlikely because this indicates that they became unaware of the health benefits of taking the stairs). In the literature there is no standardized answer to the question of how the people should be assigned to stages of change and what empirically validated surveys are most suitable (e.g., Sutton, 2001). In future work an effort could be made to compare different surveys to determine stage-of-change and to validate their outcomes in the COMBI model.

As mentioned in Section 10.2, coachees were paired with their coaches for the duration of the experiment, leading to repeated interactions regardless of prior experiences. One might object that this setup does not reflect everyday situations in which people enable apps, only to disable them thirty minutes later because they are not satisfied. However, it does simulate the situations where people give coaching — either human or computerized — a serious try because they are motivated to achieve a certain goal that they feel they cannot achieve without help.

Finally, in Section 10.5, we suggested that there might have been a general skepticism towards human involvement. This might be explained by the fact that the questions that all coachees received were repetitive and obviously (and transparently) computer-generated. While this aspect of the experiment had been explained to the coachees, it might have contributed to some people's skepticism.

10.7 Conclusion and Future Work

This paper has described a month-long experiment of e-coaching in the wild: people received coaching outside of the lab for an activity with known health benefits. The coaching messages that coachees received were either computer-generated or written by a human coach, but they were sometimes purposefully given the wrong information about the identity of their coach (human or computer). The underlying model that was used to identified bottlenecks to people's behavior change is part of the fully autonomous e-coaching system eMate. eMate also facilitated all interfacing between coachees and coaches (i.e., messages, questions, progress overview).

From the results we conclude that the belief that coachees had about the identity of their coach had no effect on the effectiveness of the coaching. Although research is needed to establish that this results holds in different domains, it is a first indication that e-coaching can be successful without the (transparent) involvement of a human coach. Secondly, we conclude that people do have a bias towards human coaching with regard to judging how much positive influence coaching had on their behavior. To explain this, we have hypothesized that people are more familiar with humans than computers in a coaching role. This hypothesis will have to be tested in future work. Lastly, people showed no difference in trust towards human coaches or computerized coaches. Overall, people judged both types of coaches as averagely trustworthy. This is an important finding, as trust is the key ingredient of successful coach-coachee relationships (Gyllensten and Palmer, 2007). Moreover, this judgement is further evidence that, at least with regard to trust, people do treat computers as social actors.

Acknowledgment

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 helpful comments and her contributions to the content of the messages.

References

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares and T. Urdan (Eds.), *Self-efficacy beliefs of adolescents*, Adolescence and Education, pp. 307–337. Information Age Publishing.

Bickmore, T. W. and R. W. Picard (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions of Computer-Human Interaction 12*(2), 293–327.

Blount, S. (1995). When social outcomes aren't fair. *Organizational Behavior and Human Decision Processes 63*(2), 131–144.

Bode, C., B. J. Thoolen, and D. T. D. de Ridder (2008). Het meten van proactieve copingvaardigheden. Psychometrische eigenschappen van de utrechtse proactieve coping competentie lijst (UPCC). *Psychologie & Gezondheid 36*, 81–91.

Cialdini, R. and K. Rhoads (2002). The business of influence. In J. Dillard and M. Pfau (Eds.), *Persuasion handbook: Developments in theory and practice*, pp. 513–542. London, UK: Sage.

van Dongen, K. A. and P.-P. van Maanen (2006). Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. In *Proceedings of the Human Factors and Ergonomics Society's 50th Annual Meeting*, Volume 50, San Francisco, CA, pp. 225–229.

Endler, N. and D. Parker (1999). *Coping Inventory for Stressful Situations (CISS): Manual (2nd ed.)*. Toronto: Multi Health Systems. Dutch translation by De Ridder D. T. D. and Maes, S.

Fogg, B. J. (2003). *Persuasive Technology: Using computers to change what we think and do*. San Francisco, CA: Morgan Kaufmann Publishers.

Fogg, B. J. and C. Nass (1997). Silicon sycophants: The effects of computers that flatter. *International Journal of Human-Computer Studies 46*(5), 551–561.

Gyllensten, K. and S. Palmer (2007). The coaching relationship: An interpretative phenomenological analysis. *International Coaching Psychology Review 2*(2), 168–177.

Klein, M., N. Mogles, and A. van Wissen (2013). An intelligent coaching system for therapy adherence. *IEEE Pervasive Computing 12*(3), 22–30.

Klein, M. C. A., N. Mogles, and A. van Wissen (2011). 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). Lecture Notes in Computer Science*, Volume 7065, pp. 104–116. Springer Verlag.

Kool, L., J. Timmer, and R. van Est (2013). Keuzes voor de e-coach: Maatschappelijke vragen bij de automatisering van de coachingspraktijk. Technical report, Den Haag, NL.

Lee, J. R. and C. I. Nass (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In D. Latusek and A. Gerbasi (Eds.), *Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives*, pp. 1–15. IGI Global.

Madsen, M. and S. Gregor (2000). Measuring human-computer trust. In *Proceedings of the 11th Australian Conference on Information Systems*, pp. 6–8.

Mageau, G. A. and R. J. Vallerand (2003). The coach-athlete relationship: A motivational model. *Journal of Sport Sciences 21*(11), 883–904.

Mayer, R. C., J. H. Davis, and F. D. Schoorman (1995). An integrative model of organizational trust. *Academy of Management Review 20*(3), 709–734.

McAuley, E. and K. S. Courneya (1994). The subjective exercise experiences scale (SEES): Development and preliminary validation. *Journal of Sport and Exercise Psychology 16*(2), 163–177.

Nass, C. and Y. Moon (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues 56*(1), 81–103.

Nass, C., Y. Moon, and P. Carney (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology 29*(5), 1093–1110.

O'Keefe, D. J. (2002). *Persuasion: Theory and research*. Thousand Oaks, CA: Sage Publications.

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

Pruitt, D. G. and P. J. Carnevale (1993). *Negotiation in social conflict*. Belmont, CA: Thomson Brooks/Cole Publishing Co.

Reed, G. R., W. F. Velicer, J. O. Prochaska, J. S. Rossi, and B. H. Marcus (1997). What makes a good staging algorithm: Examples from regular exercise. *American Journal of Health Promotion 12*(1), 57–66.

Reeves, B. and C. Nass (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places (CSLI Lecture Notes Series, no. 63)*. Center for the Study of Language and Informatics.

Ryan, R. M. and J. P. Connell (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology 57*(5), 749–761.

Sutton, S. (2001). Back to the drawing board? A review of applications of the transtheoretical model to substance use. *Addiction 96*(1), 175–86.

Warner, T. (2012). E-coaching systems: Convenient, anytime, anywhere, and nonhuman. *Performance Improvement 51*(9), 22–28.

van Wissen, A., Y. Gal, and M. V. D. B. A. Kamphorst (2012). Human-agent teamwork in dynamic environments. *Computers in Human Behavior 28*(1), 23–33.

van Wissen, A., J. van Diggelen, and V. Dignum (2009). The effects of cooperative agent behavior on human cooperativeness. In *Proceedings of AAMAS*, pp. 1179–1180.

Yoo, K.-H., U. Gretzel, and M. Zanker (2013). *Persuasive Recommender Systems: Conceptual Background and Implications*. SpringerBriefs in Electrical and Computer Engineering. Springer.