TRUST IN HUMAN-COMPUTER RELATIONSHIPS:

Do cross country skiers have trust towards a physical intelligent tutoring system as an accurate feedback on performance?

Karolina Thorsén, Anna Lindström
We would like to thank Juan Carlos Nieves for supervising us through the thesis. Esteban Guerrero Rosero for making changes in the existing physical intelligent tutoring system, and Roger, Joel and Stefan at Sport Sciences Lab and Umeå School of Sport Science (USSS) for assistance when performing the tests. And of course, the 6 participants who decided to participate in the study. Thank you for being a part of making our thesis possible!
Abstract

Trust is one of the attitudes that can affect the intentions and the behavior of a human using a system. Misusing a system can have safety- as well as economic consequences, this is why it is important that a user develops calibrated trust towards a system. In this report, the research question is: how much trust cross country skiers have towards a physical intelligent tutoring system (PITS)? Six biological males (age 24 to 50) roller skied on a mechanical treadmill and received feedback from the PITS on a TV-screen. Experience of using the PITS was evaluated with an instrument in a semi-structured interview. The instrument measured the participants overall perceived trust (OPT) for the system, and the participants were asked to further their thoughts about the statements. The data was transcribed, coded, and categorized in a thematic analysis. The result showed that a majority of the participants had low OPT for the PITS, and the thematic analysis showed that the minority with higher levels of OPT focused on the choice of an elite skier as the reference skier. One of the problems with the instrument was that it was developed for evaluation in long term usage, and not first time usage as in this study. The result of this report can be used for further development of the PITS and a reminder on why trust needs to be considered when creating user experiences.

Keyword: trust, autonomous system, human-computer-trust, HCT, physical intelligent tutoring system, PITS, user experience, UX, cognitive science

Abstrakt


Nyckelord: tillit, autonoma system, människa-dator-tillit, fysiskt intelligenta handledningssystem, FIHS, användarupplevelse, UX, kognitiv vetenskap
Trust in human-computer relationships:
Do cross country skiers have trust towards a physical intelligent tutoring system as an accurate feedback on performance?

To reach expert levels of performance, extensive training is not enough. People who has spent more time on deliberate training compared to those focused on general training, are often the best performers. Deliberate training assumes that achieving expert level of performance takes time and improves by receiving suitable training, preferably designed by a coach or a teacher (Ericson, 2006). With the digitalization happening in society, the role of a coach can also be automated to contribute to deliberate training (Goldberg, 2016). With that in mind, more individuals could be able to reach their full potential. For tutoring of cognitive learning without humans, an intelligent tutoring system (ITS) has been used in the academic setting (Anderson, Boyle, & Reiser, 1985). From ITS, the concept of physical intelligent tutoring system (PITS) was developed. PITS focus on tutoring users in the psychomotor domain by giving feedback to improve performance. A user performing a physical task will get feedback from the PITS on e.g. their technique in double poling. In this way, a user can improve their skills and performance (Goldberg, 2016). For PITS, there are great challenges when it comes to capture and processing data for behavioral information related to certain domains of skills (Goldberg, 2016).

In a study performed at Umeå University by Guerrero, Nieves, Sandlund, Lindgren, and Söderström, (2017), a prototype for a PITS (Nordic Telemedicine Center, 2017) has been used for improving cross country skiers double poling technique. The Guerrero et al. (2017) PITS has two different usage area, it could either be used by cross country skiers to get instant feedback as a substitute for a human coach, or be used by human coaches to analyze the skiers performance after the session. The area in this study, was the usage of the cross country skiers instant feedback. The PITS is based on inputs from a Microsoft Kinect camera that collects information about the movements and angles of a users limbs. A machine learning algorithm compares and analyzes the users double poling movements in real time to the double poling movements of a professional skier, that has previously been recorded for the PITS. In this way, the system can calculate and give feedback on how similar a user's technique is to the reference elite skier. Output is then visualized on a screen in an augmented reality (AR) version in real time. The screen allows the user to see their limbs represented as a stick figure and receive feedback about their performance in comparison to a reference elit skier, see Figure 1.
In sports, researchers have taken advantage of the benefits of using AR as a feedback tool for enhancing the performance of athletes techniques as well as avoiding injuries (Onate et al., 2001). The widely accepted definition for AR describes it as adding virtual elements to the user's real world as another layer enhancing the opportunities to gain information and interact (Azuma, 1997).

With the digitalization in society, less collaborations in daily life are constrained to being human-to-human. A user being introduced to a digitalized coach, will create a human-to-computer relationship. In a good human-to-computer relationship, the computer is dependent, predictable and will perform as expected (Lee & See, 2004). There are factors in human-to-computer relationships that also can be found in relationships between humans, such as trust. In this report the focus is on measuring the amount of trust users have to the PITS.

There have been a lot of studies in recent year about trust, in both human-to-human and in human-to-computer relationships. Studies have reached different conclusions on the nature of trust, it has been defined as either an attitude, a behavior, an intention or a belief (Lee & See, 2004). To unite these competing definitions, a framework was established by Ajzen and Fishbein (Lee & See, 2004). This framework states that experiences and available information creates beliefs, and emotional evaluations of these beliefs lead to attitudes which evolves from cognitive functions where analytic, analogic and affective processes work together. These attitudes lead to intentions that create certain behaviors. As an attitude, trust can guide the users reliance and compliance on an agent, and work as a social decision heuristic under uncertainty (Lee & See, 2004). Reliance is when the user trusts the system that no action is needed since no feedback is given by the system to imply that. Compliance is when the user find the information given by the system trustworthy enough to follow.
Although, it should be noted that trust is one of many factors that might affect a behavior. Sometimes the reason for a person not following through her intentions is because of external factors such as limitations of time or challenges of using for, e.g. a certain system (Lee & See, 2004).

In this report, trust is defined as an attitude where the users are confident in that an agent, which can be a computer or a human, will help them to achieve goals also in environments and situations where they feel vulnerable from ambiguity and uncertainty (Lee & See, 2004). When trust is considered an attitude, previous direct or indirect experience similar to the situation will guide the persons trust, whether it is experienced by themselves or if they have been presented with the information (Lee & See, 2004). This will set the person with a certain level of initial trust, general for the situations they get involved in and also influence the intake of information (Lee & See, 2004).

According to Mayer et al. (1995), vulnerability is a key element for the development of trust, and trust is rather about being willing to take a risk, than to actually take a risk. So if a person is in a situation where they are not willing to take a risk, there is no need for trust.

Lewandowsky et al. (2000) also studied additional variables and their effect in combination with trust. The variables were studied together with trust and self confidence to determine if they would affect the users usage of automation. One of the variables studied was advanced knowledge of automation failure. The researcher could see how the user’s trust for the automation did not get reduced if they knew about the faults of the automation in beforehand. These findings were of importance since previous studies by Lee and Moray (1992) had showed how trust would decline with the presence of faults.

Except for using a system accordingly, a user can have poor calibration with overtrust or distrust in a system. If a user have overtrust in a system, s/he will expect the system to do more than it actually can, which will lead to misuse and overuse of the system. An example of this is the autopilot system in Tesla cars. The system can in some environments drive the car without the driver, but the driver is informed that they still need to have their attention on the road and be able to take control over the car if something unexpected happens (Tesla, 2018). Even provided with this information, some users have had overtrust in the autopilot and not providing the attention needed when using the autopilot. This overtrust has lead to car accidents that in some cases have ended fatally (Business insider nordic, 2017; CNBC, 2018; The Guardian, 2018). In the other end of the spectrum, a user can have distrust and underestimate what the system can do. This can manifest itself in disuse where the user is not using the system at all or only using a portion of its functions (Lee & See, 2004). Situations where the user can develope distrust for a system can vary. It can for e.g. be when the system does not reach the expected usability the user has based on the first impressions, or in situations where the user observe the system performing errors without explanations about why the error occurs. Knowing why errors occur has shown to being able to increase the users trust in the system (Dzindolet, Peterson, Pomranky, Pierce, Beck, 2003; Lewandowsky et al., 2000).

For optimal usage of a system, one want to have calibrated trust where the users trust is appropriate and adjusted to the functions of the system is important (Lee & See, 2004).
Figure 2, calibrated trust is present when the user’s level of trust for the system, matches the systems level of capabilities. For example, a user takes a photo with her/his phone. After the photo is taken s/he locks the screen of the phone and trust the phone to automatically save the photo as it normally does. After an update of the software, s/he notice that the phone no longer automatically saves it and s/he therefore no longer have trust for the phone to automatically save it.

To establish calibrated trust to a system, there are three expectations that needs to be fulfilled, according to one of the first theories in Trust in Automation (TiA) (Drnec, Marathe, Lukos, & Metcalfe, 2016). These are technical competence, persistence, and fiduciary responsibility and their roles will be of different importance during the users interaction stages with the system. The technical competence is in greatest importance in the first stages of interacting with the system for the first time. This is when the user rely on previous experience to guide expectations on the automation to perform what it was designed for. Persistence, could also be explained as predictability, is when the users can expect the system to reliably perform similar behavior in similar situations in the future. Fiduciary responsibility is the users expectation of the system that will impact the users allocation of tasks. When the user gain experience with the system, the user will learn what the functions of the system was designed for and what it can be expected to be responsible for. Because of this, the user will use less personal resources and allocate more responsibility on the system.

Trust is one of the attitudinal factors that will affect the intentions and finally the behavior of the user to use, disuse, misuse or to abandon a system (Lee & See, 2004). This report will study the degree of overall perceived trust cross country skiers have for the physical intelligence tutoring system (PITS) for feedback on performance, developed by Guerrero et al. (2017).

![Figure 2](image_url)

*Figure 2. The relationship between the users trust and the capability of the system, shown here as a graph. This shows that to much trust in the user versus the capability of the system creates overtrust and the reverse creates distrust. Calibrated trust is desirable, where a user's trust in the system correlates with the systems abilities (Lee & See, 2004). With regard to the example given with the phone taking photos, the user constantly retained calibrated trust by minimizing trust as the capabilities declined. This can be visualized in the graph where the staples from the trust and automation capability axis meets.*
Method

Cross country skiers were invited to try the PITS and participate in a semi-structured interview. A psychometric instrument constructed by Madsen and Gregor (2000) was used to measure users subjective overall perceived trust for the PITS. The instrument was created specific for measuring humans-computer trust (HCT) with intelligent systems in percentage and could also give information whether the overall perceived trust was based on more cognitive or affect-based aspects. In this study, the focus was to find out participants overall perceived trust and not the specific aspect. Why cognitive or affect-based could be of interest was because earlier studies have shown that affect-based aspects have had most indications for overall perceived trust (Andersson, Malm, & Thurén, 2003). The instrument by Madsen and Gregor (2000) was chosen because of its reported empirical validity and reliability.

To get as much insight into the participants thoughts as possible, considerations of both semi-structured interviews after the participants had tried the PITS, and thinking-aloud interviews during the usage of the PITS were discussed. Because a thinking-aloud interview could distract the participants in their performance and the use of the system, the choice fell on semi-structured interviews.

After the participant had used the PITS for 10 minutes, they were asked to agree or disagree to 25 statements from 5 different constructs, as well as furthering their thoughts about the statements. The statements were translated into Swedish from English to lower the barriers of understanding the statements, since the users were expected to have a higher understanding of Swedish than English. Even though the participants were encouraged to further their thoughts, there were no guarantee they would mention what improvements of the PITS they wished to see, what changes they considered as important, how much trust they considered themselves having for the PITS. It was not certain the participants would give information about if their perceived trust was similar to the trust the instrument would suggest them having, which could lead to a discussion about the validity of the instrument and the participants capability of measuring their trust, and if they believed their trust could improve in a future version with suggested changes applied. Without these answers, the authors would only be able to guess when looking for explanations for the participants levels of trust and what improvements to make for the PITS. That was why three additional questions were added after the statements. The questions were:

1) Considering the system you tried today, what improvement do you think should or needs to be done?
2) How do you rate your trust for the system today, from a scale 1 to 10?
3) If the system would be improved with the changes you suggested, how much trust do you think it would be possible to get, from scale 1 to 10?

The percentage of overall perceived trust the participants had to the PITS was calculated by dividing the amount of times participants agreed with the statements from the instrument, to the total amount of statement. The textual data would be analyzed in a thematic analyze to add to the discussion about what might have affected the participants overall perceived trust.
for the PITS, what improvement they wished to see, and how their trust could improve if changes were implemented. With the thematic analysis repetitive patterns and gainable insights for improving the PITS was expected to be revealed from the information the participants gave.

**Selection of participants**

Participants who had cross country skied before and also had experience with roller skiing and performing double poling, were encouraged to participate in the study. A message request was sent out to cross country skiers through an email list given by staff at Umeå School of Sport Science (USSS), and in two cross country skiing groups on Facebook with a total of 1500 members (see Appendix A). 6 biological males in the age from 24 to 50 years decided to participate in the study by signing up on a document. Most of them had skied over 16 years, and the minimum experience was 7 years. Two of the participants had never done roller skiing on a motor driven treadmill before. The rest had roller skied on a motor driven treadmill 1-6 times. The number of participants were limited because of low interest to participate in the study. Because of participants had signed up with only first names in the sign up-document, personal messages were sent out to skiers with two certain names in the two cross country skiing groups which led to two extra participants signing up.

**Instruments and material**

**The PITS.** The PITS was developed by Guerrero et al. (2017) to improve cross country skiers double poling technique. It could either be used by cross country skiers receiving instant feedback in real time, or in retrospective by a human coach analyzing the skiers performance along with the skiers. This study focused on the real time respond/feedback of the PITS, and to this a new front-end of the platform was developed. Some of the features of the new user interface was suggested by the authors of this report. In the new version, feedback about temperature in the face, name, age, average speed, elbow angle and Swedish and English statements about the performance depending on the performance level; was replaced with a percentage of polling and recovery performance changing from red, yellow to green depending on the performance percentage, difficulty level was added with a mark on chosen level, and a Swedish statement of the performance saying “Fantastic!” when the participant reached 100% in poling and/or recovery phase. The authors could manually adjust the difficulty level of the PITS for each participant from level 1 to level 5. The comparison of the cross country skier’s and the reference elite skier’s technique would be visualized on a TV-screen in front of the cross country skier. Users were able to see: themselves, a stick figure representing themselves, and a percentage of how accurate their double poling technique matched the reference elite skier. Both the original and adjusted version the PITS would receive data of the skiers body and limb postures with a Microsoft Kinect camera, and a machine learning algorithm compared the cross country skiers data with the data of a reference elite skier before returning the feedback. If the participant would perform 100% in a performance level, green dots on the stick figure would specify the correct
movements of the limbs and the participant would receive written feedback saying “Fantastiskt!” (Figure 3).

**The instrument.** The instrument was developed in Australia by Madsen and Gregor (2000). It consisted of 5 different constructs relevant to HCT with 5 statements each, for the participants to agree or disagree on. According to Madsen and Gregor (2000) the instrument was “the first of its kind to be specifically designed to investigate the HCT and shown empirically to be valid and reliable.”. The amount of agreements or disagreement was used to estimate the users overall perceived trust for the system in percentage. This was done by dividing the amount of total statements to amount of agreements. If a participant did not want to agree or disagree to the statements in the instrument, that statement would be discarded from the total amount of statements in the calculation.

The instrument was created for users that had months to years experience of a system, but the participants in this study, would use the system for the first time during 10 minutes. The instrument was originally in English, but translated into Swedish for this study, since the participants would be assumed to have Swedish as their first language (Appendix D).

![Figure 3. The feedback the participants would receive when they performed 100% on a double poling movement. The participant is anonymized behind back squares in the screenshot. The green circles would specify the correctness of the movements.](image)

**Pilot study**

A pilot study was conducted to test the setup of the Kinetic camera, the TV-screen, and the PITS. A cross country skier with previous experience from roller skiing on a motor driven treadmill participated in the pilot study. The pilot study showed that the setup worked, but before the actual study could take place; the system needed some final adjustment of the code. A second version of the system was programmed and tested, and adjustments were
being done also for the second version. The authors tried the final version of the system by making double pooling movements on the non moving motor driven treadmill without roller skis on. This gave enough information about how the feedback was delivered in the final version of the system and deemed workable for the study. The instrument was tested by the authors of the study, to see if there were unclarity about the statements after translation, and how concepts could be furthered explained if needed.

**Procedure**

Before the evaluating of the PITS, the participants were informed about the participation through messages on the social network site Facebook and they were also asked to fill in a survey about their background experience of roller skiing on a motor driven treadmill, cross country skiing and receiving feedback from computational devices or human coaches (Appendix B).

The test was performed in a laboratory environment at Sport Sciences Lab at Umeå School of Sport Science (USSS). The participant met up one by one. While the participants started by warming up on a training bicycle for 5 minutes, they were once again informed about the aim of the study and the procedure of the evaluation (see Appendix C). They were informed that their participation was non-compulsory and that they could cancel whenever they wanted without further explanation. When they stood on the motor driven treadmill (Rodby RL3500E 3500 x 2500 mm) at 1 degree inclination with the roller skiing equipment on, they were informed about the meaning of the information on the screen. The participants had a safe rope around their waist as a safety procedure if they would fall, they used skis and poles from the lab and worn their own ski boots. The participants could see their feedback on a SONY 45” inch screen in a distance around 3 meter and in a height of around 2 meters (Figure 4). Under the screen, there was also a floor length mirror. The camera was a Microsoft Kinect 2.0 (field view 70° horizontal × 60° vertical, depth range between 0.8 m and 4.0 m) and stood two meter in front of the participant roller skiing on the motor driven treadmill. The participants were encourage to focus on the balance for the first minutes and ignoring the screen, and after that take more notice of the screen. The double poleing (dp) was performed for 10 minutes. The speed on the motor driven treadmill started on 0 and went slowly up to around 15 Km/h. The participant gave constant feedback whether they wanted to lower or higher the speed during the whole test, to one of the authors who managed the speed on the side of the treadmill. The lowest constant speed for a longer period of time was 15 km/h and the highest speed was 22 km/h. The difficulty level of the PITS was ranging from 4 up to 5 during the process, were 5 was the most challenging level of the PITS for the participant to get 100% in poling and recovery performance. The level was set and changed manually by one of the authors during the study.
The semi-structured interviews were held in the same building but in another room, straight after the participants used the PITS. The choice of the semi-structured close to the experience with the PITS, was so the participants would remember as many details as possible. The participants were offered a glass of water and then informed about the procedure of the semi-structured interview before it began. During the interview, one person held the semi-structured interview and the other one took notes. The interviews were recorded with the acceptance of the participants, and they were informed that they could cancel their participation without further explanations. The participants were asked to agree or disagree to 25 statements in 5 different constructs from the instrument by Madsen and Gregor (2000), as well as furthering their thoughts about the statements and answering three additional questions in the end of the interview. After the test, the participant received a chocolate bar, a banana and a small juice packet as a gift for their participation. They were not informed about the gift beforehand.

**Encoding data**

After transcribing the interviews, the participants were given a chance to read through the transcribed interview and give feedback if they wanted to change or add anything (Appendix E). This was made to improve the reliability of the data (Hedin, 2011). All six participants received the opportunity to read the transcribed version of their semi-structured interview. Out of the six participants three replied and approved the transcribed interview. One responded with no need to read through it, and two did not respond at all. The textual
data were then analyzed through three levels of coding where keywords were drawn out. This was done to organize and reduce the amount of data (Langdridge, Hagger-Johnson, 2009). In level one, nine descriptive codes were developed through discussions and applied to all the transcribed interviews. Some codes were revised during the process, for e.g. if a code could be merged into a more general code, or there was a need to add or develop a code to avoid confirmation bias of what the authors looked for. In level two, sixteen interpretative codes were developed through discussions which encouraged interpreting the participants statements into codes. Deliberation was taken to look for negative cases and alternative explanations that contradicted already established codes.

In the third pattern coding level, whole segments of statements were coded through relevant theoretical ideas that would be grounded in psychological theories. Some of the codes that was created was for example the framing effect of an elite skier as the reference skier (Levin, Schneider & Gaeth, 1998), and the cognitive overload from distributing the participants attention between instructions and performance (Starfish Therapies, 2012). These melted into the thematic session where high level concepts were drawn out from the data, but in the thematic session a tenth theme was added because of the need for a category that collected the participants positive thoughts about the PITS (Appendix F).

Ethical considerations

Consideration were taken to be careful about the data the participants would conduct to the semi-structured interviews and surveys. An individual code number was randomly selected for each participant after the interviews to certain their anonymity.

Results

The data from the semi-structured interviews were coded in three levels by keywords assigned to the sentences of the transcribed interviews so they could be analyzed and categorized in a thematic analyze. This was expected to give insight into why users had trust or distrust for the PITS and if there where repetitive patterns. One of the additional questions asked for the participants own perception of trust for the PITS rated from 1 to 10. This was later compared to the percentage of overall perceived trust from the instrument, by curiosity from the authors to see if the participants own ratings of perceived percentage of trust was similar to the instruments calculations (Figure 5). Participants self rated trust for the PITS of today, in a future version with applied changes, along with the overall perceived trust according to the instrument of Madsen and Gregor (2000) was compared and can be seen in Figure 5.
Figure 5. Overview of participants overall perceived trust in blue, according to the instrument by Madsen and Gregor (2000). The participants self rated trust in red, and participants imagined self rated trust if their suggested improvements were done to the PITS in orange. Participant D01 did not contribute to all questions.

The thematic analysis gave an overarching view which represented the participants thoughts about the PITS and suggestions of improvements. Figure 6 presents the different themes and the amount of times they were present in each participants’ textual data. The themes included:

1. Cognitive overload
   Participants faced cognitive overload when distributing attention between the balance on the treadmill, technique of the performance and receiving feedback on the screen.

2. System improvements
   The participants saw the need for improvements in front- and backend in the system.

3. Improvement for receiving feedback
   The participants saw the need for improvements in placement of technical devices.

4. Perceived Intelligence
   The participants did not perceive the systems ways of handling input and output as as a more intelligent way than what a human coach could do.

5. Framing effect
The choice of an elite skier as the reference person had a positive impact on the participants trust for the system, and no mention or question was made about the skills of the developers who coded the system.

6. Contextual Constraints of Usability
   The participants did not believe the system could help them become better skiers in the real world.

7. Expert vs. Novice
   The participant thought that experience of using the system and/or skiing was a factor that would influence the participants’ experience with the system.

8. Distributed vulnerability
   The participants were positive to use the system among other tools to improve their cross country skiing, but not on its own.

9. Confirmation bias
   The participants would compare the systems feedback to their own thought about their performance to validate the correctness of the system.

10. System Positivity
    The participants expressed positive thoughts about usage of the PITS.

Figure 6. Overview of the themes in the thematic analysis from the categorizing of participants interviews. The X-axis represent the amount of times each theme was mentioned by the participants, the participants are separated by colours.
In the theme Distributed vulnerability, participants were positive to use the PITS among other tools to improve their cross country skiing, but not on its own. In the theme Framing, the choice of an elite skier as the reference person had a positive impact on the participants overall perceived trust for the system, and no mention or question was made about the skills of the developers who coded the system. The two participants with the highest overall perceived trust for the PITS, mentioned that the choice of an elite skier as the reference was of great importance for their trust to the system.

The participants talked about their self confidence for improving their cross country skiing technique by their own, and they all reported a high confidence. The theme Confirmation bias revealed that the participants compared the systems feedback to their own thought about their performance, to validate the correctness of the system. For some participants, this meant that they distrusted the system because they got the same feedback on their double poling movements no matter if they performed what they perceive as good or bad technique. One participant viewed the feedback as accurate and mentioned that if the result would have shown that the performance was in 10% of the reference of the elite skier, then the participants trust would have been reduced.

The three additional questions in the semi-structured interviews revealed that the participants had many suggestions for improvements, both for the placement of technical devices and for improvements in the front- and backend of the system to lower the cognitive load and give more valuable feedback.

Examples:

I would like to have more of e.g. “move poles further forward” or “longer movement for the poles”, “less crouching with the knees” […] analyse the whole body's movement pattern so I get more technical correct movement. (D01T2)

I would like the referent skier to ride beside me on the screen as a stick figure. In this way I could see how he skies compared to me and what limbs I need to adjust. More concrete things to change and not just red and green. You know you are searching for something but don’t really know how to achieve it. (D06P5)

The second question about their own ratings of perceived trust to the PITS was similar to the percentage from calculating the result of the instrument of Madsen and Gregor (2000). In the third question, all the participant thought their trust could improve to 70-100% if their suggested improvements would be implemented to the PITS. The themes for what it would take to get this level of trust could be found in the theme System improvements; where the participants saw the need for improvements in front- and backend in the system, and in Improvement for receiving feedback; where the participants saw the need for improvements in placement of technical devices.

Examples:
The first detail I thought of during the test was that when the screen was in that place, I couldn’t ski 100% as I wanted to. (D03B)

I thought it would be interesting if there were more point of measurements. And that it would be better if the square where the values shows up wasn’t transparent, but maybe had a square behind it, so it would possible to see what the values are. (D05B)

It was like looking at a great movie on a cruise boat with 70 people standing in front of me. And you can see that it’s some kind of information on the screen. But what, why and how I should use it, does not exist. (D06P1)

**Discussion**

**The selection**

Although both men and women were encouraged to participate in the study, only biological men did. The importance of both sexes participating in the study was due to that the physiology might differ between biological men and women, and maybe between the reference elite skier and the participants. The biological sex of the reference elite skier is not known to the authors. There was a homogenous group where all participants had great experience of cross country skiing and a big self-confidence to judge their own performance. Participaters mentioned that it might be different for someone with less experience and on a lower performance level. It would have been interesting to find participant with less experience, and see if they would report the same amount of overall perceived trust in the coaching system, and if they would have gotten as positive feedback from the PITS. For this study, people with experience of roller skiing on a motor driven treadmill were encouraged to participate. The reason experience of roller skiing on a motor driven treadmill was encouraged was to lessen the cognitive load of the participant by performing a new physical task (Starfish Therapies, 2012). It is likely that cross country skiers with little experience did not feel encourage to participate in the study since it was encouraged that people that had tried roller skiing on a treadmill signed up. The experience of roller skiing on a motor driven treadmill is likely to be constrained to athletes performing sports on higher levels.

**The instrument**

The instrument could give a percentage to estimate the users overall perceived trust for the system, but as studied by Drnec, et al. (2016) different expectations will have different importance during different stages. Therefore, a percentage of trust from the participant might have greater impact depending on for how long the user have used the system. One of the instrument’s limitations was that the users should have experience with the system from a few months to several years. The participants in this study, were considered experts when it came to cross country skiing, but novices when it came to using the PITS. They would answer the instrument after using the PITS once for 10 minutes. That was not enough of time to get the best possible outcome of using the instrument. The original thought was to contact
the skiers that had experience of the system for the development of it, but it was not possible to establish that contact.

Example:

I can rely on the system to function properly. (Interviewer)
I don’t know. Since I don’t know how it should work since I haven’t seen it before.
(D04R4)

The instrument was developed in Australia, and there has been findings suggesting that since the nature of trust differ between cultures, a validation is encouraged when transferring findings of trust in automation from one cultural to another (Lee & See, 2004). No such validation was done for this test. In this test, the instrument was used by Swedish citizens in Sweden. The instrument was therefore translated into Swedish from English to lower the language barrier for the user. In the translation, there is a risk that some meanings were changed. A big difference to how the instrument intentionally was used was that the participants did not get to read the statements and fill in their answers on a paper. Instead, the statements were delivered and answered orally. Concepts that would occur in the interview were explained in the beginning, and would later rely on the participant to remember, or to ask to be reminded, for the meaning of the concepts.

Before the participant used the PITS, they were informed that researcher from the Computer Science department had created the PITS. They were not however informed that the authors of this paper definitely were not a part of the Computer Science department or had any part in developing the system. The participants could therefore have gotten the impression that the authors who interviewed them after the test was not neutral to the system and could get offended depending on how the participants expressed their thoughts. Because of this, participant might have confirmed to social desirability bias when they answered questions about the system in front of the authors, restricting the participants from expressing thoughts of negative nature about the PITS (Langdridge & Hagger-Johnson, 2009). Interpersonal factors such as the age, gender and the authority role of leading the interview were interpersonal variables that was not possible to minimize; but consideration was taken in the translation to make sure the words did not get to complicated or presupposed an academic background, to make the participants feel comfortable, and explaining that they would be anonymous throughout their participation.

The setup

The setup was mentioned by all of the participants, due the placement, distance and size of the TV-screen. The TV-screen was placed on a high level on the wall, and not in the height where the participants normally had their gaze when performing double poling. Because of the distance and size of the TV-screen, some of the participants did not see everything on it. One participants also mentioned to barely see the feedback about the recovery phase due to that his head was tilted down for those movements. Some of the participants also mentioned that they rather would have had sound as a feedback.
Examples:

Should be just a little bit bigger then, I think I have pretty good eyesight and I think I had a hard time anyways. It is hard to get the information when you are training, when I double pole. (D02B)

The first detail I thought of during the test was that when the screen was in that place, I couldn’t ski 100% as I wanted to. [...] The contrasts of the color on the screen was not ok, there were things that got away, hard to distinguish, it has to be presented more clear. (D03B)

Some kind of intensified beeping system. Such as when you are correct, the whole frequens will be louder. (D01R1)

In retrospective, it would have been interesting to see if and how participants overall perceived trust would have been affected by a more considered setup. Seeing more of the PITS would be no guarantee for neither, depending on what the participant based their trust on the most. In this study, suggestions that did not get applied for the new front-end version was to have written feedback on the TV-screen that would differ from e.g. “You can do better”, “Keep going” to “Fantastic!” depending on how similar the participants performance were to the reference elite. For the final version the only feedback the participant could receive was “Fantastic” if the participant performed close to the reference elite skier. Some of the participants did not find this feedback useful, because it was not constructive or informative enough, since it did not specify what was better in one good double poling movement than another. It would have been interesting to see how and if the intended feedbacks would have affected the participants trust and evaluation of the PITS.

Examples:

I didn’t get any feedback I could use in a constructive way. I only got fantastic all the time regardless of how I doubled poled. (D04R2)

Fantastic? hahaha [...] No, there are competent people that can give me better tips than fantastic. (D04T3)

The result

When it came to the thematic analysis, the reliability of the distinction between the categories could have been higher if the authors let another person place the participants quotes in the 10 different categories (Hedin, 2011). When comparing, the result from the instrument with the participants rated trust, the participants grading was translated from the scale 1 to 10, into a percentage. It is not certain that the participants would have graded their trust the same if they had graded it in percentage from the start. Also one participant graded it to 0 although the instructions was to grade it from scale 1.
The participants had great experience of cross country skiing with four of the six participants cross country skiing over 1000 kilometers of cross country skiing each year (Figure 7). The participants had performed cross country skiing for seven years and onwards, and three of the six participants had been cross country skiing for over 16 years. In the result from the instrument (Madsen & Gregor, 2000) two of the participants had HCT at 72% and 84%, compared to the other participants that had HCT from 0% to 48%. From the semi-structured interviews, data also shown that the participants that had the highest percentage of HCT viewed the PITS as a tool among other already existing tools to improve their technique, and not as the only tool. This is important to acknowledge, because the definition of trust in this paper is that an user are confident in that an agent, which can be a computer or a human, will help them to achieve goals also in environments and situations where they feel vulnerable from ambiguity and uncertainty (Lee & See, 2004). According to Mayer et al. (1995), vulnerability is a key element for the development of trust, and trust is rather about being willing to take a risk, than to actually take a risk. If a person does not position themself in a situation where they are willing to take a risk, there is no need for trust. When a participant evaluate the PITS in the context of “a new tool to use with my other tools” the vulnerability the participant faces gets distributed among the other tools.

Ungefär hur många mil längdskidor åker du per år? Skate och/eller klassisk.

6 svar

Figure 7. Answers to one of the questions in the questionnaire on how many Swedish mils (1 mil = 10km) the participant ski each year.

As well as the lack of intrinsic values and motivation, studies have shown differences in the effect of perceived trustworthiness depending if the beholder is thought to be a human or a machine. When a person wants to restore trust in a human-to-human relationship, acknowledging one’s own perceived reduced trustworthiness is of importance. A logical way to manifest this insight could be by delegating a task to another person. The PITS was not able to let the user know if her/his performance level was lower or higher regarding to what
the existing performance levels of the PITS could evaluate. There is a possibility that the participants would have had a greater overall perceived trust if the PITS acknowledged that their performance would be to advance to be able to give them anything else than 100% and “Fantastic!” as a feedback for each double poling movement.

Examples:

It is not working on the level I ski. (D04P1)

I am not sure that it will solve a problem, not as I experience it. I can imagine that it is a little different if you are less experienced skier maybe. That it could be more value in what you get, that it could guide you on someway then, but not in my case. (D03F5)

Lewandowsky et al. (2000) studied if trustworthiness would lead to different behaviors of a human user if the collaboration partner were thought to be a human or a machine. First, they investigated if the user would judge their own trustworthiness based on the amount of faults they made. Rationally, the more fault they did, they perceived their trustworthiness as reduced in the perspective of their collaboration partner, whether the partner was a human or a machine. They also showed that if the collaboration partner was thought to be a human and perceived trustworthiness were low, the operator were more likely to allocate the task to the collaboration partner. If their trustworthiness were perceived as high, they were more likely to keep the task. Although, if the collaboration partner were thought to be a machine, perceived trustworthiness did not matter for allocating the task.

As mentioned before, if participants were aware of faults before using the system, the faults would not affect the trust for the system (Lewandowsky et al., 2000). The participants in this report’s study, were aware of that the only feedback they would receive were “Fantastic!”, but not that they probably would receive the feedback every time they made a double poling movement. If they had been aware of the system's desensitivity this could have affected the level of trust they had. The user tried out a prototype, but it was never described as a prototype. The framing effect of the PITS as a prototype, could have affected the trust in both ways.

So, do the cross country skiers have trust for the physical intelligence tutoring system for feedback on performance? In general terms, the current PITS does not inflict a lot of overall perceived trust in most of the participants. Moreover, the semistructured interviews showed that the participants that had the higher overall perceived trust in the system thought of the system as another tool in their collection. They also did not put trust in the system, but to the reference elite skier.

On one hand, it is not likely that the participants actually would use the PITS in its current state, because of their low trust towards it. On the other hand, basing the trust on the knowledge of the mismatch between what the PITS can do and what they need to improve their double poling technique, is a sign of effective development of calibrated trust for the PITS, which is important for not misusing PITS. In future improvements of the PITS, users
capability to develop calibrated trust should be prioritized and retained, but improvements need to be done to the PITS so the users overall appropriate trust to the PITS is high enough to motivate an usage. Based on our experimental results with cross country skiers, we hypothesize that if a human (e.g. a skiing coach) can do the same job, the users will not choose an autonomous system before the human. In the sense of using the PITS as one of other tools to improve performance where less trust is needed, there is still a chance that it will be used. From this study, the suggestions to improve the overall perceived trust to the system, lays in improving the quality of feedback from the system, and present it on equipment that does not constrain the user from the task of performing.

Future studies

To retain users trust if and when errors occur, faults that can occur in the PITS should be presented for the user before using the system (Dzindolet, Peterson, Pomranky, Pierce, Beck, 2003; Lewandowski et al., 2000). Still, caution need to be done so the user do not develop overtrust to the system, because of this. Explanations about errors should only be applied so the user can develope calibrated trust. Therefor it might be of interest for the user to know what input the system need, what the systems decisions are based upon and what the output means when function accordingly. Future studies need to look into how the explanation of errors increasing trust can be used to reach calibrated trust, and not overtrust.

Because the overall perceived trust of the participants in this study were low towards the PITS, it would be interesting to see if the participants could get an higher overall perceived trust in an improved future version of the PITS, or if the first initial interaction and experience with the PITS would still influence their overall perceived trust with the PITS. This could be tested in a between group design, where one group have had previous interactions with the PITS, and one group interacts with the new and improved version of the PITS for the first time. The different and individual level of trust users have for intelligent systems, and how considerations could be taken to adapt to the users level of trust to generate a calibrated trust, could also be considered more thoroughly. It would also be interesting to see novice cross country skiers trying the PITS over a period of time, to see if the PITS could help participants with less experience and lower self confidence to improve their double poling technique.

Reference list

Andersson, J., Malm, M., & Thurén, J. (2003). Systemtilltro (FOI Rapport FOI. FOI-R--1121--SE) Retrieved from the website of FOI:
https://www.foi.se/rapportsammanfattning?reportNo=FOI-R--1121--SE


Hedin, A. (2011). En liten lathund om kvalitativ metod med tonvikt på intervju. Retrieved 2018 May 09 from https://studentportalen uu.se/uusp-filearea-tool/download.action%3FnodeId%3D459535%26toolAttachmentId%3D108197%26cd=1%26hl=sv%26ct=clnk%26gl=se


22
Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A
typology and critical analysis of framing effects. Organizational Behavior and Human
Decision Processes, 76(2), 149-188. doi: 10.1006/obhd.1998.2804
Brisbane, Australia: Information Systems Management Research Centre
doi:10.5465/amr.1995.9508080335
Skiing, a data-driven automatic assessment of double poling technique. Retrieved 2018
April 12 from
ncepts/item/230-kinesis-kinetic-smart-evaluation-for-cross-country-skiing-a-data-driven-a
umatic-assessment-of-double-poling-technique
Improve Acquisition of Motor Skills. Retrieved 2018 May 14 from
ng-and-strategies-to-improve-acquisition-of-motor-skills/
The Guardian. (2018) Tesla car that crashed and killed driver was running on Autopilot, firm
says. Retrieved 2018 April 20 from
https://www.theguardian.com/technology/2018/mar/31/tesla-car-crash-autopilot-mountain-
view
from https://www.tesla.com/sv_SE/autopilot

Appendix A

The first message request to get participants to the study

Hej!

Visst är det några av er här som tidigare har testat att åka rullskidor på rullband?
Vi är nämligen två studenter från Kognitionsvetenskapliga programmet som påbörjat vårt examensarbete där vi utvärderar ett coachande längdskidsåkningssystem som forskare på datavetenskapliga institutionen tagit fram. Och nu behöver vi hjälp med att testa detta!

Systemet analyserar stakrörelser och ger feedback på hur de överensstämmer mot lagrade rörelser utförda av referensgrupp. Vi är nyfikna på hur själva interaktionen mellan skidåkare och systemet ser ut, och undersöker detta genom att ställa frågor om upplevelsen efter att du testar systemet. Det finns alltså inga rätt eller fel, utan det är din upplevelse av systemet som är av intresse för oss.

Deltagande

1. Du kommer åka rullskidor och staka på ett rullband i 10 minuter.
3. Efter stakningen kommer du att få frågor om upplevelsen.

Förhandskunskaper: du har åkt rullskidor på rullband vid något tidigare tillfälle och har vana av längdskidåkning där du också har använt dig av stakning

Vid deltagande kommer du att behöva ta med egna pjäxor. Rullskidor och stavar finns i Sportlabbet.

Tidsåtgång: ca 1 h
Plats: Sport Sciences Lab, Umeå

När:
Måndag
2018-04-23
07:50-09:00
Måndag
2018-04-23
09:50-11:00
Måndag
2018-04-23
11:50-13:00
Måndag
2018-04-23
13:00-14:10

Så deltar du:
Följ länken här: [https://docs.google.com/...](https://docs.google.com/.../1AV8jwTBLS6JgpZacrxxwaAM900.../edit...), för att skriva upp när du har möjlighet att delta.
Du kan också skriva upp när du har möjlighet att vara reserv. Fler tider kan komma senare under veckan, vilket vi då påminner vi här i inlägget.

Stort tack för att ni tagit er tid att läsa! Och för att ni om möjligt skriver upp er och/eller tipsar en/några skidåkande vänner om detta!

/Anna Lindström och Karolina Thorsén, Kognitionsvetenskapliga programmet på Umeå universitet

https://www.facebook.com/groups/334665733308132/

Appendix B

Survey about participants background experience relevant for the study.

https://docs.google.com/forms/d/e/1FAIpQLSf_yIbFUQO3ljgxHp-uKBs43oY5vSgaFWgoYj0xoVQ4fBSTIQ/viewform?usp=sf_link

Appendix C

Information to participants when participating in the study.

Information på plats:
Utvärdering av ett coachande längdskidsåkningssystem som forskare på datavetenskapliga institutionen tagit fram.

Systemet analyserar stakrörelser och ger feedback på hur de överensstämmer mot lagrade rörelser utförda av referensgrupp. Vi är nyfikna på hur själva interaktionen mellan skidåkare och systemet ser ut, och undersöker detta genom att ställa frågor om upplevelsen efter att du testar systemet. Detta är frivilligt och du har möjlighet att avbryta testet när du vill utan förklaring.

Efter 5 minuters uppvärmning på en cykel ska du få köra stakning på rullbandet i 10 min och få feedback av det coachande längdskidsåkningssystemet. Den kommer jämföra dina rörelser med en elitskidåkares bästa prestation. Detta kommer gälla hela kroppen och vinklar på olika lemmar. Du kommer få se en skärm i realtid av dig själv och en digital streckgubbe som representerar dina lemmar.

När du börjar staka dig fram kommer du få skriftlig feedback på skärmen från systemet hur bra din teknik är.

När de står på bandet och kan se skärmen:
Bredvid din streckgubbe kommer du också få information om

- Totala prestation, du får skriftlig feedback på din stakteknik.
Appendix D

The instrument in Swedish and English

<table>
<thead>
<tr>
<th>Swedish</th>
<th>English</th>
</tr>
</thead>
</table>

1. Upplevd pålitlighet
   Pålitligheten i systemet, vid vanligen upprepat och konsekvent användande.

<table>
<thead>
<tr>
<th>Swedish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1 - Systemet gav mig alltid de råd jag behövde för att ta ett beslut.</td>
<td>1 - The system always provides the advice I require to make my decision.</td>
</tr>
<tr>
<td>R2 - Systemet fungerade på ett tillförlitligt sätt.</td>
<td>The system performs reliably.</td>
</tr>
<tr>
<td>R3 - Systemet reagerade på samma sätt, vid samma förhållanden, vid olika tidpunkter.</td>
<td>The system responds the same way under the same conditions at different times.</td>
</tr>
<tr>
<td>R4 - Jag kunde lita på att systemet fungerade som det skulle.</td>
<td>R4 - I can rely on the system to function properly.</td>
</tr>
<tr>
<td>R5 - Systemet analyserade problem konsekvent.</td>
<td>R5 - The system analyzes problems consistently.</td>
</tr>
</tbody>
</table>

2. Upplevd teknisk kompetens
   Systemet uppfattas att utföra uppgifterna noggrant och korrekt baserat på den input systemet tar in.

<table>
<thead>
<tr>
<th>Swedish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 - Systemet använder lämpliga metoder för att ta beslut.</td>
<td>T1 - The system uses appropriate methods to reach decisions.</td>
</tr>
<tr>
<td>T2 - Systemet har god kunskap om denna typ av problem inbyggt i det.</td>
<td>The system has sound knowledge about this type of problem built into it.</td>
</tr>
<tr>
<td>T3 - De råd som systemet producerar är lika bra som vad en högt kompetent person skulle kunna ge.</td>
<td>The advice the system produces is as good as that which a highly competent person could produce.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>T4 - Systemet använder informationen jag förser det med på ett korrekt sätt.</td>
<td>The system correctly uses the information I enter.</td>
</tr>
<tr>
<td>T5 - Systemet använder sig av all den kunskap och information det har tillgång till, för att ta fram en lösning på problemet.</td>
<td>The system makes use of all the knowledge and information available to it to produce its solution to the problem.</td>
</tr>
</tbody>
</table>

### 3. Upplevd förståelse

**Du kan skapa en mental modell och förutse systemets framtida beteenden.**

<table>
<thead>
<tr>
<th>U1 - Jag vet vad som kommer hända nästa gång jag använder systemet eftersom jag förstår hur systemet beter sig.</th>
<th>I know what will happen the next time I use the system because I understand how it behaves.</th>
</tr>
</thead>
<tbody>
<tr>
<td>U2 Jag förstår hur systemet kommer hjälpa mig med beslut jag behöver fatta.</td>
<td>U2 - I understand how the system will assist me with decisions I have to make.</td>
</tr>
<tr>
<td>U3- Även om jag kanske inte vet exakt hur systemet fungerar, vet jag hur jag använder det för att fatta beslut om problemet.</td>
<td>U3 - Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.</td>
</tr>
<tr>
<td>U4 -Det är enkelt att följa med i vad systemet gör.</td>
<td>U4- It is easy to follow what the system does.</td>
</tr>
<tr>
<td>U5 - Jag förstår vad jag ska göra för att få de råd som jag behöver från systemet nästa gång jag använder det.</td>
<td>U5-I recognize what I should do to get the advice I need from the system the next time I use it.</td>
</tr>
</tbody>
</table>

### 4. Förtroende

**Som användaren menas detta att du har förtroende för systemets framtida förmågor att kunna presterå/göra bra ifrån sig också i oprövade/nya situationer.**

<table>
<thead>
<tr>
<th>F1 - Jag litar på råden från systemet, också när jag inte vet helt säkert att det är korrekt.</th>
<th>F1-I believe advice from the system even when I don’t know for certain that it is correct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2- När jag blir osäker inför att fatta ett beslut litar jag hellre på systemet än mig själv.</td>
<td>F2 - When I am uncertain about a decision I believe the system rather than myself.</td>
</tr>
<tr>
<td>F3-Om jag inte är säker på ett beslut, har jag förtroende för att systemet kommer att ge den bästa lösningen.</td>
<td>F3 - If I am not sure about a decision, I have faith that the system will provide the best solution.</td>
</tr>
<tr>
<td>F4-När systemet ger ovanliga råd är jag säker på att rådet är korrekt.</td>
<td>F4 - When the system gives unusual advice I am confident that the advice is correct.</td>
</tr>
</tbody>
</table>
F5- Även om jag inte har någon anledning att förvänta mig att systemet kommer att kunna lösa ett svårt problem, känner jag mig fortfarande säker på att det kommer göra det.

F5 - Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.

5. Personlig tillgivenhet som användare tycker du om systemet på så sätt att det är trevligt att använda och det matchar dina preferenser. Du föredrar och är förtjust i att använda det, och är fäst vid systemet.

Personal Attachment to the system comprised of: liking meaning that the user finds using the system agreeable and it suits their taste and loving meaning that the user has a strong preference for the system, is partial to using it and has an attachment to it.

<table>
<thead>
<tr>
<th>P1 - Jag skulle känna en känsla av förlust om systemet inte var tillgängligt och jag inte längre kunde använda det.</th>
<th>I would feel a sense of loss if the system was unavailable and I could no longer use it.</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2 - Jag känner mig hängiven till att använda systemet.</td>
<td>P2 - I feel a sense of attachment to using the system.</td>
</tr>
<tr>
<td>P3 - Jag känner att systemet passar mitt sätt att fatta beslut.</td>
<td>I find the system suitable to my style of decision making.</td>
</tr>
<tr>
<td>P4 - Jag gillar att använda systemet för att fatta beslut.</td>
<td>P4 - I like using the system for decision making.</td>
</tr>
<tr>
<td>P5 - Jag föredrar personligen att fatta beslut med systemet.</td>
<td>P5 - I have a personal preference for making decisions with the system.</td>
</tr>
</tbody>
</table>

Appendix E

Transcribed interviews

https://drive.google.com/file/d/1YFoP09Wuc9FEKl9On5Tc07EHA7mhYVvL/view?usp=sharing

Appendix F

Thematic analysis of the transcribed interviews

https://drive.google.com/file/d/1-Y-wxwzjbwjWg0GCOhC5nySuqghj2pHV/view?usp=sharing