THE DIAGNOSIS OF SELF-EFFICACY USING MOUSE AND KEYBOARD INPUT

MSc Thesis

Maarten Dijkstra

ICA-3755770

Juli 2013

Supervisors:

Dr. P.J.M. Wouters

Dr. H. van Oostendorp

Universiteit Utrecht

Master Game and Media Technology

Dept. of Information and Computing Sciences

Utrecht University

Abstract

Self-efficacy is a student's belief in his or her own capabilities regarding the completion of a specific task. Students with high levels of self-efficacy are proven to be more effective learners. If serious games, intelligent tutoring systems and computer enhanced learning in general can diagnose self-efficacy, it could lead to improved tutoring strategies, consequently improving the learning experience and process of the student. This research investigated the diagnosis of self-efficacy levels at runtime using mouse and keyboard input, the default communication channels of computer enhanced learning. In an empirical experiment, small to medium significant correlations were found between mouse movement and self-efficacy levels for the variables *Distance difference*, *Number of pauses*, *Time difference*, *Pause time* and *Question time*. Linear multiple regression revealed that mouse movement variables were able to predict 17% of the levels of self-efficacy, in which the *Time difference* was the largest and only significant contributor. This means that *Time difference* can be used to partially diagnose self-efficacy levels. In contrast, no correlations were found between typing performance and self-efficacy levels.

Acknowledgements

This thesis could not have been realised without the help of the people who surrounded and supported me during my Master programme and my research. First I would like to express my gratitude towards my supervisors Dr. Pieter Wouters and Dr. Herre van Oostendorp, who gave me the freedom to explore my interests while their expertise helped me to avoid the caveats of doing empirical research. I would also like to thank my girlfriend, who supported me, believed in me and listened to countless hours of babbling and more than a few utterances of frustration. Furthermore, I would like to thank my parents for supporting me through my studies and lastly I would like to thank the participants of my experiment, who sacrificed their time to help me with this research.

1. Introduction

Serious games are popular in the educational field today and research into these games that educate and entertain at the same time almost unanimously agrees that serious games have the potential to greatly improve the learning experiences of students. Games (whether serious or not) inherently teach their players to understand the game and the rules that define the game world. Good games even incorporate learning principles that are strongly supported by contemporary research in cognitive science (Gee, 2003; Chen, 2005). But assessment within serious games is still difficult and frequently only assesses if the player completes the game. This can be insufficient since a student could have only learned the rules of the game without understanding the educational content (Chen, 2005).

Where serious games focus on the completion of the game or level, most e-learning is only concerned with the correct solution to a problem and gives the student feedback on their performance in relation to this correct solution. Although it might seem that traditional education assesses in the same way, it uses human tutors to assess its students. Where e-learning only assesses the answer itself, human tutors use errors, student responses, and features of interaction (e.g., the timing of a student's responses and/or the way in which a response is delivered) as diagnostic evidence to assess and tutor a student (Derry & Potts, 1998).

To address this issue, much research has been done to assist or tutor the student in e-learning, resulting in the creation of Intelligent Tutoring Systems (ITS) and a variety of tutoring strategies. These studies also developed various methods to acquire information about the student to determine the most appropriate tutoring strategy, ranging from simple heuristics to advanced statistical predictions. However, human tutors indicate that they do not use such complex evaluation systems but simply classify the students based on their performance and motivation. So instead of using probabilities or statistical calculations like some e-learning systems, they combine knowledge of the student's motivational state with knowledge of the domain to assess the student's overall competence (Derry & Potts, 1998).

Human tutors indicate that a large component of a student's motivation is their confidence (de Vicente & Pain, 2002). Research combining confidence and learning indicates that a student's self-efficacy (a student's belief in their own capabilities regarding the completion of a specific task) is key to his or her engagement and learning efficiency (Linnenbrink & Pintrich, 2003). For humans it is intuitively clear which behaviour shows more or less confidence and this is taken into account when asked to assess a student (Derry & Potts, 1998). However, e-learning systems or serious games cannot assess the student in the same way human tutors do, since they normally only receive their

information through mouse and keyboard input and therefore can be considered blind and deaf compared to human tutors. Despite the proven importance of self-efficacy and the statements of human tutors who indicate that confidence is a large component of their motivational assessment, little to no research has been done to acquire the level of self-efficacy of a student. Most studies (del Soldato, 1993; Derry & Potts, 1998; de Vicente & Pain, 1998; de Vicente & Pain, 2002; Beal & Lee, 2005) research motivation as a whole, without establishing how each component of motivation can be measured separately. Of the studies that do (McQuiggan & Lester,2006 and McQuiggan et al. , 2008), no research has been done to acquire self-efficacy from mouse and keyboard input, the default input channels of computer enhanced learning. The diagnosis of a student's self-efficacy level could inform about a student's overall motivation in more detail and with more certainty and consequently allow for more appropriate adaptations to improve the student's self-efficacy levels. In turn these improved self-efficacy levels could improve a student's learning process, since self-efficacious student learn more efficiently (Linnenbrink & Pintrich, 2003).

This research aims to diagnose levels of self-efficacy using keyboard and mouse input at runtime of students by answering the following research question:

To what extent can levels of self-efficacy be diagnosed using mouse and keyboard input in order to improve a student's learning experience?

It is expected that the keyboard and mouse input of students with a high level of self-efficacy will be different from their input when they have a low level of self-efficacy. An experiment will be performed to assess whether and to what extent this applies for students who are answering quiz-like multiple-choice and open-ended questions. During the experiment the students will answer quiz-like multiple-choice and open-ended questions while their mouse movement and typing performance is recorded and they indicate their self-efficacy level for each question.

The rest of this thesis is structured as follows. The relevant prior research concerning this study is presented in the related work section after which the design of the experiment (including the measurements and data analysis) is discussed in the method section. The outcomes of this experiment on self-efficacy diagnosis are described in the subsequent results section and their implications and additions to the research field are discussed in the conclusions and discussion. Lastly, the future work section justifies its name and presents suggestions for further research.

2. Related work

The potential of e-learning and serious games has inspired many researchers to create more personal and effective educational systems. Although the research field of computer enhanced learning is still relatively young (with the first computer assisted instruction dating back to 1960 (Sirohi, 2007)) some major branches have separated and developed into new research fields. These research fields will be discussed in this section, starting with serious games, continuing to e-learning systems and motivational diagnosis and ending with mouse and keyboard input.

2.1.Serious games

Serious games studies almost all agree that serious games have the potential to improve the learning experiences of present day students. Games teach their players how to play the game and what the rules within the game world are. Some games even incorporate learning principles that are supported by contemporary research in cognitive science (Gee, 2003; Chen, 2005). For instance, these games give information "on demand" and "just-in-time", in the contexts of the situation at hand and directly linked to a user's purposes and goals at that time. Good games also remain challenging for a player, operating at the outer and growing edge of a player's competence which is known as being in a state of "flow", a state where players are fully invested in the game, forget their surroundings and even the concept of time and don't want to stop playing (Csíkszentmihályi, 2000).

Games also gradually increase the difficulty of the problems that need solving. They confront players in the initial game levels with problems that are specifically designed to teach players what solutions or strategies will work well when they face more complex problems at higher levels (Gee, 2003). This is further emphasized by a reward system or scoring, that not only indicates to a player which actions are good or bad, but also which actions are relevant or not. This allows a player to quickly understand which actions or elements are relevant to learn, remember and apply.

These benefits are difficult to achieve in the traditional education of schools. A human tutor who is responsible for a class of students cannot guide or tutor them all personally because it would take too much time. Therefore, classrooms usually operate at the lowest common denominator, so every student can keep up (diSessa, 2000 in Gee, 2003). Furthermore, content matter is generally explained using general situations, which might make it more difficult for a student to apply this content to his or her own situation, whereas games can explain the content matter in specific (practical) situations.

Despite these inherent benefits, compared to normal games, serious games have two extra hurdles to cross in order to be successful: they need to present (possibly dull or boring) educational content

in an entertaining fashion and also need to assess if and to what extent the player has mastered the content. This is still a very difficult task since it is not sufficient to assume that players have mastered a subject when they completed the game like traditional games do. The student could for instance have learned how to play the game itself, regardless of the educational content. To improve the assessment within serious games and other forms of computer enhanced learning, more information is needed about the player and his or her decisions in the game. However, as McQuiggan et al. (2008) indicates, the task of gathering information about the player is more difficult when learning environments are more complex and serious games can be of a complex nature and involve actions that are not relevant for the assessment of the educational content. For serious games to be considered as serious options for the education of next generations, assessment within these games needs to be improved (Chen, 2005).

2.2.E-learning & ITS

Compared to serious games, which is a relative new research field, e-learning is more matured and has extensively researched the problems that serious games face.

E-learning is commonly defined as computer enhanced learning and has almost been around since there were computers. The potential of computers to enhance and assist humans while learning was noticed and researched since the 1960's and has resulted in many applications of computer enhanced learning (Sirohi, 2007). A more detailed definition of e-learning is given by Garrison & Anderson (2003:p.52): "Broadly defined, e-learning is networked on-line learning that takes place in a formal context and uses a range of multimedia technologies".

Researchers agree that e-learning (like serious games) has advantages over traditional education by default. It can be used whenever and wherever the student likes to use it and doesn't have to depend on teachers and their time schedules. Furthermore, e-learning is (mostly) addressed to an individual student and can be repeated or used until the student comprehends and masters the content. So it also doesn't depend on the schedule or pace of a classroom or fellow students. This individual usage also allows for immediate feedback and provides a clear distinction between right and wrong which can aid the learning process of the student. (Zhang et al., 2004).

However, e-learning needs to do more than digitise the original content and present it using a computer since this can lead to frustration, confusion, and reduced learner interest. For example, some e-learning systems only present content using text like a digital textbook, which may lead to boredom, disengagement and can prevent a good understanding of a topic (Zhang et al., 2004). Newer and more advanced e-learning systems use multimedia to enrich the content and present this

content in various ways to allow a student to interact with the content to better understand it. Video, 3D models, simulations, time-lapses and many other interactive representations can be used to provide more clarity or at least offer the student different ways to look at the content. This is an improvement but it requires an efficient structuring of this multimedia content. For example, locating a particular segment within a long instructional video may be ineffective and time-consuming and might negate the benefits of video as an educational medium (Zhang et al., 2004).

So the use of multimedia content in e-learning systems can enrich the learning experience of a student, but it is still necessary to structure the content so students are not prohibited in their learning process. This inspired researchers to look beyond the simple delivery of digital content and create more intelligent e-learning systems, which resulted in the Intelligent Tutoring System (ITS). These systems tailor the learning structure, format and tutoring strategy to the needs of the student, comparable to the adaptations a real life tutor would make (Derry & Potts, 1998).

The type and complexity of the adaptations an ITS can make, varies per ITS. Some provide step-by-step monitoring of a student's solution, allowing for more accurate assessments, while others only monitor when the student is done with a problem entirely. The latter is consequently only able to provide feedback after the problem, while the first can aid the student with feedback during the problem and can direct the student towards the correct solution. The ITS could change its tutoring strategies and for instance provide more empathetic feedback or provide less feedback when it disrupts a students learning experience. It can vary the type of support (for instance visual or textual support), when this support is given and can choose to provide students with feedback on their progress or not (Conati, 2009). But for an ITS to do this, it needs information to decide if, when and how to adapt the learning experience of a student. This information can be divided into four knowledge types: knowledge about the educational content, knowledge about the relevant tutoring strategies, knowledge about the possible presentations (of the educational content) given the available output channels and knowledge about the student (Conati, 2009).

Of these knowledge types, knowledge about the student is different from the other three types, because it is something that cannot be prepared or defined before the student interacts with the ITS or serious game. The characteristics of and knowledge about a student needs to be collected and stored during the interaction with the student. This collection of knowledge and characteristics is commonly called a "student model" and is considered to be very valuable since it drives many (if not all) decisions that are made by the ITS to tailor the learning experience to the student. An exemplary architecture for an ITS which incorporates a student model is the architecture of O'Shea et al. (1984) from Nawana (1990).

This architecture, which is depicted on the right, shows the different components that contain the four types of information and how they are connected. The Teaching administrator provides interaction information to the student history component, which is the collection of all the actions the student has performed so far and could for instance contain all the answers a student provided. This history is used to update or feed data to the student model, which in turn is used to determine a teaching strategy. If for instance a student answered all questions regarding multiplications wrong, a new teaching strategy could be chosen where multiplication is explained again or more thoroughly. The chosen teaching strategy is used by the teaching generator component to prepare a new question, assignment or other type of content and the teaching administrator component presents this content given the available communication channels.

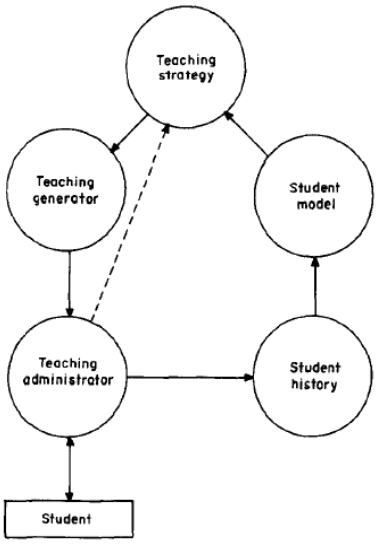


Figure 1: The learning system architecture of O'Shea et al.

Despite its importance, the construction of a robust student model with just the limited communication channels of mouse and keyboard is still very difficult (Nwana, 1990). As mentioned before, an ITS could be considered both blind and deaf when it comes to the clues that students express and which imply knowledge about the student that needs to be included in a student model. Some ITS's solve this lack of information by asking the student to indicate their preference on a number of indicators when an ITS needs this information or after a fixed period of time (de Vicente & Pain, 1998). However, this still requires the student to stop learning for a (brief) moment and do something of a completely different nature, which is likely to disrupt the learning experience of that student. Another method of creating and updating a student model is to infer a student traits based on statistical analysis of his or her behaviour while interacting with the ITS. This behaviour can include all kinds of actions and reactions to events like answering correctly, asking for help, giving up, skipping instructions and/or reading extra background material.

As ITS's increase in complexity, the data that is collected increases equivalently. The large amounts of data and the huge data sets that can and are being collected have fuelled the creation of a new research field called Learning Analytics. This field approaches the problem of creating and updating a student model from a "big data" perspective, where "big data" stands for the huge data sets generated by computers in general these days. Techniques from research into statistics, business intelligence, web analytics, data mining and social network analysis all contribute to interpret the

vast amounts of data that are generated for a single student. This implies a large system that tracks students over longer periods of time to generate these large data sets, but the techniques and current research is promising. However, since the research field is still young, it will take some time for it to catch up with established research fields and contribute to motivational diagnosis research. Therefore, this research focuses on improving the more traditional but mature research field of student modelling using relatively small data sets. In this field, there is still enough progress to be made, since ITS's in current research cannot assess a student in the same way a human tutor does.

2.3.Motivation & Self-efficacy

As mentioned earlier, there are some important differences between a human tutor and an ITS. Despite the limited communication channels, most ITS's store information about a student in a student model. However, this is mostly done by evaluating the cognitive and informative goals set by the ITS. This means that if a student performs well, it is assumed he or she has a good understanding of a topic. Some researchers however argue that cognitive and informative assessment in itself is not enough to ensure an appropriate and stimulating learning experience. *"Students who are anxious, angry, or depressed don't learn; people who are caught in these states do not take in information efficiently or deal with it well.".* This indicates that the motivational state of a student can greatly influence their learning efficiency, which means this is relevant information for an ITS to incorporate in the student model. This is also argued by Derry & Potts (1998), who indicate in their book that human tutors use the features of an interaction like the timing and delivery of student's responses as diagnostic evidence for adapting their tutoring. *"Expert human tutors, it would appear, devote at least as much time and attention to the achievement of affective and motivational goals in tutoring, as they do to the achievement of the sorts of cognitive and informational goals that dominate and characterize traditional computer-based tutors.*" (de Vicente & Pain, 1998).

So an ITS needs the ability to diagnose motivational goals as well as cognitive goals in order to adapt tutoring strategies and improve a student's learning process and experience. However, motivational diagnosis is still relatively new and complex and the motivational state of a student ranges a broad spectrum of components like curiosity, confidence, interest, tiredness, boredom, expectation and many more (de Vicente & Pain, 1998). This research will focus on self-efficacy, a measure of self-confidence which specifically indicates the self-confidence of a person to perform a task (Bandura, 2006).

Bandura (2006) indicates the importance of self-efficacy, since it influences the decisions and motivation of people in general*: "Efficacy beliefs influence whether people think erratically or strategically, optimistically or pessimistically. They also influence the courses of action people choose*

to pursue, the challenges and goals they set for themselves and their commitment to them, how much effort they put forth in given endeavours, the outcomes they expect their efforts to produce, how long they persevere in the face of obstacles, their resilience to adversity, the quality of their emotional life and how much stress and depression they experience in coping with taxing environmental demands, and the life choices they make and the accomplishments they realize."

Although other forms of self-knowledge can also influence the decisions and motivation of students while they are learning, self-efficacy is well-suited for an ITS because it is task-specific. For instance, a student can be self-confident about his or her ability in math but have low levels of self-efficacy regarding linear algebra or even more specific tasks within linear algebra. An ITS cannot decide to adjust its tutoring strategy and for instance skip an assignment based on the general belief of the student that he or she is good in math. It needs information that is specific for the educational content or even a single assignment within that content. Self-efficacy can easily be confused with other form of self-knowledge beliefs like self-concept, self-esteem, perceived control, stability, and self-crystallization. Although the main difference between these beliefs and self-efficacy is that self-efficacy is task-specific, some other differences are also worth mentioning.

The first self-knowledge belief that could be confused with self-efficacy is self-concept. Self-concept is a more general self-descriptive construct that incorporates many forms of self-knowledge and can be seen as the construct that contains the other believes that were mentioned. Self-efficacy can also be confused with self-esteem or self-worth, which consists of emotional reactions to a student's accomplishments like feeling good or bad about themselves after performing a task and whether he or she accepts or respects him or herself respectively. Self-confidence is the belief of a student that he or she has the ability to accomplish goals competently and to their own satisfaction. As mentioned before, self-confidence seems the same as self-efficacy, but is domain specific where self-efficacy is task specific.

Stability indicates how easy or difficulty it is to change the student's self-concept, and depends on how crystallized or fixed the student's self-beliefs are. These self-beliefs become more crystallized with repeated similar experiences. This means the student will believe he or she is bad at math when they regularly fail to complete math problems. And as this happens more and more, it will be harder to convince this student that he or she can be good at math and the student could give up trying to improve at math. Closely connected to this is perceived control, which indicates how much control a student thinks he or she has on the outcome of situations like their performance on a test. The student might for instance believe that the teacher is against him or her and therefore has a low level of perceived control and possibly feel that he or she will fail no matter what he or she does.

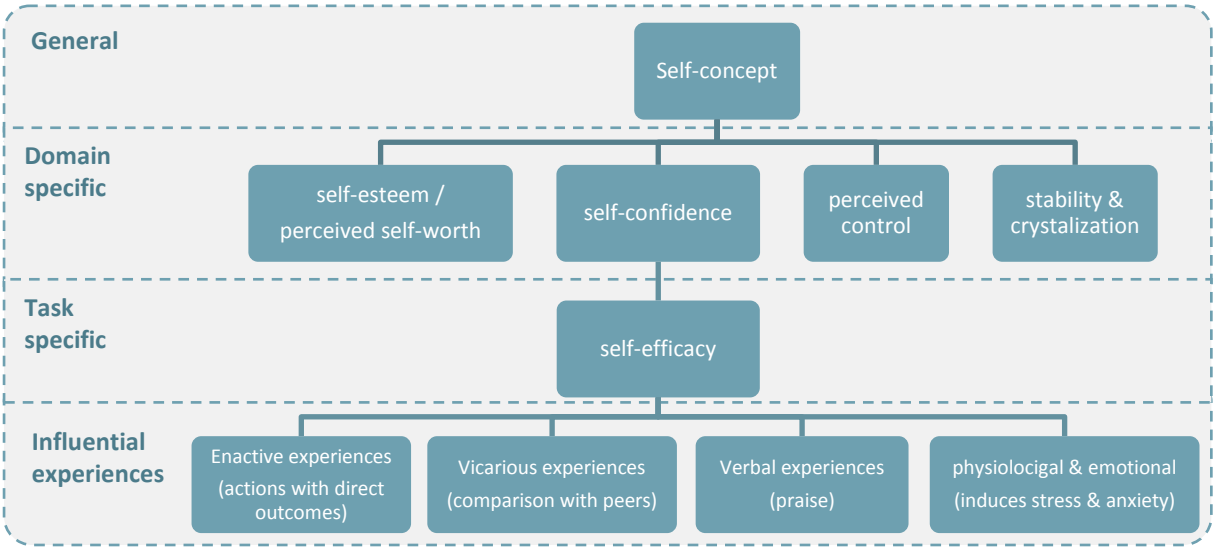


Figure 2: an overview of the hierarchy of self-knowledge beliefs

As is illustrated in the hierarchical overview which is depicted above, self-efficacy is influenced by four types of experiences: enactive experiences, vicarious experiences, verbal experiences and affective reactions. Enactive experiences involve actions that have a clear and direct outcome. These are typically the most influential and account for most of the self-efficacy level for a specific task. Vicarious experiences are comparisons with peers, tutors, teachers etc. This is normally a more environmental influence, since it depends on the people that surround the student. However, when interacting with an ITS or serious game, a tutor or virtual peer in the ITS or characters in the serious game can function as one of these influential people. Verbal experiences are similar, since in this type of experience, the student experiences the outcome through a persuader's description and this again depends on the people that surround the student and can similarly be affected by an ITS. The last influential experience consists of physiological and emotional responses to situations. An example of such a situation is an exam or unannounced pop quiz and can induce stress and anxiety. These responses are typically physically manifested in the form of an increased heart rate and sweaty palms. These experiences are task-specific, like self-efficacy, and can vary for each situation or activity the student experiences. So when a student has a history of negative experiences while trying to solve quadratic equations, this would result in low self-efficacy levels but does not influence the self-efficacy levels of integration problems, where the student might feel very proficient (McQuiggan & Lester, 2006).

In everyday life, the flow of influences would be from the bottom towards the top of the belief hierarchy, where the four types of experiences influence the student's level of self-efficacy. If multiple experiences with similar self-efficacy levels (be it high or low) occur it could influence the student's self-confidence and so on. How much these beliefs influence each other (and if self-efficacy for instance influences perceived control) is not clear and still a subject of debate and in need of more research (Zimmerman, 2000; Schunk, 1991).

Preliminary research did however find that self-efficacy levels seem to influence students' effort, their persistence, how they make choices, how resilient they are when confronted with failure and the level of success they might achieve. Furthermore, self-efficacy is recognized in educational research as a predictor for motivation and learning effectiveness. An ITS (or serious game) that would incorporate an accurate model of self-efficacy could be able to improve the learning process and experience of a student through their self-efficacy levels (Zimmerman, 2000; de Vicente & Pain, 2002).

Linnenbrink & Pintrich (2003) indicate that low self-efficacy levels or negative confidence levels can present themselves as anxiety and cause students to avoid a task or give up more easily than students who have higher levels of self-efficacy. This relation has been used in an ITS by Soldato (1993) and labels students as less confident when they give up in the middle of the task or indicate they would rather do a different task when given the choice. Although Gregersen & Horwitz (2002) indicate that perfectionism can also cause this anxiety, it still causes the same fear of failure, which could be reduced when a student develops a higher level of self-efficacy. Linnenbrink & Pintrich (2003) continue and agree with Zimmerman (2000) that self-efficacy influences emotion in general, where high levels of self-efficacy cause happiness and low levels of self-efficacy cause anxiety. This translates to students working harder and trying longer when they have high levels of self-efficacy and also indicate to be more inclined to ask for or receive help and tutoring, where students are more self-conscious when they have low levels of self-efficacy and indicate to be more reluctant to ask for or receive help because they feel embarrassed. An ITS with the capabilities to diagnose self-efficacy could prevent these situations by increasing self-efficacy levels using appropriate tutorial strategies.

As stated before, the order and magnitude of influences between self-knowledge believes are still the subject of debate and research, but high levels of anxiety have been negatively related to the learning process and learning performance, regardless of a student's age, origin or level of education. Furthermore, strong efficacy beliefs about false or faulty subject knowledge are of negative influence on a students learning process. This means that when students believe they know everything or enough of a subject matter, they will be less motivated for and less attentive during education regarding that subject. Closely related to this is the idea that a student's self-efficacy judgements should be calibrated to reflect their actual performances. This ensures that an ITS does not blindly follow the self-efficacy judgements of the student, since he or she might be overconfident and overestimate his or her skill level regarding a subject. These students are inclined to try tasks that are outside the range of their level of expertise, which might lead to failure and reluctance to try new tasks, in turn decreasing the learning effectiveness. The same applies to students who underestimate

their own skill and only try tasks that they can easily accomplish and therefore are not challenged enough. All this knowledge is important for an ITS or serious game to obtain and incorporate in tutoring decisions or the adaptation of game mechanics, since it explains a student's low motivation and possible reluctance to ask for help (Linnenbrink & Pintrinch, 2003). This situation could easily be inferred by an ITS or serious game when it has an indication of the student's level of self-efficacy and could be corrected using appropriate tutorial strategies. Which tutorial strategy is appropriate and evokes the expected or wanted motivational state within the student is an entirely different challenge and is beyond the scope of this research.

In sum, self-efficacy has been proven to be an important component of a student's learning process and can cause significant improvements in learning and levels of self-efficacy itself when it is used to determine an appropriate tutorial strategy by an ITS (Boyer et al., 2008).

2.4.Measuring of motivation & self-efficacy

But how can self-efficacy be measured so an ITS can choose a tutorial strategy or a serious game adapt its game mechanics? Humans are by nature trained to assess the body language and facial expressions of other humans and read their motivational state (among other traits). For an ITS or serious game, this diagnosis is not so easy. Traditional diagnosis of personal, emotional or motivational states is commonly done using psychometric instruments like questionnaires and self reports. However, the reliability of a questionnaire increases by its length, which in turn increases the time consumption for a student and causes the diagnosis to become an obstruction to the primary task of educating the student (Khan et al., 2008).

The default communication channels in a normal ITS setup are limited to keyboard and mouse input and on-screen output. Some researchers try to overcome these limitations by incorporating additional communication channels in the form of special sensors to measure the physiological responses of a student (Person et al. 1999 in de Vicente & Pain, 2002; McQuiggan & Lester, 2006; McQuiggan et al., 2008). The four most common sensors used in these studies are the Galvanic Skin Response (GSR) Sensor, the Blood Volume Pulse (BVP) sensor, the Respiration sensor and the Electromyogram (EMG) sensor (de Vicente & Pain, 1998). Another example of additional input can be found in the research of Wang et al. (2006), where eye-tracking is used to adopt empathic tutoring agents to a student's reactions and inferred motivational state. For instance, when the student seems to lose concentration, an agent shows mild anger or alerts the student to keep focussed.

But like the questionnaires, although these sensors are suitable for experiments in a laboratorial setting, they do not comply with the requirement of e-learning where a student can use it anywhere

and anytime he or she pleases. It would be more beneficiary to diagnose the motivational state of a student using only the default computer input channels: keyboard and mouse input. Since most ITS's model knowledge of the student to some degree, there is some research that incorporated some form of confidence or self-efficacy measurement that did not require additional sensors.

A straightforward method of determining efficacy is to let the students themselves indicate their own efficacy with regard to a task on a self-efficacy scale. Bandura (2006) drafted a guide to construct valid self-efficacy scales to do just that. These scales are likert-scales from 0 to 100 with intervals of 10. Bandura (2006) emphasizes that the student needs to have a good understanding of the task at hand, before he or she can indicate his or her self-efficacy level regarding that task. Other than this prerequisite, the self-efficacy scale needs to meet the same conditions as other self reports like avoiding socially desirable answers and ambiguous questions.

Self reports are confirmed by Beal & Lee (2005) as an efficient source of information about student states and do not require expensive or intrusive instrumentation or equipment that cannot be used in public school classrooms. (Also, self reports constitute something close to ground truth: if students say that they are in a bad mood, or that they do not feel that they are any good in math, we are inclined to take them for their word.) However, as mentioned before, inquiries of any kind to students may disrupt their learning process and is therefore ill-suited. De Vicente & Pain (2002) point out that even the use of passive self reports (where students can indicate their self-efficacy level whenever they want) pose the risk of being inaccurate, since the student can be to engaged with a task and forget to update the self-efficacy level or attempt to please the tutoring system by providing exaggerated values on the self report.

Soldato (1993) applied a different method and used a numerical value instead of a self-efficacy scale to indicate the confidence of a student at a given moment in time. This value is not indicated by the student, but is incremented or decremented by the ITS depending on the student's response to a pre-task query and the performance during a task. For instance, students might indicate before the beginning of a task that they think the task is to difficult for them, indicating a low efficacy and consequently, the efficacy value is decremented. The value is also decremented when students ask the tutoring system for help before trying to perform the task. Although this method might function for the majority of the students, it relies on a generalised model of motivation and assumes that all students have the same behavioural pattern in a given motivational state i.e. all students with low self-efficacy levels ask for help. But as de Vicente & Pain (1998) indicate, human tutors know that each individual student has its own characteristics and can react differently to the adaptations of the tutor. Although this study will attempt to find general characteristics of self-efficacy in mouse and

keyboard input, the final goal is to facilitate individualized self-efficacy diagnosis at runtime. This means that a future computer enhanced learning system would be able to diagnose self-efficacy for a single student at runtime and based on the behavioural data that was collected for this particular student. But before this can be realized, it needs to be clear what behavioural characteristics need to be collected.

2.5.Mouse and keyboard input

Since it has been established that self-efficacy scales in whatever form (digital or pencil-and-paper) do not fulfil the requirement of measuring self-efficacy without intruding on the learning experience of the student, how can these requirements be met? Some of the previously discussed studies contain indicators that could provide a way to measure efficacy without bothering the student.

To improve assessments in e-learning systems, de Vicente & Pain (2002) did experiments where human tutors evaluated the actions of students (based on their on-screen behaviour alone) and found that confidence was the largest factor in a tutors motivational evaluation of a student. The way the tutors described the on-screen behaviour of the student gives more concrete indicators for the behaviour that they characterised as confident: *"Participant: Well, [...] he is hovering the mouse over the answers each time, he wasn't randomly moving the mouse, he is looking for the answer, [...] and that he didn't take a long time to answer the questions. [...] So, I would increase the satisfaction here, just for the fact that he did it with confidence."* This quote and the findings of de Vicente & Pain (2002) lead to the presumption that the level of self-efficacy of students is reflected in their mouse movements and typing performance. De Vicente & Pain (1998) also inferred this possibility and recommend the exploration of the existing communication channels (i.e. mouse and keyboard input) for future research.

But no research was found that actually studies the characteristics of self-efficacy in mouse and keyboard input. Zimmerman et al (2003) published a rationale for the measurements of mood through mouse and keyboard input while Khan et al. (2008) went in a slightly different direction and found significant correlations between personality tests and mouse and keyboard input. However, since personality is not the same as motivation, the research of Khan et al. (2008) is not comparable to this research. The research of McQuiggan & Lester (2006) and McQuiggan et al. (2008) is the only found research that induces levels of self-efficacy. However, it uses physiological measurements, needs a pre-test, machine learning process and physiological equipment to do so (although the results are impressive) and does not measure mouse and keyboard input.

All in all, no behavioural keyboard and mouse characteristics for self-efficacy were found in prior research. Consequently, general characteristics of mouse and keyboard input were used as variables to establish which characteristics reflect a student's level of self-efficacy. This led to the following hypotheses:

1. ***A student's level of self-efficacy is reflected in his or her mouse cursor movements.***
2. ***A student's level of self-efficacy is reflected in his or her typing performance.***
3. ***A student's level of self-efficacy can be diagnosed using his or her mouse or keyboard input.***

where hypotheses one and two test if there are correlations between levels of self-efficacy and mouse and keyboard input and hypothesis three evaluates if it is consequently possible to predict a student's level of self-efficacy using mouse and keyboard input. This would allow educational systems like serious games or ITS's to diagnose and incorporate self-efficacy levels in their tutoring decisions while using the default communication channels of mouse and keyboard input. The measurement of self-efficacy has been discussed in the previous subsection, but how can these general mouse and keyboard characteristics be measured? Although little to no research was found to extract motivational meaning from mouse and keyboard input in order to assess students like human tutors do, research and analysis on mouse movements and typing performance was found extensively in other research areas.

2.6.Mouse movements

In the Human Computer Interaction (HCI) research area, researchers have been trying to quantify the movement of humans who perform pointing tasks on computers and other devices. Mouse cursor movement is a vital part in this field and used in many studies. One of the measurements used in those studies is Fitts' law for movement time (Fitts, 1954; Fitts & Peterson, 1964). This law originated from the desire to quantify the pointing behaviour of humans and is currently most used to measure if the layout of a user interface is efficient. The efficiency is measured in the time period that is needed to use the interface i.e. click the buttons or navigate within the interface. Fitts' law can be summed up into a single statement: A bigger and closer object (to the cursor), is easier to move to (Fitts, 1954).

Since the HCI research field mainly focuses on the improvement of the interaction between humans and computers, most research that involves mouse movements is done to increase the pointing performance of humans. A good example is Gajos et al. (2012), who used Fitts' law (among other measurements like trajectories and jerk profiles) to obtain lab-quality mouse movement measurements from a normal (domestic) computer setup. They did this to be able to differentiate

between deliberate and accidental mouse movement for people with motor disabilities in order to compensate for their accidental mouse movements during their computer interactions. Although this is a form of user modelling using mouse movements, it does not reveal anything about the motivational state of the user. The same can be said about the research of Pusara & Brodley (2004), who used mouse movements characteristics (distance, angle, and speed) and mouse events (single and double clicks of either a left, right or a middle mouse button, and the mouse wheel movements) to re-authenticate a known user for security purposes. This differentiates between different users through their use of the mouse, but again does not provide any knowledge about the motivational state of that user or their level of expertise in a subject.

Of the research that does describe something similar to the diagnosis of self-efficacy or any other part of the motivational state of a student using his or her mouse input, the first is the research of Khan et al. (2008), who correlated mouse and keyboard usage to the outcome of personality tests. Their measurement of mouse usage was the number of clicks in a given window (during a given mood rating) and the average and standard deviation times between all the events (keyboard and mouse input and window switching) and they did not record any positional or movement data for the mouse. The second article is a rationale of Zimmerman et al. (2003), who indicate they intent to research mood through mouse and keyboard input and will record the mouse coordinates, not just the mouse clicks. Lastly, as mentioned before, the research of McQuiggan & Lester (2006) and McQuiggan et al. (2008) is the only research that induces levels of self-efficacy but uses input from a paper-and-pencil pre-test and physiological sensors and does not record mouse input.

Although these studies do not indicate characteristics of motivation or self-efficacy within mouse input, they do supply general characteristics of mouse input which can be tested for correlations with self-efficacy.

2.7.Typing performance

Like the research into mouse movements, keyboard input was studied in the HCI research area to model and improve keyboard behaviour of humans in different situations and environments. The basis for keyboard performance is the research of Card & Moran (1980), who studied the performance of computer users who were using a keyboard to perform a typing task and measured the number of keystrokes, the time between each keystroke and the number of errors made by those users. Umphress & Williams (1985) and Leggett & Williams (1988) had the same intention as Pusara & Brodley (2004) and studied the potential of keyboard characteristics as a method of user authentication. Like the research of Card & Moran (1980), they measured the number of keystrokes, the time between keystrokes, the number of words per minute and the number of corrections made

by the keyboard user. Lastly, a study comparable to Gajos et al. (2012) is the research of LoPresti et al. (2006), who studied the influence of disabilities on typing performance. This study set out to determine the difference in error occurrences between normal and disabled users in order to compensate for these differences. To study this, the error rates for a number of common errors among users with disabilities were measured. An example is the “Long key press error”, where a key was pressed long enough to generate repeats and which occurred often among users with disabilities but rarely among users without a disability.

Similar to the previously mentioned research regarding mouse input, research could not be found for keyboard input in the context of motivational diagnosis. As previously mentioned, Khan et al. (2008) correlated mouse and keyboard input to personality tests and recorded the number of backspace and delete keystrokes, the number of alphabetic and numeric keystrokes, the number of all other keystrokes and the delay between each of those keystrokes, but again this study is not comparable to this research.

Finally and again as mentioned before, McQuiggan & Lester (2006) and McQuiggan et al. (2008) induced levels of self-efficacy, but did not record any keyboard data. Except these studies, no research was found to infer any kind of emotional state from keyboard input. This lack of research could be attributed to a lack of open ended questions in computer enhanced learning in general, which in turn could be explained by the difficulties that evaluation of open ended questions present at this time. Mödritscher et al. (2006) indicate that assessment of open ended questions is still mostly done by human experts since methods based on artificial intelligence are still too limited. However, Mödritscher et al. (2006) also indicate that open ended questions like sentence completion, short answers and essays are essential to evaluate high-level learning objectives since it requires the students to formulate their own answers and thus assimilate the knowledge they gained. Therefore, the role of the keyboard and open ended questions for the future of computer enhanced learning cannot be ignored. This is why this study also included keyboard input in order to find keyboard characteristics that indicate levels of self-efficacy.

Again, these studies do not supply motivational characteristics within keyboard input, but do supply general characteristics which can be tested for correlations with self-efficacy.

In conclusion: self-efficacy and mouse and keyboard characteristics have been studied separately, but a relationship between the two has not been researched up till now. This study will be conducted to find these relations and provide an indication of the characteristics of self-efficacy levels within mouse and keyboard input. This could in turn facilitate the diagnosis of levels of self-efficacy of

students and allow computer enhanced learning systems to improve the students learning experience and efficiency.

3. Method

To answer the research question stated in the introduction: *To what extent can levels of self-efficacy be diagnosed using mouse and keyboard input in order to improve a student's learning experience?* an experiment was performed to answer this question by proving or disproving the following hypotheses:

1. ***A student's level of self-efficacy is reflected in his or her mouse cursor movements.***
2. ***A student's level of self-efficacy is reflected in his or her typing performance.***
3. ***A student's level of self-efficacy can be diagnosed using his or her mouse or keyboard input.***

Hypothesis 1 and 2 will assess whether a student's level of self-efficacy is reflected in their mouse and keyboard input and in which variables it is reflected, after which hypothesis 3 assesses whether and to what extent these variables can diagnose or predict self-efficacy levels.

The experiment that was performed to test these hypotheses consisted of a mouse and a keyboard part which both consisted of a calibration and quiz section. A schematic overview of the experiment is depicted in the system section, which will discuss its setup in more detail. Although the intention of this research is to facilitate self-efficacy diagnosis for serious games and computer enhanced learning in general, the lack of previous research forced the simplification of the experiment setup in order to prevent interference caused by the possible complexity of an ITS or a serious game. Therefore, a simplistic quiz-like setup was chosen, which is designed to isolate the answering behaviours of the participants but does not offer educational content like an ITS or serious game would do.

3.1.Participants

Participants were drafted from the family, friends and acquaintances of the researcher via email and Facebook. This allowed the participants to invite their own friends and acquaintances to participate, reaching more potential participants than the researcher originally had access to. This resulted in the participation of 60 participants of which 3 participants stopped after the first part of the experiment. The participants with a personal connection to the researcher were not considered biased to this experiment due to several factors. Although some of the participants might have shown extra motivation during the experiment compared to participants without a personal connection, this cannot influence their self-efficacy, only their determination. The personal connection with the researcher also cannot influence the factual knowledge of the participants, so they answered with

the same factual knowledge and self-efficacy as they would during an experiment conducted by a stranger. Furthermore, only a few people knew about the goals and design of the experiment and these people were excluded from participation. To minimize socially desirable answers or results, the participants participated in the experiment anonymously and were only asked for their gender, age and education level (VMBO, HAVO, VWO, MBO, HBO, WO). These demographics were used to evaluate the quality and diversity of the participant population. Lastly, the participants received no compensation of any kind except the gratitude of the researcher.

3.2.System

As mentioned at the beginning of the method section, the experiment consisted of a mouse and a keyboard part, each with a calibration and then a quiz where the participants answered 20 factual knowledge questions. To capture the behaviour of the participants, a number of variables were stored for each interaction with the experiment. These variables will be discussed in the following paragraphs and a summary of these variables is included in the measurements section.

For the purpose of this experiment, a webpage was built (in Dutch) to capture and store all the variables that describe the mouse and keyboard behaviour of the participants. This was a single webpage and used Javascript to show the consecutive parts of the experiment, which was hosted on a privately funded webhosting and was publically accessible through the internet. The recorded data and variables were stored in a text file on the same server, using AJAX calls and a simple PHP log script. A schematic overview of the system within the webpage is depicted on the right.

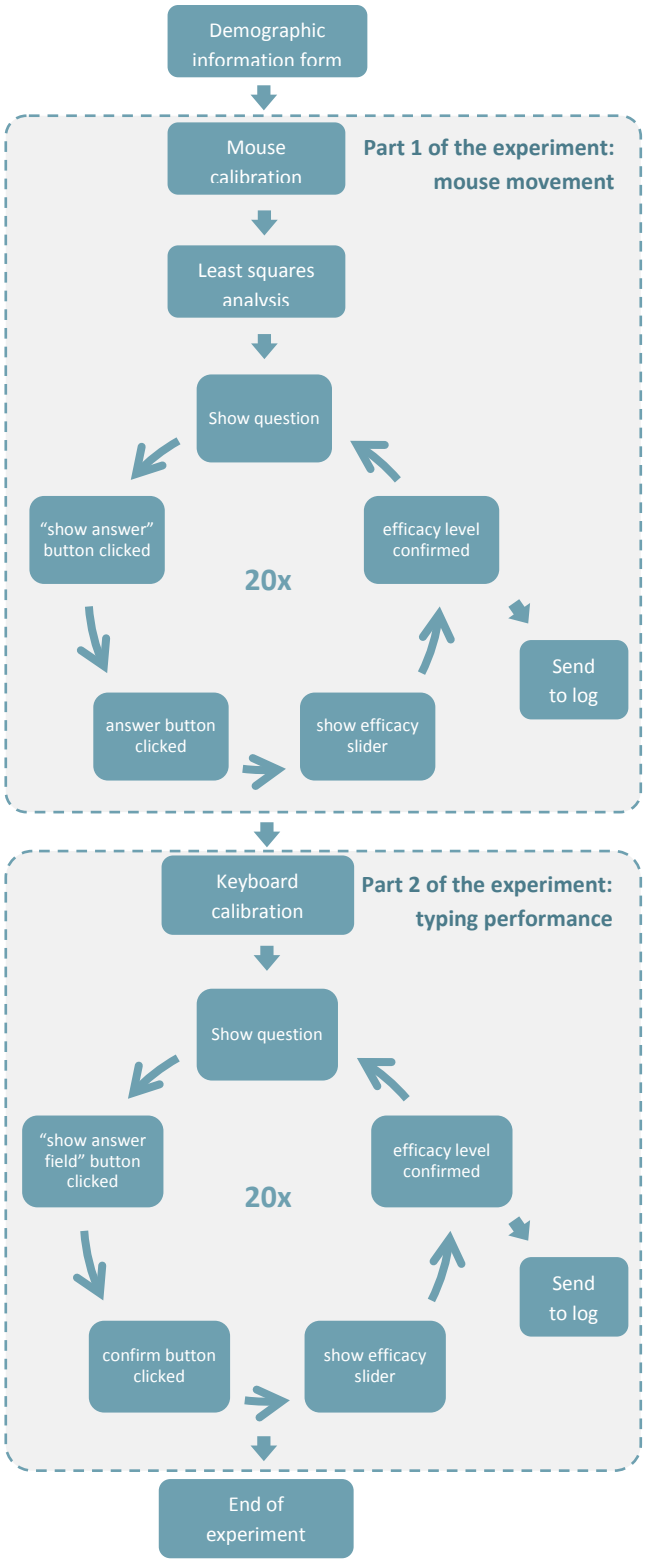


Figure 3: Schematic system overview

Before the actual experiment began, the participants were asked to enter their demographic data in a HTML form which contained input elements for gender, age and level of education. It also contained a self-efficacy slider where the participant could indicate their general self-confidence with regard to answering factual knowledge questions. An image of this form is depicted below.

Figure 4: the demographic data form at the beginning of the experiment

Variable	Description
Date and time	The date and time when a question in the experiment was stored
Gender	The gender of the participant
Age	The age of the participant in years
Education	The educational level of the participant
Pre-test self-confidence	An indication of self-confidence, indicating how confident a participant felt about answering questions in general before the experiment began

Table 1: Demographic variables

The next part of the experiment consisted of the mouse calibration process which was based on the Fitts' law demonstration by Wichary (2005) for the Vrije Universiteit, where the participant was presented with 2 onscreen circles and was instructed to click one circle and then the other, creating a start and endpoint for the Fitts' law calculation. Fitts' law is used to determine the average mouse movement of a user on a particular device, by using the distance and the size of an onscreen target to calculate the average amount of time that this particular user on this particular device would need to point at and click on a target (of a certain size). Simply put, Fitts' law states that a small target that is far away will take more effort to click on than a large target that is close by. Since the onscreen buttons for the answers to the multiple-choice questions in the next part of the experiment were all the same size, the calibration only needed to involve circles of the same size. The participant was presented with 20 pairs of circles to create enough data points to calculate the coefficients for Fitts' law. The original formula for Fitts' law is:

$$MT = a + b\log_2(2A/W)$$

Equation 1: The original Fitts' law equation

However, MacKenzie (1992) improved this formula, changing it to:

$$MT = a + b\log_2(\frac{A}{W} + 1)$$

Equation 2: The improved Fitts' law equation from MacKenzie (1992)

where MT is the movement time, A is the distance between the start point and the target, W is the width of the target and a and b are the coefficients. These coefficients were approximated using Least-squares analysis on the 20 data points from the calibration. The calculated coefficients were stored for future usage during the quiz section for the mouse.

Variable	Description
Fitts' law coefficients	The coefficients within Fitts' law, approximated using Least-squares analysis of the 20 calibration data points.

Table 2: Mouse calibration variables

After the mouse calibration, the participant was presented with 20 multiple-choice questions (one at a time) which he or she could answer by clicking on an answer button which corresponded with the answer they believed to be correct. The layout of the questions resembled the well-known quiz show "So you want to be a millionaire" but to separate the time it took the participants to read each question and the time to actually answer a question by clicking on one of the four answer buttons, the answers for each question were hidden from the participant and were shown when the participant clicked on the "show answers" button. This also provided a relatively constant starting point for the mouse cursor, making comparisons of answering times and travelled distances feasible. An example of the visual layout of a question (after the "show answers" button was clicked to show the answers) is depicted below.

Figure 5: example of the visual layout of a multiple choice question

When the “show answers” button was clicked, a timer started and the position of the mouse cursor was recorded every 100 milliseconds. Like Pusara & Brodley (2004), this interval was based on the assumption that 100 milliseconds is short enough (for humans) not to miss any details from a participants mouse movements, but long enough (for even a slow computer, since the participants participated using their own computer) to allow the computer to perform the collection and possible calculation of each variable at every interval. Although this last assumption was not tested, no evidence was found in the resulting data that proved this assumption wrong. To determine the travelled distance per interval, the Euclidian distance between subsequent mouse cursor positions was calculated and stored along with the mouse cursor positions.

If the Euclidian distance was smaller than two pixels (to prevent minute mouse movements caused by natural hand movements) for at least 200 milliseconds (two intervals), this was flagged as a pause in the mouse movements (*Number of pauses*). When this occurred, the preceding movement (if any) was stored and the length (*Pause time*) and position of the pause was recorded. To provide extra insight in the answering behaviour of the participants, it was monitored and stored whether or not pauses occurred over one of the answering buttons (*Pauses on buttons*). If the mouse cursor moved again, the length of the pause was stored and the mouse cursor positions were stored as movements again. These movements were stored using their start point and end point, the travelled Euclidian distance and the elapsed time in milliseconds.

The multiple-choice questions were designed to alternate in difficulty between easy and difficult, invoking answers with high- and low levels of self-efficacy respectively. The easy questions were based on common-sense knowledge that most (Dutch) participants would possess and would allow them to answer without difficulty and with a high self-efficacy. An example of an easy question is “How many eggs are in a dozen eggs?“, which is common knowledge for most people. The difficult questions were based on “did you know” statements, which is factual knowledge that most participants would not possess and could cause them to doubt their answer, possibly think longer or show non-deliberate mouse movements like moving the mouse cursor around and hovering the answer buttons. In conclusion, the participants could answer with a low level of self-efficacy. An example of a difficult question is “How wide is the mouth of the statue of liberty?” which is clearly not common knowledge for most people.

The level of difficulty of these questions was tested in a pilot experiment, to confirm that the participants could indeed answer the easy questions easily and would have (much) more difficulty answering the difficult questions. This pilot experiment consisted of four participants who would participate in the pilot experiment under supervision of the researcher and provide feedback on

every unclear aspect or element of the experiment. After the feedback of the pilot participants, some questions were replaced with better suited ones, to ensure that the requested level of difficulty was met and the topics of the questions were not biased by the interests of the researcher. The resulting final list of questions can be found in Appendix A. The pilot experiment was also used to fine-tune the written instructions of the experiment, since these instructions were all the information the real participants would have and ambiguities could cause participants to be confused or unclear in what they were asked to do.

When a participant answered the multiple choice question by clicking on an answer, the position of the mouse, combined with the previously stored start position (on the "show answers" button) were used to calculate the shortest distance from start point to endpoint. This was the Euclidian distance between these two points and represented the fastest way to move from the start position to the end position (on the chosen answer button). The stored Euclidian distances for each interval (of 100 milliseconds each) during the mouse movement, were summed to create the actual travelled distance of the mouse cursor and the difference these two distances was also stored (the *Distance difference*).

The end position, combined with the calculated Fitts' law coefficients and the size of the chosen answer button, was also used to calculate the Fitts' law movement time, indicating the average time it would take this person, on this device to move the mouse cursor from the start position on the "show answers" button to the end position on the chosen answer button. The calculated movement time represented the time in which the participant would move the mouse cursor from the starting point to the ending point in a normal situation, without hesitation or other factors that would influence the movement time. When the participant clicked an answer button, the elapsed time (since the "show answers" button was clicked) was stored and the difference between the predicted "normal" mouse cursor movement time and the actual time it took the participant to move the mouse to the chosen answer and click was stored (the *Time difference*).

Variable	Description
Question time	The time period from the moment the question was shown up to the moment the participant clicked on an answer
Distance difference	The difference between the shortest distance and the total distance.
Time difference	The difference between the Fitts' predicted time and the actual time.
Read time	The time the participant waited before clicking the "show answers" button
Number of pauses	The number of pauses during the movements of the mouse cursor, where a pause is the lack of movement for longer than 200 milliseconds
Pause time	The total pause time for a particular question
Pauses on buttons	The number of pauses that occurred on an answering button

Table 3: Mouse question variables

For clarification a visual representation of these variables with regard to a multiple choice question is depicted below.

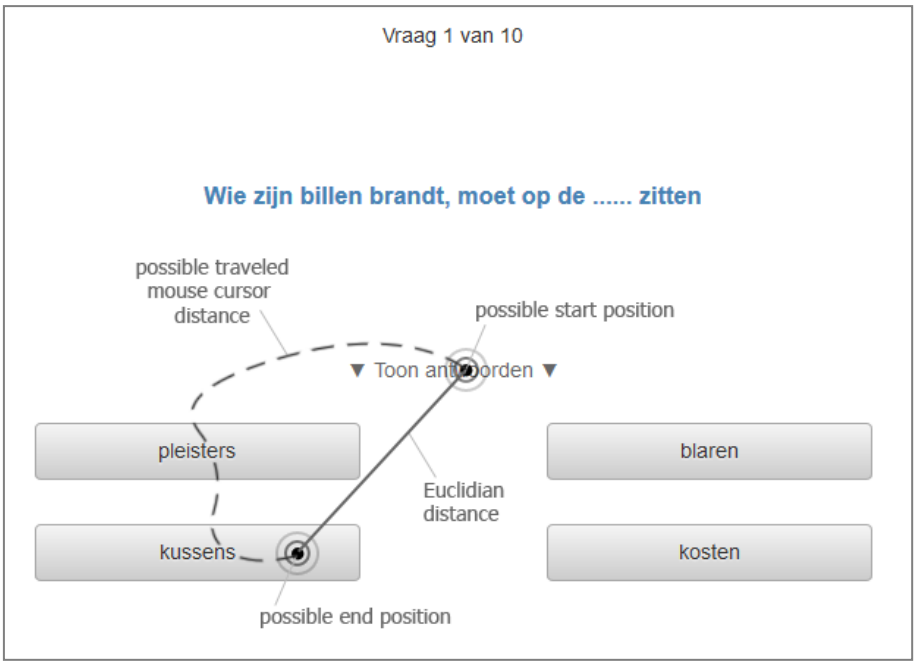


Figure 6: A visual representation of the variables obtained from the multiple choice questions

After each question, a self-efficacy scale (created using the guidelines set by Bandura (2006)) was presented which allowed the participants to indicate their level of self-efficacy from 0 to 100, with intervals of 10. To label the scale, but not influence the participant in their sense of self-efficacy, only the 0 and 100 values on the scale were labelled with “Ik gokte” (“I guessed”) and “Ik wist het zeker” (“I was certain”) respectively. The scale was presented as an onscreen horizontal slider, allowing the participant to move the handle to the predefined positions (intervals of 10) between 0 and 100. This scale is depicted below.

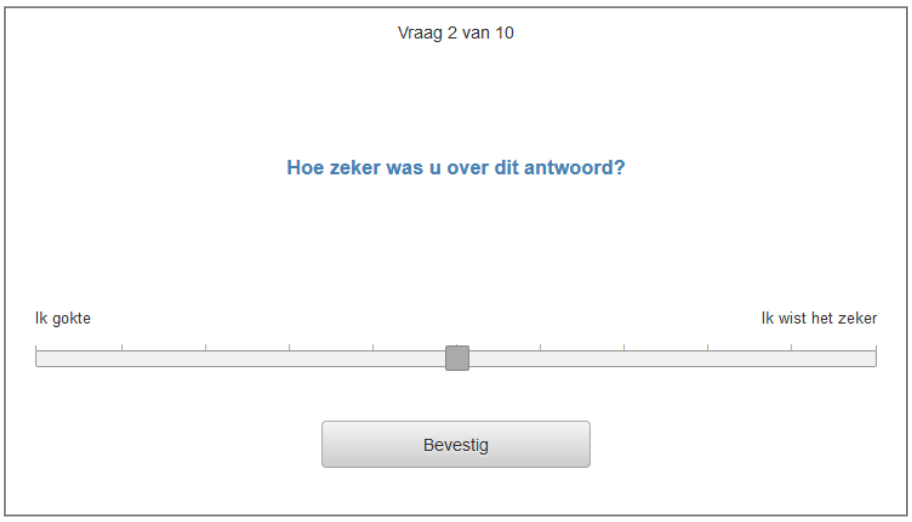


Figure 7: The self-efficacy scale used in the experiment

At the moment the self-efficacy scale was presented, another timer was started to record the time it takes the participant to indicate their self-efficacy level on the scale. Next to this time, the number of times the participant changes the handle of the slider was stored, so when a participant confirmed their self-efficacy level for a question, the time, number of changes and the value of the slider were stored.

Variable	Description
Self-efficacy value	The level of self-efficacy as indicated by the participant for a particular question
Self-efficacy time	The time the participant needed to indicate their self-efficacy level.
Self-efficacy changes	The number of times the participant changed their indication of self-efficacy.

Table 4: Self-efficacy scale variables

When the participant confirmed their self-efficacy level, all the previously recorded and stored variables were sent to the web server to be appended to the log file, creating a line of variables for each answered question, preceded by a timestamp and the demographic data of the participant for participant differentiation in the case two participants generated results at the same time. After the results were sent to the web server, the next question in line was presented, until the participant answered all of the 20 multiple choice questions and completed the first part of the experiment.

Now the participant was presented with the second part of the experiment, regarding keyboard input. This second part of the experiment consisted of a calibration and 20 fill-in-the-blanks questions, where the participants used their keyboard to type their answers in a blank answering field. The calibration was used to calculate average values of typing speed and accuracy for this particular participant and consisted of an example text that the participant duplicated by typing it into the blank answering field.

During the typing process, the delay between two keystrokes was recorded and the average of these delays is calculated and stored (*Average keystroke delay*). All the keys that were pressed were also stored, making it possible to follow the typing (and correcting) behaviour of the participant. When the participant used the backspace key, this was considered a correction and the number of these corrections was stored separately (*Number of corrections*). Lastly, the total typing time (*Type time*) was stored and the typed text was compared with the example text using the Levenshtein distance. The Levenshtein distance is a string metric to measure the difference between two words. It results in the minimum number of single-character alterations required to change one word into the other, indicating the number of uncorrected errors made by the participant. Although Soukeroff (2003) improved on this metric, it was considered a sufficient error rate for this experiment, since the number of errors in the final answer is only one of many factors that could indicate a low self-efficacy and was suspected to be less decisive than the keystroke delay or the number of corrections made.

When the participant completed the keyboard calibration i.e. completed copying the example text and clicked the button to confirm their entry, the first fill-in-the-blank question was presented. While the participant entered their answer, the same variables as for the calibration are stored: the number of keystrokes, the latency between those keystrokes, the number of corrections made, the Levenshtein distance and the total answering time. When the participant then clicked the confirmation button, he or she was presented with a self-efficacy scale similar to the first part of the experiment. Like before, when the participant confirmed their indication of their self-efficacy level using the confirmation button, all the stored variables were sent to the web server and logged.

Variable	Description
Question time	The time period from the moment the question was shown up to the moment the participant clicked on an answer
Read time	The time the participant waited before clicking the “show answers” button
Number of keys	The number of keys the participant pressed. This can deviate from the number of characters in the answer, since the participant could have made corrections.
Answer string	The answer the participants typed
Average keystroke delay	The average time between the individual keys the participant pressed to type his or her answer
Number of corrections	The number of corrections made by the participant (using the backspace key)
Type time	The time the participant needed to type their answer
Levenshtein distance	The number of characters in the answer of the participant which need to be changed to result in the correct string of characters of the correct answer, indicating the error rate of the given answer

Table 5: keyboard question variables

3.3.Measurements

A number of variables were recorded in the experiment. However, not all variables are expected to be relevant for the diagnosis of self-efficacy levels. Using the hypotheses stated in the related work and method sections, the following variables are expected to be of influence in the assessment of the hypotheses and in turn the diagnosis of self-efficacy. Although most of these variables have been discussed in the previous section, a short description of each variable is given for clarity, starting with the self-efficacy value and continuing with the variables for each input separately.

Self-efficacy value

This is the level of self-efficacy indicated by the participant regarding their answer to a specific question. This represents the certainty a participant felt when he or she answered that question.

Pre-test self-confidence value

This is the level of self-confidence indicated by the participant regarding their general answering capabilities, indicating how sure a participant felt about answering questions in general. This is a self-confidence level and not a self-efficacy level, since this indication is given regarding general question

answering, which is on the domain level, not on the task-specific level of self-efficacy, as explained in section 2.3 regarding motivation and self-efficacy.

Self-efficacy reflected in mouse movements

Distance difference

This is the difference in distance between the shortest (Euclidian) distance (between the start point where the participant clicked on the "show answers" button and the end point where the participant clicked the answer of their choice) and the distance the mouse cursor actually travelled, both depicted in figure 6 in the previous section. This variable represents the extra or unnecessary movement a participant might use to choose his or her answer.

Number of pauses

The *Number of pauses* indicates the number of times where the mouse cursor did not move for longer than 200 milliseconds and for no more than two pixels. Each pause was timed, which contributed to the *Pause time*. The *Number of pauses* variable represents the times when a participant might for instance be thinking, reading or in doubt of his or her answer.

Pauses on buttons

For every pause, the location of the mouse cursor was compared to the locations of the answering buttons and if a pause occurred on one of these locations, the *Pauses on buttons* variable was incremented. Therefore this variable indicates the number of pauses that occurred on one of the answering buttons and could indicate a situation where a participant chose one of the answers but changed his or her mind.

Time difference

The *Time difference* is the difference between the actual movement time and the Fitts' law movement time (i.e. the average amount of time that this participant on this device would need to move the cursor to the answering button as predicted using the coefficients acquired from the mouse calibration). This difference represents the extra time needed for a participant to move the mouse to an answer button compared to their normal or average mouse movement time and might for instance indicate when a participant moves the mouse cursor more slowly than usual.

Pause time

The *Pause time* is the accumulated time of the number of pauses. Each pause lasts a period of time larger than 200 milliseconds but this can vary for each pause. The *Pause time* indicates the total amount of time the participant paused during the answering of a single question. Like the *Number of pauses*, this variable could indicate when a participant might be reading or thinking.

Read time

Although the *Read time* is technically not a part of the answering time, it could indicate if a participant takes longer to read a question that he or she finds difficult and for instance reads the question twice. The *Read time* starts when the question is shown and ends when the participant clicked the "show answers" button.

Question time

The total time a participant needs to answer a single question is represented in the *Question time* and might for instance indicate that participants take longer to answer difficult questions. The *Question time* starts when the question is shown and ends when a participant clicks on an answer.

Self-efficacy reflected in typing performance

Number of corrections

A commonly used component of typing performance is the number of corrections a keyboard user performs while typing a word or sentence and indicates if the participant corrected his or her answer and how many times they needed to correct. This variable could also indicate when a participant changes his or her mind and erases their answer entirely and starts over.

Relative average keystroke delay

The Relative average keystroke delay is the average time between keystrokes while answering a question, relative to the average time between keystrokes from the keyboard calibration. This ensures that the average keystroke delays are comparable between participants. It could for instance indicate if a participant took longer to enter certain characters but also could indicate if he or she changed her mind or needed extra time to think.

Type time

The Type time is the total amount of time needed by the participants to type their answer and is measured from the moment the participant clicks the "show answer field" button until he or she clicks on the "confirm" button to confirm their answer.

Read time

The read time for the keyboard is measured in the same manner as the *Read time* for the mouse movements and describes the time needed for the participants to read the question before they click on the "show answering field" button.

Question time

Like the *Read time* variable, the *Question time* variable is measured in the same manner as the

Question time variable for the mouse movements.

A complete list of all the collected variables can be found in appendix B.

Diagnosing self-efficacy

To assess to what extent self-efficacy levels can be diagnosed using mouse or keyboard input, linear multiple regression analysis was used to predict self-efficacy levels when significant correlations were found. In this analysis, the self-efficacy levels were the dependent variable and the variables from the previous two subsections were the independent variables. This means that when large and significant correlations were found for mouse input, the linear multiple regression model for mouse input attempted to predict self-efficacy levels using the *Distance difference*, *Number of pauses*, *Pauses on buttons, Time difference, Pause time, Read time and Question time.* Likewise, when large and significant correlations were found for keyboard input, the linear multiple regression model for the keyboard input attempted to predict self-efficacy levels using the *Relative average keystroke delay, Number of corrections, Type time, Read time and Question time.*

3.4.Data analysis

For the data analysis, the statistical software SPSS from IBM was used. Using SPSS, the results from the experiment were first cleared of invalid values like negative values on a timescale. Then extreme outliers that were a clear result of unwanted behaviour during the experiment were removed. These outliers consisted mostly of extremely long answering times (for instance an answering time of 12 minutes in relation to an average answering time of less than a minute) and unrealistic distance values (again compared to average distance values). To be able to compare the values of every participant, these values needed to be relative to the calibration values. This means that only the differences in time, distance etc. were compared and not the measured times and distances themselves, since they are specific and unique for each participant. For these variables, Pearson product-moment correlations were calculated to assess the relationship between the variable and the self-efficacy levels which were indicated by the participant.

Depending on the correlations found in the first part of the analysis, linear multiple regression will be used to assess the ability of the variables to diagnose or explain levels of self-efficacy. These analyses will be done for the mouse and keyboard input separately, since the question types and collected variables were not comparable. Preliminary analyses were conducted and transformations were performed to ensure no violations of the assumptions of normality, linearity, multicollinearity and homoscedasticity. These transformations can be found in appendix D. The results from these analyses are presented in the next section.

4. Results

As discussed in the related work section, no research was found that diagnoses self-efficacy levels using mouse and keyboard input and consequently no characteristics are known of these levels. The experiment that is described in the previous section was performed to find these characteristics and assess to what extent those characteristics could diagnose self-efficacy levels.

Self-efficacy reflected in mouse movements

To assess whether levels of self-efficacy are reflected in mouse movements and which variables are characteristic for self-efficacy, the Pearson product-moment correlation coefficient was used to determine the relationship between self-efficacy levels and the mouse movement variables (*Distance difference*, *Number of pauses* and *Pauses on buttons, Time difference, Pause time, Read time and Question time*). The results are summarized in Table 6.

Table 6: Correlation coefficients between mouse movement variables and self-efficacy levels

Variable	Pearson Correlation coefficients	Significance (2-tailed)
Time difference	- 0.369	0.000
Pause time	- 0.332	0.000
Number of pauses	- 0.286	0.000
Question time	- 0.267	0.000
Distance difference	- 0.148	0.000
Read time	- 0.020	0.503
Pauses on buttons	- 0.032	0.279

Of these correlations, the correlation coefficients of the *Time difference* (the difference between the actual time and the predicted Fitts' law time based on the data from the mouse calibration) and the *Pause time* were the largest (of medium strength according to Pallant (2011: p.134)) with coefficients of -0.369 and -0.332 respectively. The correlation coefficients of the *Number of pauses*, *Question time* and *Distance difference* were small with values of -0.286, -0.267 and -0.148 respectively. All the correlations were negative in direction and significant at a level of $p < 0.001$. The two variables that did not have a significant correlation with self-efficacy, were the *Read time* (-0.020 with $p = 0.503$) and the *Pauses on buttons* variables (-0.032 with $p = 0.279$).

Self-efficacy reflected in typing performance

Similar to the assessment for mouse movements, keyboard typing performance was assessed for characteristics of self-efficacy by using the Pearson product-moment correlation coefficient to determine the relationship between self-efficacy and the typing performance variables (*Relative average keystroke delay*, *Number of corrections, Type time, Read time and Question time*). The results are summarized in Table 7.

Table 7: Correlation coefficients between typing performance variables and self-efficacy levels

Variable	Pearson Correlation coefficients	Significance (2-tailed)
Read time	0.033	0.284
Type time	- 0.026	0.389
Number of corrections	- 0.019	0.537
Question time	0.018	0.557
Relative average keystroke delay	- 0.011	0.557

In contrast to the mouse movement results, no significant correlations with self-efficacy were found for any of the typing performance variables. This means no significant correlation was found for the *Relative average keystroke delay* (-0.011 with p = 0.557), the *Number of corrections* (-0.019 with p = 0.537), the *Type time* (-0.026 with p = 0.389), the *Read time* (0.033 with p = 0.284) or the *Question time* (0.018 with p = 0.557). Along with the insignificance of these relationships, the correlation coefficients are almost undetectable, further confirming the lack of relationship between typing performance and self-efficacy levels.

Diagnosing self-efficacy

The correlation coefficients indicate that mouse movements do reflect self-efficacy levels but typing performance does not. Linear multiple regression was used to assess the predictability of self-efficacy using the variables for mouse input (*Distance difference*, *Number of pauses*, *Pauses on buttons, Time difference, Pause time, Read time, Question time*). During preliminary analyses to ensure no violation of the assumptions of normality, linearity, multicollinearity and homoscedasticity, the independent variables *Time difference, Pause time* and *Question time* correlated highly (more than 0.70) with each other, violating the multicollinearity assumption (Pallant, 2011). Since the *Time difference* had the strongest correlation with self-efficacy and intuitively indicates differences in mouse behaviour, *Time difference* was chosen as the time variable in the linear multiple regression analysis. Furthermore, to prevent violations in normality, the variables were transformed using logarithms or square roots to improve their normal distribution. These transformations can be found in appendix D. The linear multiple regression using the transformed variables resulted in a model which explains 17% of the variance of self-efficacy ($R^2 = 0.175$, adjusted $R^2 = 0.172$, $p < 0.001$). The beta values are summarized in Table 8.

Table 8: Beta values for the linear multiple regression for mouse movement

Variable	Standardized Beta Coefficient	Significance
sqrt Time difference	- 0.415	0.000
log Pauses on buttons	0.060	0.041
log Distance difference	0.038	0.201
log Number of pauses	- 0.030	0.410

Time difference had the largest and largest significant beta value in the model with a negative beta value of -0.415 and a significance at a level of $p < 0.001$. The other variable with a significant beta value was the *Pauses on buttons* (with $p = 0.041$), but the value of 0.060 indicates that its contribution was negligible. The *Distance difference and Number of pauses* were not significant with beta values of 0.038 (with $p = 0.201$) and -0.030 (with $p = 0.410$) respectively.

Given the lack of significant correlations for all typing performance variables, no linear multiple regression analysis was performed to assess whether the variables of keyboard input could predict or diagnose self-efficacy.

5. Conclusions & Discussion

5.1. Conclusions

This research was conducted to answer the question whether levels of self-efficacy can be diagnosed using mouse and keyboard input at runtime and in an unobtrusive way. Small to medium significant correlations between self-efficacy and mouse movement variables were found and the linear regression model contained significant beta values**.** No significant correlations were found between self-efficacy and typing performance variables and therefore no linear regression analysis was performed for typing performance. The following subsections will discuss the results of the experiment in further detail and in the same order as the hypotheses and the results section.

Self-efficacy reflected in mouse movements

As presented in the results section, the *Distance difference*, *Number of pauses*, *Time difference*, *Pause time* and *Question time* all had significant negative correlations of small to medium strength with self-efficacy. The negative correlation coefficients indicate that one variable is small when the other variable is high and vice versa i.e. if the time difference increases, the self-efficacy will decrease. Given the small to medium strength of the correlations between mouse movement variables and self-efficacy, hypothesis 1, stating that self-efficacy levels are reflected in mouse movements, is proven.

Only the correlation coefficients between the independent variables *Read time* and *Pauses on buttons* and the dependent variable self-efficacy were barely detectable and insignificant. The lack of correlation between the *Read time* and self-efficacy might be explained by the fact that the answers for the question were not visible during the *Read time* period and therefore the participant did not know what his or her answering options were and how self-efficacious they were about those answers.

Although it was expected that the *Distance difference* variable of mouse movements would reflect self-efficacy levels more strongly, a possible explanation for this conclusion could be that participants first choose an answer out of the possibilities and then moved their mouse cursor directly to the answer they had chosen. A visual reconstruction of the mouse movements seems to confirm this. Most pauses occurred in the vicinity of the “show answers” button or in the beginning of the mouse trajectory, after which the mouse cursor moves directly to an answer with no or only a few pauses.

Lastly, the “show answers” button, which was introduced to separate the act of reading from the act of answering, seemed to have worked, although a small portion of the answering time is likely to be attributed to the time needed for reading the answers.

Self-efficacy reflected in typing performance

Although significant correlations were found for mouse movement, there were no significant correlations found between typing performance and self-efficacy. Therefore, hypothesis 2, stating that self-efficacy levels are reflected in typing performance, is disproven.

Unrelated to the two hypotheses, but relevant to mention was the lack of correlation between the pre-test self-confidence levels and self-efficacy levels. This indicates that there was no relation between the (domain-specific) self-confidence level for general question answering and the (task-specific) self-efficacy levels for specific question answering. This confirms the research of Bandura (2006), which indicates that self-efficacy is task related and self-confidence is domain related and these two self-believes could be different for different tasks.

Diagnosing self-efficacy

In order to assess whether the characteristics of mouse movements could predict self-efficacy levels, linear multiple regression analysis was used. Using the variables for mouse movement as independent variables and self-efficacy levels as dependent variable, the resulting model explained 17% of these self-efficacy levels. In this model, the *Time difference* had the largest beta value and proved that self-efficacy levels can be partially predicted and therefore diagnosed using mouse movements. The significant beta values within the model indicate that the main indicator for self-efficacy seems to be time, since the three largest contributors for the linear multiple regression model were measures of time (*Time difference*, *Pause time* and *Question time)*. Despite their correlations with self-efficacy, the linear multiple regression model indicates that the *Distance difference* and *Number of pauses* are not good indicators for - or measure of self-efficacy, since both variables did not have significant beta values in the model used to predict self-efficacy.

Since there were no significant and/or strong correlations between typing performance and self-efficacy, linear multiple regression was not used and no prediction could be done to test the ability of typing performance to predict self-efficacy. In the end this means that hypothesis 3, regarding the diagnosis of self-efficacy levels can only be confirmed for mouse movement.

This means that self-efficacy levels could be diagnosed using mouse input, allowing computer enhanced learning systems to incorporate the self-efficacy level of a student into the decision making process when determining how to improve the learning experience of the student. In this diagnosis, self-efficacy levels present themselves most in the time it takes the student to answer the question and less in the movements of the mouse cursor.

5.2.Discussion

The goal of this research was to determine whether levels of self-efficacy are reflected in mouse and keyboard input and to what extent these variables can diagnose self-efficacy levels at runtime and in an unobtrusive way. It found significant correlations between mouse movements and self-efficacy levels but no correlation between typing performance and self-efficacy levels. This is somewhat consistent with the findings of Khan et al. (2008), who found that the trait Conscientiousness (with self-efficacy as a sub-trait) had the largest correlation with mouse clicks.

As mentioned in the method section, the lack of previous research forced the simplification of the experiment design in order to prevent interference caused by the complexity of an ITS or a serious game. In this simplified design, no learning content was offered but the participants were assumed to possess a certain amount of factual knowledge to answer the questions. Although McQuiggan & Lester (2006) also used a multiple-choice answering method, an online genetics tutorial was used as a basis for the participant's knowledge, which is a better approximation of an ITS or serious game environment. Similar to this study, McQuiggan & Lester (2006) implemented the self-efficacy slider after each question to allow participants to indicate their self-efficacy level, despite the guideline of Bandura (2006). This guideline indicates that self-efficacy reports should precede the actual task and requires the participant to understand the nature of the task in order to provide an accurate indication of his or her self-efficacy level. The effects of the differences in methodology are not clear and are in need of further research.

This research excluded the answering performance of the participants from the correlation calculations and linear multiple regression analysis. This was done because the questions in the experiment were designed to invoke answers with high- and low self-efficacy levels by alternating in difficulty between easy and difficult, respectively. This design consequently creates a relationship between the questions (and their answers) and levels of self-efficacy. This relationship is assumed to be reflected in the correlations between the answering performance and self-efficacy for both mouse and keyboard input. These answer performance variables were the *Answer* (indicating the answer the participant clicked) for the mouse movements and the Levenshtein distance (which indicates the number of errors made by counting the number of characters in the participants answer that need to be changed to transform it into the correct answer) for the typing performance. These variables both did show the expected relationship with self-efficacy with correlation coefficients of 0.683 (at a level of $p < 0.001$) and -0.526 (at a level of $p < 0.001$) respectively. This means that high self-efficacy levels associated with correctly answered questions in most cases for mouse movements and higher levels of self-efficacy were associated with fewer errors in the second part of the experiment. When the participants' answering performance was included in the linear multiple regression analysis for

mouse movements, it explained 52% of the variance of the self-efficacy levels ($R2 = 0.526$, adjusted $R2 = 0.524$, $p < 0.001$) as opposed to the 17% of the model without answering performance. In the model the other variables had the same beta values as for the original model, indicating that their contribution to the model remained the same. Since the inclusion of the answering performance variables greatly increases the accuracy of the linear multiple regression model, it is worthwhile to investigate the role of answering performance further in future research.

Although these answering performances were not used in this research, it proved to be a difficult task to evaluate open-ended questions, which is also indicated by Mödritscher et al. (2006). The Levenshtein distance proved to be a strict assessment method for open questions when compared to evaluations by human tutors which represents the majority of evaluations in e-learning today. To illustrate this, there were 86 cases with different notations or variations, which would be considered correct by a human tutor. Examples of this are: “geurstof” instead of “geur”, “nieuw zeeland” where “nieuw-zeeland” was correct and “vegetariers” which should be “vegetariërs”. This could have been prevented by using multiple notations of the correct answer or by filtering special characters, however, although a more robust assessment method is needed, this is beyond the scope of this research.

Furthermore, there were a number of participants who used a different correction method than the backspace key. These different correction methods were not foreseen nor accounted for in the design of the experiment. Consequently, these corrections were not recorded and the participants were never instructed to correct using the backspace key. The difference in the number of keystrokes and the length of the resulting answer string gave an indication of the number of correction but had no significant relation with self-efficacy levels.

Unexpectedly, the *Average keystroke delay* from the calibration was almost always longer than the A*verage keystroke delay* during the answering process. This however, could be attributed to the calibration process of copying text compared to self-constructed texts, since the brain acts as a buffer during text reproduction but can only hold a small number of characters at a time (Leggett & Williams, 1988). This means that the calibration for the keyboard input, where the participants copied an onscreen text, required the participants to read a sequence of characters, remember these characters, type them and then read the next sequence of characters. This produced a higher *Average keystroke delay* than a participant who types an answer from memory.

The research of McQuiggan & Lester (2006) and McQuiggan et al. (2008) was the study most similar to this study, as mentioned earlier in the related work section, and induced self-efficacy levels using a pre-test, machine learning process and physiological sensors for heart-rate and skin resistance. The

naïve Bayes model and the decision tree (both machine learning methods) correctly classified respectively 72% and 83% of the cases without the physiological sensors. Despite this impressive result, this research did not use mouse and keyboard input, but based its predictions on the machine learned models at run-time. In future research, these methods could be combined to create a best of both worlds method which could correctly classify levels of self-efficacy without the need for pre-test information.

5.3.Future work

This research confirmed the difficulties of assessing open-ended questions as indicated by Mödritscher et al. (2006), who indicate that learning results from these open-ended questions cannot be measured using technology-based methods without hard efforts. Further research to evaluate the results of this research should find a more robust measurement method for both the error rate as the number of corrections. Further research could also analyse more characteristics of mouse movements and for instance account for target overshooting, jerk profiles, changes in direction and speed etc. It could also be beneficiary to collect more data from a smaller group of participants to model the mouse movement for each participant separately and in more detail in order to find user specific patterns.

More research is also needed to determine the role of answering performance for the diagnosis of self-efficacy. Next to research in an environment where the questions were not designed to induce levels of self-efficacy, research could be conducted to calibrate a student's level of self-efficacy to their performance. For instance, 7.9% of the answered questions regarding mouse movement and 14.4% regarding typing performance were indicated by the participant with a high self-efficacy level (of 60 or higher) while answering incorrectly, possibly indicating false confidence or overconfidence. Some of the typing performance cases could be explained by the difficult assessment of the open-ended questions and only approximately 1% of the participants answered incorrectly while indicating a self-efficacy level of 100. However, this observation should still be the subject for future studies, since Linnenbrink & Pintrinch (2003) indicate that overconfident students show behaviour that decreases their learning effectiveness and efficiency. This also applies to students that underestimate their own abilities. In this study, 15.6% of the participants in the first part of the experiment indicated a low self-efficacy level (of 50 or lower) while answering correctly and 6.3% even indicated a self-efficacy level of zero while answering correctly. However, one could argue that these participants just guessed correctly and were lucky. In either case, more research is needed to provide more insight and depth to the diagnosis of self-efficacy and motivation in general.

As this research was meant to provide a first indication for the diagnosis of self-efficacy using only mouse movement and keyboard input, further research will be needed to validate its results in similar and more complex situations like ITS's and/or serious games.

For example, existing serious games where the mouse is used to make decisions, like the "Code Red Triage" serious game from the research of Van der Spek (2011), could be extended to track mouse movements and answering times. This game uses an in-game menu (depicted on the right) to allow players to choose the appropriate actions in the triage process for casualties of a mass casuality disaster. To extent a game to research self-efficacy diagnosis, the variables with significant correlations from this research and a players self-efficacy levels need to be recorded.

Figure 8: In-game triage menu (van der Spek, 2011)

The best method to implement is to first inform the player of the situation, then let the player indicate his or her self-efficacy value regarding the task at hand and then record his or her mouse behaviour in the situation itself. However, occasionally it is difficult to inform the player of a situation without giving away information that might aid the player or give away the solution to the problem. In that case, the self-efficacy indication by the player could occur after the answering behaviour.

The behavioural variables that were relevant in this study were the *Distance difference*, *Number of pauses*, *Time difference*, *Pause time* and *Question time*. These variables can easily be recorded in a game like "Code Red Triage" using a similar implementation as the experiment of this study. When a player is presented with a casualtee, he or she is informed of the state of the casualtee and the environment in which the triage is taking place (enough light and visibility etc.). At this point a self-efficacy scale could be presented to let players indicate their level of self-efficacy regarding the triage actions for this person. When the player has indicated his or her self-efficacy level, he or she performs the triage actions in the order they think is correct. During these triage actions, the previously mentioned variables are recorded and stored. When the game is over, the self-efficacy values can be correlated to the mouse behaviour to evaluate the results of this study.

These alterations could be done to most serious games and ITS's that use the mouse for decision making, question answering or problem solving. Initially this future research would still need to incorporate an indication of self-efficacy by the player or student in order to further establish the relationship between mouse input and self-efficacy, but once this relationship has been established, the impact of the diagnosis of self-efficacy on tutorial strategies or game mechanics can be researched.

Self-efficacy has been proven to influence emotions, where high levels of self-efficacy can cause happiness and low levels of self-efficacy could cause anxiety. This translates to students working harder and trying longer when they have high levels of self-efficacy and also indicate to be more inclined to ask for or receive help and tutoring, whereas students with low levels of self-efficacy are more self-conscious and indicate to be more reluctant to ask for or receive help because they feel embarrassed (Linnenbrink & Pintrich, 2003). An ITS or serious game that would be able to correctly diagnose self-efficacy levels and use appropriate tutorial strategies or game mechanics to improve these levels, could improve the learning experience and efficiency of a student or player.

References

Bandura, A. (2006). Guide for constructing self-efficacy scales. In: Pajares, F., & Urdan, T. C. (red). *Self-efficacy beliefs of adolescents* (pp. 307–337). Charlotte: Information Age Publishing.

Beal, C. & Lee, H. (2005). Creating a pedagogical model that uses student self reports of motivation and mood to adapt ITS instruction. *Workshop on Motivation and Affect in Educational Software, in conjunction with the 12th International Conference on Artificial Intelligence in Education*.

Boyer, K. E., Phillips, R., Wallis, M., Vouk, M., & Lester, J. (2008). Balancing cognitive and motivational scaffolding in tutorial dialogue. *Proceeding. ITS '08 Proceedings of the 9th international conference on Intelligent Tutoring Systems*: 239-249.

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM* 23(7): 396-410.

Chen, S., & Michael, D. (2005). Proof of learning: Assessment in serious games. *Gamasutra.* Retrieved from http://www.gamasutra.com/features/20051019/chen_01.shtml at 4th of June, 2013.

Conati, C. (2009). Intelligent tutoring systems: new challenges and directions. *Proceeding IJCAI'09 Proceedings of the 21st international joint conference on Artificial intelligence*: 2-7.

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety: Experiencing flow in work and play*. San Francisco: Jossey-Bass.

Derry, S. J., & Potts, M. K. (1998). How tutors model students: A study of personal constructs in adaptive tutoring. *American Educational Research Journal 35*(1): 65-99.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology 47*(6): 381 - 391.

Fitts, P. M., & Peterson, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology 67*(2): 103 - 112.

Gajos, K., Reinecke, K., & Herrmann, C. (2012). Accurate measurements of pointing performance from in situ observations. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*: 3157-3166.

Garrison, D. R. & Anderson, T. (2003). *E-learning in the 21st century a framework for research and practice.* Open Universiteit Nederland. Retrieved from: http://portal.ou.nl/documents/89037/89380/Garrison+%26%20Anderson+%282003%29.pdf (16-07-2013)

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE) 1*(1): 20-20.

Gregersen, T., & Horwitz, E. K. (2002). Language Learning and Perfectionism: Anxious and Non-Anxious Language Learners' Reactions to Their Own Oral Performance. *The Modern Language Journal 86*(4): 562-570.

Khan, I. A., Brinkman, W. P., Fine, N., & Hierons, R. M. (2008). Measuring personality from keyboard and mouse use. *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*: 38-45.

Leggett, J., & Williams, G. (1988). Verifying identity via keystroke characteristics. *International Journal of Man-Machine Studies 28*(1): 67-76.

Linnenbrink, E. A., & Pintrich, P. R. (2003). THE ROLE OF SELF-EFFICACY BELIEFS INSTUDENT ENGAGEMENT AND LEARNING INTHECLASSROOM. *Reading &Writing Quarterly 19*(2): 119-137.

McQuiggan, S. W., & Lester, J. C. (2006). Diagnosing self-efficacy in intelligent tutoring systems: An empirical study. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*: 565-574.

Mcquiggan, S. W., Mott, B. W., & Lester, J. C. (2008). Modelling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction 18*(1-2): 81-123.

Mödritscher, F., Spiel, S., & García-Barrios, V. M. (2006). Assessment in E-Learning Environments: A Comparison of three Methods. Conference paper presented at *Society for Information Technology & Teacher Education International Conference 2006*.

Nwana, H. S. (1990). Intelligent tutoring systems: an overview. *Artificial Intelligence Review 4*(4): 251-277.

Pallant, J. (2011). SPSS survival manual: A step by step guide to data analysis using SPSS. Maidenhead: Open University Press.

Pusara, M., & Brodley, C. E. (2004). User re-authentication via mouse movements. *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*: 1-8.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational psychologist 26*(3-4): 207-231.

Simpson, R., Koester, H., & LoPresti, E. (2006). Evaluation of an adaptive row/column scanning system. *Technology and disability 18*(3): 127-138.

Sirohi, V. (2007). E-learning: An Overview. Conference paper presented at *The International Conference on e-Government (iceg'07)*.

Soldato, T. D. (1993). *Motivation in tutoring systems*. Doctoral dissertation. University of Sussex.

Soukoreff, R. W., & MacKenzie, I. S. (2003). Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric. *Proceedings of the SIGCHI conference on Human factors in computing systems*: 113-120.

Spek, E. D. van der (2011). *Experiments in serious game design: a cognitive approach*. Doctoral dissertation. Utrecht University.

Umphress, D., & Williams, G. (1985). Identity verification through keyboard characteristics. *International journal of man-machine studies 23*(3): 263-273.

Vicente, A. de & Pain, H. (1998). Motivation diagnosis in intelligent tutoring systems. *Intelligent Tutoring Systems, Lecture Notes in Computer Science 1452*(1998): 86-95.

Vicente, A. de, & Pain, H. (2002). Informing the detection of the students' motivational state: an empirical study. *Intelligent Tutoring Systems, Lecture Notes in Computer Science 2363*(2002): 933-943

Wang, H., Chignell, M., & Ishizuka, M. (2006, March). Empathic tutoring software agents using real-time eye tracking. In Proceedings of the 2006 symposium on Eye tracking research & applications (pp. 73-78). ACM.

Wichary, M. (2005). Fitts's Law demonstration. http://fww.few.vu.nl/hci/interactive/fitts/. Visited on: 4th of July 2013.

Zhang, D., Zhao, J. L., Zhou, L., & Nunamaker Jr, J. F. (2004). Can e-learning replace classroom learning? *Communications of the ACM 47*(5): 75-79.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology 25*(1): 82-91.

Zimmermann, P., Guttormsen, S., Danuser, B., & Gomez, P. (2003). Affective computing--a rationale for measuring mood with mouse and keyboard. *International journal of occupational safety and ergonomics 9*(4): 539-551.

Appendices

Appendix A: Complete (Dutch) list of questions

id	vraag MOUSE	Antwoord A	Antwoord B	Antwoord C	antwoord D	zekerheids-inschatting
1	Wie zijn billen brandt moet op de ..... zitten	Pleisters	Blaren	kussens	kosten	hoog
2	Op hoeveel km afstand van het aardoppervlak vliegt een TV satelliet	39 km	40 km	41 km	42 km	laag
3	Hoeveel millimeters gaan er in drie centimeter	15	20	25	30	hoog
4	Hoeveel tijdzones zijn er in de wereld	24	20	16	32	laag
5	Hoeveel letters heeft het alfabet	23	24	25	26	hoog
6	Waar groeit 75% van alle ananassen ter wereld	Mexico	Argentinië	Hawaï	China	laag
7	Om de hoeveel meter staan de hectometerpaaltjes langs de Nederlandse snelwegen	100m	150m	200m	500m	hoog
8	Hoe breed is de mond van het Amerikaanse vrijheidsbeeld	80 cm	70 cm	100 cm	90 cm	laag
9	Hoeveel eieren gaan er in een dozijn	12	11	10	13	hoog
10	Hoe noemt men inwoners van Korfoe	Korfeten	Korfioten	Korfoeneten	Korfers	laag
11	Welke Amerikaanse stad heet "De geboorteplaats van de wolkenkrabber"	New York	Chicago	Washington	New Orléans	laag
12	Hoeveel jaren zitten er in een millennium	100	1000	10000	2000	hoog
13	In welk land kun je een kop koffie betalen met Cruzeiro’s	Peru	Argentinië	Brazilië	Bolivia	laag

14	Wat is het hoogste getal van vijf cijfers	99999	00000	88888	55555	hoog
15	Wat is de wetenschappelijke naam voor een ringvormig koraaleiland	Schiereiland	Etol	Rif	Atol	laag
16	Wat voor wegen hebben namen die beginnen met een A, zoals de A1 en A73	Tolwegen	Snelwegen	Zandwegen	Parallelwegen	hoog
17	Wat is de middelste naam van Margaret Thatcher	Hilda	Matilda	Greta	Theodora	laag
18	Hoe noemt men grillen in de open lucht met behulp van houtskool	Buitenkeuken	Skottelbraai	George Foreman Grill	Barbeque	hoog
19	Hoe noemt een kapitein de buitenkant van zijn schip	Boeg	Huid	Romp	Zeil	laag
20	Hoe noemt men mensen die niet kunnen lezen en schrijven	diabeten	digibeten	anachoreten	analfabeten	hoog
	KEYBOARD					
	De weggelaten woorden dienen door de proefpersoon ingevuld te worden			antwoord		
21	Iets goedkoops koop je volgens het spreekwoord voor een _____ en een ei			appel		hoog
22	Als een echtpaar 15 jaar getrouwd zijn, noemt men dit ook wel een _____ huwelijk			kristallen		laag
23	In Groot-Britannie is _____ de voertaal			Engels		hoog
24	De keuken op een schip noemt men een _____			kombuis		laag
36	De drie (meest gangbare) soorten autobrandstof zijn benzine, LPG en _____			diesel		hoog
26	Bij fotosynthese zetten planten _____ en koolstofdioxide om in voedsel.			zonlicht		laag
27	Bij het snijden van _____ gaan je ogen tranen			uien		hoog
28	Een muur zonder deuren of ramen noemt men een _____ muur.			blinde		laag
29	Mensen die ervoor kiezen om geen vlees of vis te eten worden _____ of veganisten genoemd.			vegetariërs		hoog
30	Brons is een legering van de metalen: _____ en tin.			koper		laag

31	Natuurgas is geurloos van zichzelf, daarom wordt er een _____ aan toegevoegd zodat mensen het toch kunnen ruiken.	geur		hoog
32	Als er bij het schaken een "rokade" gespeeld wordt, bewegen er 2 stukken in 1 beurt, namelijk de koning en de _____	toren		laag
33	De Mount _____ is de hoogste berg op aarde met een hoogste piek van 8,848 meter hoog.	Everest		hoog
34	De hoofdstad van Peru is _____.	Lima		laag
35	Het land met de meeste inwoners ter wereld is _____	China		hoog
36	De Haka is de traditionele oorlogszang van de Maori uit _____	Nieuw-Zeeland		laag
37	De voornaam van oud-president Clinton is _____	Bill		hoog
38	De Japanse drank Sake wordt gemaakt van water en _____	rijst		laag
39	De drank _____ wordt gemaakt van hop, water, maïs en gemoute gerst	bier		hoog
40	Op de nationale vlag van Niger staan de kleuren oranje, _____ en wit.	groen		laag

Appendix B: Complete variable list

Variables from the first part of the experiment regarding mouse movement:

Date and time	The date and time stamp of the moment when the data from a question in the experiment was stored
Gender	The gender of the participant
Age	The age of the participant in years
Education	The educational level of the participant (with VMBO, MBO, HBO and WO as choices, indicating the mayor Dutch educational levels)
Pre-test self-confidence	The level of self-confidence indicated by the participant regarding their general answering capabilities, indicating how sure a participant felt about answering questions in general
Question ID	The identification number of the question of which the results have been stored. These are the same as the ID's in the question list of Appendix A
Shortest distance	The Euclidian and thus the shortest distance between the starting point of the mouse cursor and it's ending point.
Actual distance	The actual distance the mouse cursor travelled to move from the starting point to the ending point
Distance difference	The difference in distance between the shortest distance and the total distance, calculated by subtracting the shortest distance from the total distance.
Fitts' time	The time predicted by Fitts' law, predicting the amount of time this participant on this particular device would need to move the mouse cursor over a given distance from a starting point to an ending point.
Actual time	The time the participant needed to move the mouse from the starting point to the ending point.
Time difference	The difference in time between the Fitts' predicted time and the actual time, calculated by subtracting the Fitts' time from the actual time.
Number of pauses	The number of pauses during the movements of the mouse cursor, where a pause is the lack of movement for longer than 200 milliseconds
Pause time	The total pause time for a particular question, comprised of the sum of the

	time of each pause.
Pauses on buttons	The number of pauses that occurred on an answering button
Efficacy change	The number of times the participant changed their indication of self-efficacy for this question
Efficacy time	The time the participant needed to indicate their self-efficacy level.
Question time	The time period from the moment the question was shown up to the moment the participant clicked on an answer
Answering performance	Indicates if the participant clicked the correct answer
Answer	The answer the participant clicked, regardless of its correctness
Reading time	The time the participant waited before clicking the "show answers" button
Question difficulty	The (expected) difficulty of the question, as indicated by the list of questions in Appendix A.

Variables from the second part of the experiment regarding keyboard input:

Date and time	The date and time stamp of the moment when the data from a question in the experiment was stored
Gender	The gender of the participant
Age	The age of the participant in years
Education	The educational level of the participant (with VMBO, MBO, HBO and WO as choices, indicating the mayor Dutch educational levels)
Pre-test self-confidence	The level of self-confidence indicated by the participant regarding their general answering capabilities, indicating how sure a participant felt about answering questions in general
Question ID	The identification number of the question of which the results have been stored. These are the same as the ID's in the question list of Appendix ….
Average keystroke delay	The averaged time between the individual keys the participants used to type their answer for this particular question
Relative average keystroke delay	The average time between two key presses, relative to the average keystroke delay from the calibration question, calculated by subtracting the

	calibration average time from the average keystroke delay from the current question
Type time	The time the participant needed to type their answer
Levenshtein distance	The number of characters in the answer of the participant which need to be changed to result in the correct string of characters of the correct answer, indicating the error rate of the given answer
Read time	The time the participant waited before clicking the “show answers” button
Question time	The time period from the moment the question was shown up to the moment the participant clicked on an answer
Number of keys	The number of keys the participant pressed. This can deviate from the number of characters in the answer, since the participant could have made corrections.
Answer string	The answer the participants typed
Number of corrections	The number of corrections made by the participant (using the backspace key)
Question difficulty	The (expected) difficulty of the question, as indicated on the list of questions in Appendix A.

Appendix C: Detailed Correlation results

Mouse movement

		efficacy	Distance difference	Number of pauses	Pauses on buttons	Time difference	Pause time	Read time	Question time
efficacy	Pearson Correlation	1	-,148**	-,286**	-,032	-,369**	-,332**	-,020	-,267**
	Sig. (2-tailed)		,000	,000	,279	,000	,000	,503	,000
	N	1122	1122	1122	1122	1122	1122	1122	1122
Distance difference	Pearson Correlation	-,148**	1	,209**	,061*	,208**	,007	-,025	,123**
	Sig. (2-tailed)	,000		,000	,042	,000	,809	,410	,000
	N	1122	1122	1122	1122	1122	1122	1122	1122
Number of pauses	Pearson Correlation	-,286**	,209**	1	,282**	,611**	,480**	,082**	,486**
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,006	,000
	N	1122	1122	1122	1122	1122	1122	1122	1122
Pauses on buttons	Pearson Correlation	-,032	,061*	,282**	1	,205**	,176**	,041	,172**
	Sig. (2-tailed)	,279	,042	,000		,000	,000	,173	,000
	N	1122	1122	1122	1122	1122	1122	1122	1122
Time difference	Pearson Correlation	-,369**	,208**	,611**	,205**	1	,951**	,084**	,751**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,005	,000
	N	1122	1122	1122	1122	1122	1122	1122	1122
Pause time	Pearson Correlation	-,332**	,007	,480**	,176**	,951**	1	,117**	,746**
	Sig. (2-tailed)	,000	,809	,000	,000	,000		,000	,000
	N	1122	1122	1122	1122	1122	1122	1122	1122
Read time	Pearson Correlation	-,020	-,025	,082**	,041	,084**	,117**	1	,717**
	Sig. (2-tailed)	,503	,410	,006	,173	,005	,000		,000
	N	1122	1122	1122	1122	1122	1122	1122	1122
Question time	Pearson Correlation	-,267**	,123**	,486**	,172**	,751**	,746**	,717**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	
	N	1122	1122	1122	1122	1122	1122	1122	1122

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

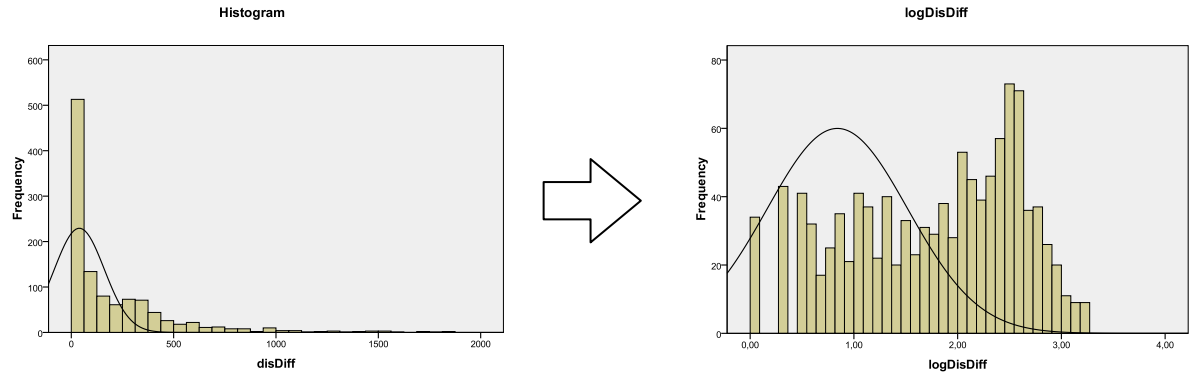
Keyboard input

		efficacy	Question time	Relative average keystroke delay	Type time	Number of corrections	Read time
efficacy	Pearson Correlation	1	,018	-,011	-,026	-,019	,033
	Sig. (2-tailed)		,557	,725	,389	,537	,284
	N	1066	1066	1066	1066	1066	1066
Question time	Pearson Correlation	,018	1	-,034	,087**	-,007	,390**
	Sig. (2-tailed)	,557		,265	,005	,811	,000
	N	1066	1068	1068	1068	1068	1068
Relative average keystroke delay	Pearson Correlation	-,011	-,034	1	,667**	,170**	-,633**
	Sig. (2-tailed)	,725	,265		,000	,000	,000
	N	1066	1068	1068	1068	1068	1068
Type time	Pearson Correlation	-,026	,087**	,667**	1	,234**	-,292**
	Sig. (2-tailed)	,389	,005	,000		,000	,000
	N	1066	1068	1068	1121	1121	1121
Number of corrections	Pearson Correlation	-,019	-,007	,170**	,234**	1	-,223**
	Sig. (2-tailed)	,537	,811	,000	,000		,000
	N	1066	1068	1068	1121	1121	1121
Read time	Pearson Correlation	,033	,390**	-,633**	-,292**	-,223**	1
	Sig. (2-tailed)	,284	,000	,000	,000	,000	
	N	1066	1068	1068	1121	1121	1121

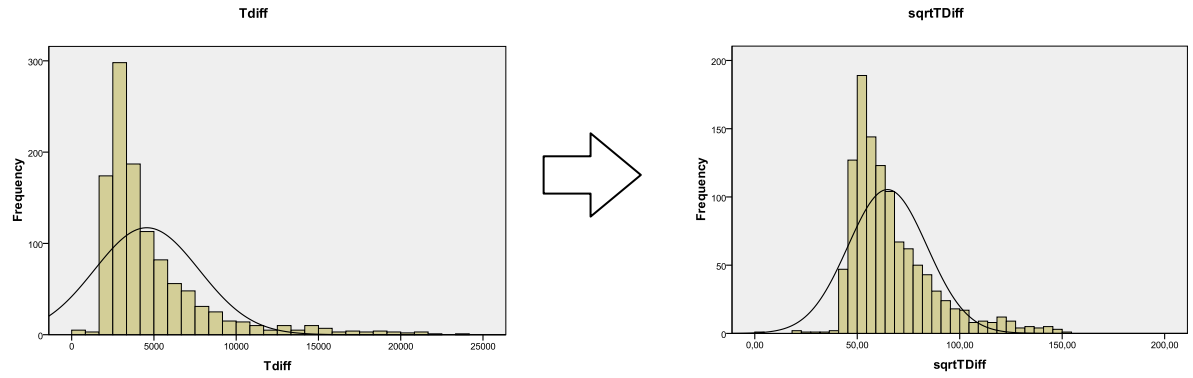
**. Correlation is significant at the 0.01 level (2-tailed).

Appendix D: Variable transformations to improve normal distribution of mouse movement variables

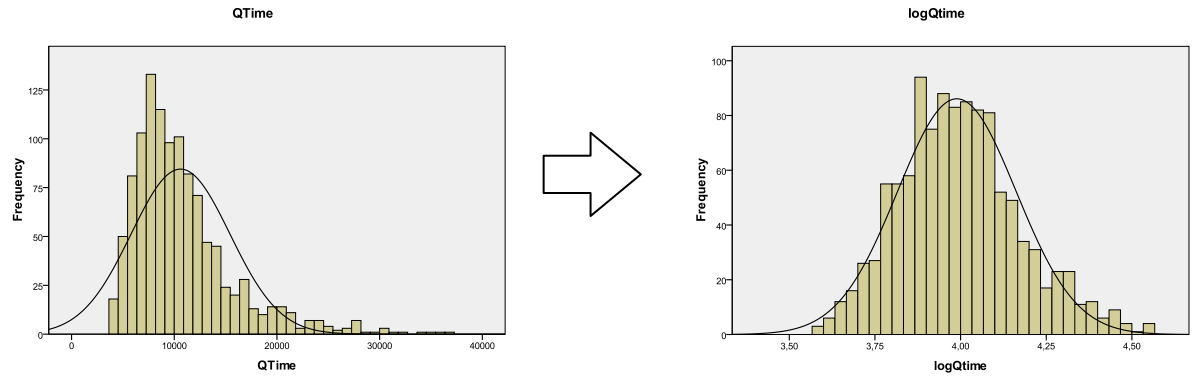
Distance difference	logarithm (Distance difference +1) (+1 to compensate for zero values)



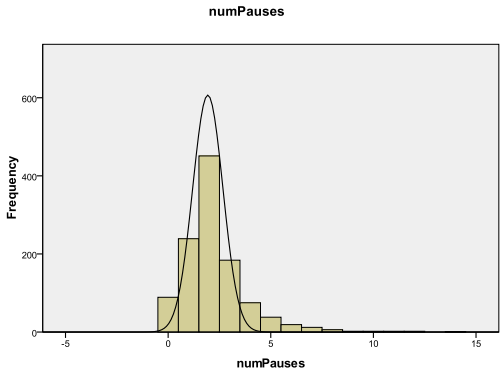
Time difference	square root (Time difference +1891) (+1891 to compensate for negative values)

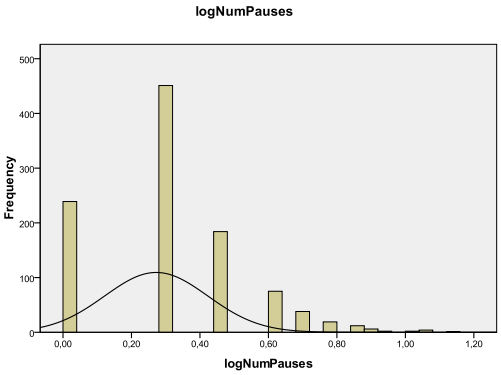


Question time	logarithm Question time

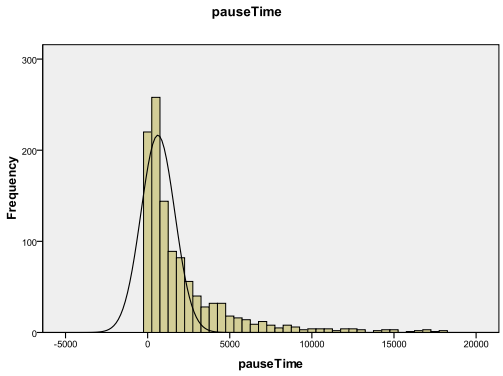


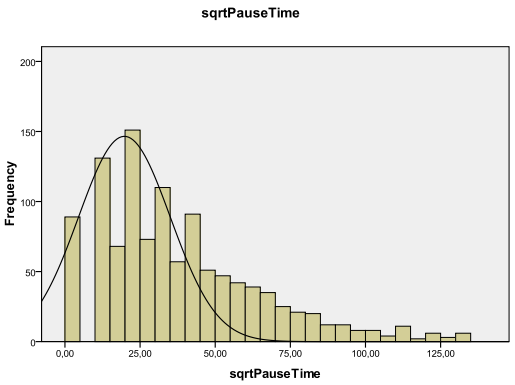
Number of pauses	log Number of pauses



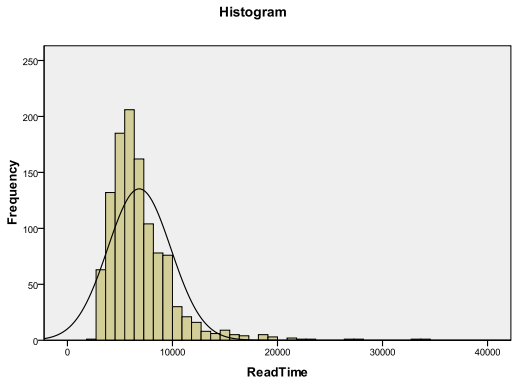


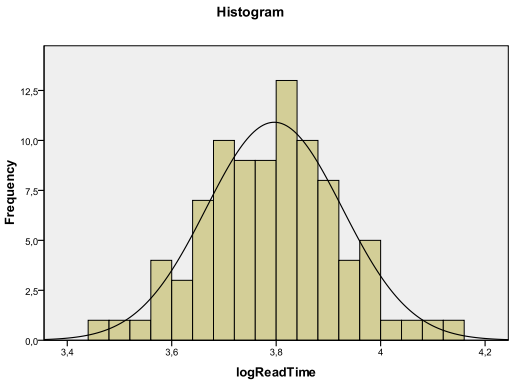
Pause time	square root (Pause time +1) (+1 to compensate for zero values)





Read time	logarithm Read time





Appendix E: Detailed Linear multiple regression results

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,418[a]	,175	,172	36,699	,175	54,455	4	1028	,000

a. Predictors: (Constant), log Distance difference, log Pauses on buttons, sqrt Time difference, log Number of pauses

b. Dependent Variable: efficacy

Correlations

		efficacy	log Pauses on buttons	log Number of pauses	sqrt Time difference	log Distance difference
Pearson Correlation	efficacy	1,000	-,036	-,247	-,413	-,046
	log Pauses on buttons	-,036	1,000	,259	,215	,010
	log Number of pauses	-,247	,259	1,000	,584	,255
	sqrt Time difference	-,413	,215	,584	1,000	,184
	log Distance difference	-,046	,010	,255	,184	1,000
Sig. (1-tailed)	efficacy	.	,111	,000	,000	,063
	log Pauses on buttons	,111	.	,000	,000	,372
	log Number of pauses	,000	,000	.	,000	,000
	sqrt Time difference	,000	,000	,000	.	,000
	log Distance difference	,063	,372	,000	,000	.
N	efficacy	1122	1121	1033	1122	1122
	log Pauses on buttons	1121	1121	1033	1121	1121
	log Number of pauses	1033	1033	1033	1033	1033
	sqrt Time difference	1122	1121	1033	1122	1122
	log Distance difference	1122	1121	1033	1122	1122

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	113,460	4,392		25,831	,000					
	log Pauses on buttons	21,621	10,566	,060	2,046	,041	-,036	,064	,058	,923	1,084
	log Number of pauses	-5,224	6,333	-,030	-,825	,410	-,247	-,026	-,023	,616	1,624
	sqrt Time difference	-,820	,069	-,415	-11,839	,000	-,413	-,346	-,335	,652	1,533
	log Distance difference	1,822	1,423	,038	1,280	,201	-,046	,040	,036	,929	1,076

a. Dependent Variable: efficacy