ARTICLE TEMPLATE

Seeing (Movement) is Believing: the Effect of Motion on Perception of Automatic Systems Performance

A. N. Author[a] and John Smith[b]

[a]Taylor & Francis, 4 Park Square, Milton Park, Abingdon, UK; [b]Institut für Informatik, Albert-Ludwigs-Universität, Freiburg, Germany

ARTICLE HISTORY
Compiled March 11, 2018

ABSTRACT
In this paper, we report on one lab study and seven follow-up studies on a crowdsourcing platform designed to investigate the potential of animation cues to influence users' perception of two smart systems: a handwriting recognition and a part-of-speech tagging system. Results from the first three studies indicate that animation cues can influence a participant's perception of both systems' performance. The subsequent three studies, designed to try and identify an explanation for this effect, suggest that this effect is related to the participants' mental model of the smart system. The last two studies were designed to characterise the effect more in detail, and they revealed that different amounts of animation do not seem to create substantial differences and that the effect persists even when the system's performance decreases, but only when the difference in performance level between the systems being compared is small.

KEYWORDS
Smart Systems; Performance Perception; User Study; Animation Cues; Visual Feedback; User Experience

Contents

1 Introduction **3**

2 Related Work **4**
2.1 Cognitive Biases 4
2.2 Transparency and Intelligibility of Software Systems 5
2.3 The role of motion in users' perception of systems 6

3 Methodology **8**
3.1 Studies' Design 8
3.2 Participants 9
3.3 Equipment 10
3.4 Procedure 10

4 Studies' Results **10**
4.1 Performance ratings 11

CONTACT A. N. Author. Email: latex.helpdesk@tandf.co.uk

4.2 Reasons for choosing one system over the other 11

5 Study 1 - Animation-performance effect in the lab 12
5.1 Study Conditions 12
5.2 Results 12
5.3 Discussion 12

6 Study 2 - Animation-performance effect on MTurk 13
6.1 Results 13
6.2 Discussion 13

7 Study 3 - Animation-performance effect: part-of-speech tagging 14
7.1 Study Conditions 15
7.2 Results 15
7.3 Discussion 15

8 Study 4 - Non-human-like animation 16
8.1 Study Conditions 16
8.2 Results 16
8.3 Discussion 16

9 Study 5 - Animation-performance effect and mental model salience 17
9.1 Results 17
9.1.1 How people think a handwriting recognition system works . . . 17
9.1.2 The system worked as participants expected 18
9.2 Discussion 18

10 Study 6 - Animation-performance effect with alternative mental models 19
10.1 Study Design 19
10.2 Results 20
10.2.1 System working according to expectations 20
10.3 Discussion 21

11 Study 7 - Amount of animation detail 22
11.1 Study Conditions 23
11.2 Results 23
11.2.1 Performance ratings 23
11.2.2 Selection of the system with the best performance 23
11.3 Discussion 23

12 Study 8 - Animation-performance effect vs real systems performance 24
12.1 Study Design 24
12.2 Results 24
12.3 Discussion 24

13 Summary and General Discussion 25

14 Implications 26

15 Further Research Opportunities 27

16 Conclusion **27**

17 Acknowledgements **28**

A Responses independent of the reward **32**

1. Introduction

There is a growing number of *smart systems* that help users to gather data and process information from sensors. These are systems that utilize some form of pattern recognition, machine learning, or more generally artificial intelligence to complete a variety of information-processing tasks. Until recent times, such smart systems were only accessible at high-cost for specialised applications (e.g. in medical fields, aviation), but more recently they have become increasingly widespread for non-specialist applications, such as apps that help people with office work (e.g. translation platforms[1] or document scanning[2]). As smart systems become available to a wider variety of users, it is important to study how non-experts interact with them. 'Smart' is a loose term, but we broadly consider systems to be smart, if they are capable to autonomously make decisions based on data that they collect or sense (such as search engines and recommender systems but also smart thermostats).

Given that smart systems can involve advanced concepts in pattern recognition (e.g. Bayesian classification Talbot, Lee, Kapoor, & Tan (2009)), or even act as black boxes (Krause, Perer, & Ng, 2016), their operation may be difficult to grasp for non-specialist users, who do not receive training (as it is common for domestic appliances).

Research in psychology and behavioural economics indicates that people's perception and decisions can be influenced by cognitive biases, implemented often through nuanced cues (Ariely, 2008; Tversky & Kahneman, 1985). As such, we are interested in whether cues, and particularly *visual animation cues*, can influence users' perception of smart systems. In particular, we focus on whether these cues affect how people rate the performance of a smart system in terms of how well it does its job. Indeed, recent research (Garcia, Costanza, Ramchurn, & Verame, 2016) demonstrated that simple *motion cues* can have quite a radical impact on people's perception of vacuum cleaning robots: when the interaction was orchestrated in such a way that participants saw the robot moving, they perceived it to clean better than a robot which worked identically, but was not seen moving. Building on this prior work, our aim is to investigate whether a similar effect can be noticed for GUI-based smart systems, through the use of *animation cues* integrated in the system interface. We argue that being aware of and understanding such biases are important for the design of interaction around smart systems. On one hand, it may reveal opportunities to improve users' perception e.g. making the system more popular or more likeable. On the other hand, and perhaps even more importantly, being aware of such biases may allow designers to avoid unintentionally deceiving users.

In this paper, we report on one lab study and seven follow-up studies on the crowdsourcing platform Amazon Mechanical Turk[3] (MTurk) designed to investigate the potential of animation cues to influence users' perception of the performance of two smart systems: a handwriting recognition and a part-of-speech tagging system. The

[1] `https://www.apertium.org`
[2] `https://www.camscanner.com/`
[3] `https://www.mturk.com`

aim of the first three studies was to observe the phenomenon, initially for a handwriting recognition system in a lab setting (Study 1), then for the same system on MTurk (Study 2), and for a different system, a part-of-speech tagging system on MTurk (Study 3). In all studies, participants were asked to compare the performance of two versions of the system, with one system containing an animation in its UI while the other one did not. The animation consisted of simply highlighting input words in red, while on the output the translated words appeared. Results indicate that indeed animation cues can influence participants' perception of the system performance. Both in the lab and on MTurk, participants reported that the system, which had animation integrated in its UI, performed better than an alternative system, which was in fact identical apart from the animation. We then report three further studies designed to try and explain this effect. Their results suggest that the effect of animation cues is related to a participant's mental model of the smart system. More precisely, if the animations are compatible with a user's expectations of how the system works, they seem to somehow provide reassurance about the system's operation, and evoke an illusion that the system works better than an alternative, which doesn't include an animation. Having identified a possible explanation for this phenomenon, we report on two further studies designed to characterize this effect in more detail. These last two studies revealed that different amounts of animation do not seem to create substantial differences, and that the effect persists even when the system's performance (with the animation) actually decreases, but only when the difference in performance between the systems being compared is small. Designers should pay attention to the fact that they could unintentionally introduce risks, as animations may lead users to rely on the results of a smart system, even when they should not, due to bad performance. All studies collected both quantitative and qualitative data, and used *consensus-oriented financial incentives* to increase ecological validity and motivate participants to provide thoughtful responses. We further summarise the relationship among the eight studies and our findings in the video accompanying this paper: `https://vimeo.com/245121953`

2. Related Work

Our research aims to analyse how simple animations can change participants' perception of smart systems' performance. To this end, in what follows, we survey prior research that has studied cognitive biases and how different framings of information impact people's perception. Then, we discuss transparency for the intelligibility of software systems Bellotti & Edwards (2001) and how their design influences how people perceive the system. Finally, we discuss prior work on the perception of motion in technology.

2.1. *Cognitive Biases*

Studies in psychology and behavioural economics have shown that people's perception of how well a system or process works can be influenced by different cognitive biases. As early as 1932 studies revealed that participants' reported perception of the *quality* of identical products (in particular nylon stockings) can be influenced by smell (Laird, 1932), or by the order in which the products are examined (de Camp Wilson & Nisbett, 1978). In both these studies participants were unanimously, or almost unanimously unaware of such bias, and instead they referred to a variety of other factors to justify their product selection. However, these factors could not have been the real reasons

for the choice as the items were identical.

In a more recent example, Tversky & Kahneman (1985) showed that people can also be influenced by the way outcomes are described to them. In a survey, participants were presented with a problem and two possible solutions. These two solutions had the same outcome, however, one emphasised its positive aspects, while the other emphasised the negative aspects. Results suggest that people had a tendency to choose the solution that emphasised the positive aspects. As another example, Ariely (2008) ran a study to analyse if the price on medicine has a placebo effect on people's perception of how they feel after they took medication. One group received the medicine with the actual price and a second group received the medicine with a 10 cents discount (off an original price of $2.50). The results showed that while almost all participants in the first group experienced pain relief from the pill, *only half* of the participants who were given the "discounted medicine" experienced pain relief. In our work, we are interested in exploring whether there are also cognitive biases that can influence people's perception of how well smart systems work.

2.2. *Transparency and Intelligibility of Software Systems*

Prior research has examined whether increasing the intelligibility of smart systems has an effect on how people understand them. In particular, previous studies have suggested that smart systems should generate and provide meaningful explanations for their actions, behaviour or outcomes (Lim, Dey, & Avrahami, 2009; Lyons, 2013; Tullio, Dey, Chalecki, & Fogarty, 2007). For example, Lim et al. (2009) ran two experiments to analyse the effect of meaningful explanations describing *why* and *why not* a context-aware application behaved in a certain way. Their findings suggest that users have a better understanding of a system's behaviour and a higher feeling of trust when it provides explanations. Moreover, Tullio et al. (2007) ran a six-week field study to analyse whether intelligibility can help office workers improve their understanding of how a system predicts their managers' interruptibility. They found that people were able to understand the system prediction better, even if the overarching structure of their mental model stayed stable during the study. Similarly, another study investigated *Laksa*, a context-aware software which used eight question type explanations (e.g. *Why, Why Not, What If*) to explain its decision to the users Lim & Dey (2011a). To evaluate the software, participants used the software in three situational dimensions (exploration, fault finding, and social awareness) that allowed the researchers to observe whether participants do or do not understand software decisions. They noted that quickly consumable explanations of a system's output are crucial and additional, richer explanations should be easily accessible. Lim & Dey (2011a) observed that prior knowledge plays an important role in both understanding of such systems and also interpreting the explanations given. The lack of previous knowledge can lead people to misunderstand or misuse a system. In addition, Lim & Dey (2011b) they ran a study with two context-aware applications (location-aware, and sound-aware). They analysed the interaction between intelligibility and application uncertainty. They found intelligibility is helpful for applications with high certainty, but it is harmful when the certainty of the application is lower, even in situations where the application behave appropriately. While such prior work shows that explanations can be useful to help people make sense of smart systems, they can also cause information overload, possibly confusing and overwhelming users (Lim & Dey, 2011a; Yang & Newman, 2013). Because of this, our aim is to understand whether it is possible to change people's

perception of smart systems without increasing their cognitive workload by, for example, providing additional cues (e.g. through animation) that can expose to users that a smart system is doing work.

Another way of improving system intelligibility is through information visualisation, which is the use of visual representations of data structures and algorithms to help people analyse data (Card, Mackinlay, & Shneiderman, 1999; Ware, 2012). The concept of information visualisation is considered a method to make a system understandable without providing explanations of its process. For example, O'Donovan et al. (2008) ran a study where participants interacted with *PeerChooser*, an interactive visualisation system for collaborative filtering. The system generated a peer-graph which is centred on the current user. The graph showed a visual representation of their peer group or neighbourhood allowing participants to manipulate connections with their neighbours. This interaction allowed participants to visualise recommendations from the system based on their preferences. Their findings suggest that a visual-interactive approach can improve the accuracy of the recommendations provided by the system and also enhance user experience (O'Donovan, Smyth, Gretarsson, Bostandjiev, & Höllerer, 2008). In our case, instead of using interactive visualisations, we explore visualisations of a system's process through motion (animations) that represents its execution of a task.

An example of a study that uses motion as a visual feedback to explain a system's decision is presented by Vermeulen (2010). In their study, animations were used to show the process that a system follows when it makes a decision, given how a user interacts with its inputs (e.g., switch) or sensors (e.g. motion detector). Findings from their study suggest that participants understood the decisions and actions taken by the system because of the explanations they received. This approach demonstrates that animation, as a feedback, can help people understand decisions made by a system. However, participants also found it difficult to track the animation at times, thereby confusing them. Building on this prior work, we are keen to further explore how people's perception of smart systems changes depending on the animation.

2.3. *The role of motion in users' perception of systems*

Research has looked at how people perceive motion of screen-based systems. Decades ago, Chang & Ungar (1993) already suggested that bringing simple cartoon animation techniques to interface elements such as icons and windows could make interfaces easier to understand and more pleasant to use. Indeed, animating progress-bars in various ways can have an effect on the perceived durations, i.e. by using a certain design it was possible to make processes appear faster to the user (Harrison, Yeo, & Hudson, 2010). Dragicevic, Bezerianos, Javed, Elmqvist, & Fekete (2011) explored various animation strategies in visualisations and focused on transitions in point clouds. Their findings show that smoothly stretching time at the endpoints of an animation is the most accurate pacing strategy for animated transitions with the possible explanation that it helps users in predicting and therefore following the motion. *Animacy*, as Tremoulet & Feldman (2000) state, is when people perceive an object as being alive, through the pattern of its movements. They found that the movement of an object does not need to be dramatic to show animacy (Fritz Heider, 1944; Reeves & Nass, 1996). As a consequence, people attribute motivations, or intention in objects' movements from the patterns that these objects follow. This means that people can infer objects' intentions through their movements (Gao & Scholl, 2011; Ju & Takayama, 2009; Michotte, 1963;

Pantelis & Feldman, 2012; Schlottmann & Surian, 1999). This has also been observed during people's interaction with physically actuated interfaces such as helium balloons (Nowacka, Hammerla, Elsden, Plötz, & Kirk, 2015) or vacuum cleaning robots (Garcia et al., 2016) and even automatic doors (Ju & Takayama, 2009). Research on animating robot behaviour in the form of robots physically displaying anticipation and reaction fostered human readability and helped people to predict robot behaviour (Takayama, Dooley, & Ju, 2011). Therefore, through designing the movement, it is possible to affect how people perceive objects. Michotte (1963) showed in their study that if two objects are in the same frame and suddenly change their direction, people can infer that both objects have a causal interaction[4]. Pantelis & Feldman (2012) ran a study with multiple objects moving around on a screen. They found that, after watching multiple objects moving on a screen, people make interpretations of the intention and behaviour of the objects. Moreover, in their experiment, people were able to distinguish if an object behaved friendly or hostile when it was moving around other objects. This body of work makes us believe that - by showing people an animation - they can be convinced that a system is working on a task. As such, we presume that people perceive a system that somehow communicates that it is doing work perform better than a system that hides how it works.

However, it has also been shown that some features of animations can confuse people and negatively impact people's perception. These features include but are not limited to: interaction between multiple objects (Gao, McCarthy, & Scholl, 2010), trajectories that are too complicated (Dittrich & Lea, 1994; Tremoulet & Feldman, 2000), unnatural movements (Popović, Seitz, & Erdmann, 2003), static backgrounds that are too complex (Gelman, Durgin, & Kaufman, 1995), or speed of how fast a feedback is displayed (Vermeulen, Luyten, Coninx, & Marquardt, 2014). Hence, it is important to ensure such issues are avoided when providing feedback about a system's execution of tasks. Additionally, Padrao, Gonzalez-Franco, Sanchez-Vives, Slater, & Rodriguez-Fornells (2016) found that the sense of agency can be perturbed when people perceive an erroneous movement of an embodied avatar.

Prior research has also analysed affective qualities of an interface depending on how the information and motion are presented on a screen (Chang & Ungar, 1993; Detenber & Reeves, 1996; Dragicevic et al., 2011; Harrison, Amento, Kuznetsov, & Bell, 2007; Park & Lee, 2010b). Park & Lee (2010a) ran a study to understand how motion (i.e. transition effects between objects) influences the affective quality of an interface to improve user experience. They presented an image viewing interface that allows users to browse through a set of photos as they shift horizontally from one to another. Their results show that motion influenced how people rated affective qualities of the interface (e.g. youthfulness, calmness, and uniqueness). Also related to the effect of animation on user emotion, Bakhshi et al. (2016) reported that social network users have a tendency to share content more frequently if it involves animations, compared to content that is purely static. In contrast to this prior work, our interest lies in observing if motion has an effect on how people perceive systems' performance rather than on people's emotions.

[4]Casual interaction is when users are not able to, or do not want to, fully engage with their devices.

3. Methodology

In this section, we summarise our methodology for the eight studies we designed and ran to understand how animation cues affect how people perceive the performance of smart systems. Even though Study 1 was conducted in the lab and the following ones were conducted on MTurk, the general structure of the experiments is the same.

The studies involved a relatively simple graphic animation, related to the system operation. Two types of systems were used in different studies: a handwriting recognition system, and a part-of-speech tagging system[5]. We chose a handwriting recognition system, a system that recognises handwritten text and converts it to electronic text (or e-text / typed text), because this is a common task that many people can relate to, at least conceptually, and it also can be simulated easily (Verame, Costanza, & Ramchurn, 2016). Moreover, we chose to use text in Filipino, a language that most users would be unlikely to know, to mimic the likely circumstances of users not being familiar with the kind of data handled by the system. In this way, rather than simply checking the system output for typos, users are required to compare the input and the output looking for differences, a task that is more attention demanding. As the part-of-speech tagging system was used only in Study 3, we refer to the section below about its specific details and the rationale for choosing it.

3.1. *Studies' Design*

For each study, a fully counterbalanced, within-participants design was used: participants were asked to evaluate and compare the performance of two or three versions of the same system. The systems being compared were always based on a similar graphical user interface. For the handwriting recognition system, illustrated in Figure A1, on the left-hand side of the screen, a scan of a page of handwritten text in Filipino (system's input) is displayed, while on the right-hand side the typed text (system's output) is shown. For the part-of-speech tagging systems, shown in Figure A2, in the centre of the screen, a piece of text is displayed in English, and the tags (e.g. 'verb') are displayed under each word. In both cases, the interface screen was preceded by a 'loading' screen, and showing a message saying that the system was processing the data for 10 seconds, to reinforce the idea that the systems were doing something in the background. We ran the studies in two conditions, no-animation and animation. In the *no-animation* condition, the systems presented the result immediately after the loading screen, and no animation was displayed. In the *animation* condition, after the loading screen, an animation was shown. Various animations were used in this condition, which we detail in each study section below. In general, the animations were designed to give an impression that the system was finishing its process. Similar to prior work Garcia et al. (2016), the duration of the animation was limited, so that it would not require users to look at the system for more than few seconds, which may otherwise become impractical for real applications. Hence, the animation was always on the last two sentences of the text.

External validity was a key factor for all the studies. Therefore, each handwriting recognition system showed the same handwritten text, and each system involved the same number of errors (four mistakes per paragraph, resulting in a total of eight mistakes across two paragraphs). The last two sentences (which included the animation) of each system contained one error. In particular, the mistakes included the usage of

[5]that is the identification of the syntactic role that each word has, e.g. *verb*, emphnoun, *adjective*, etc..

'a' where 'o' would have been correct, or 'o' for 'u', such as 'nagtuturo' for 'nogtuturo'. Equally, in the part-of-speech tagging, both systems showed the same text, and each system involved the same number of errors (eight mistakes across the whole text).

To ensure that participants would provide a meaningful and thoughtful evaluation when they choose which system they considered to have the best performance, we incorporated a *consensus-oriented reward mechanism*. In the lab study, participants were told that if they select the system which the majority identified as the one with the best performance, they will be rewarded with a £10 voucher at the end of the experiment. Because of the constraints of the MTurk platform, the reward mechanism was adjusted accordingly for the crowdsourcing studies. MTurk participants received a *fixed reward*, to compensate them for the time they spent working on our study, as well as an additional *performance-based bonus* if they selected the system which the majority identified as the one with the best performance. This performance-based bonus amounted to the same value as the fixed reward. In other words, the performance-based, consensus-oriented reward doubled the money that MTurk participants received for the study. It was awarded once all participants had completed the study.

Across all studies, participants were firstly asked to rate the individual performance of each system on a 5-point Likert scale. As mentioned above, they were then asked to select which of the systems they believe the majority of participants would choose to have the best performance and to provide a justification for their selection. The questions were deliberately framed rather generically in terms of performance ("Which algorithm do you think has the best performance?"), to leave it to each participant to decide the interpretation. We opted for such an open approach because we are interested in the general perception of this kind of systems, rather than any specific aspect (e.g. accuracy, or speed).

After the post-task questionnaire, participants were also asked (on a separate web page) which system they considered to have the best performance without taking into account what the majority of the participants would choose, nor the reward.

To make sure that participants received a fair payment, we considered the minimum wage across the different countries participants could be from (see restrictions below), and we selected the Canadian one as the one currently highest, at approximately $10 per hour. Therefore, the fixed reward was set according to the time we designed each study to last. In Table A1, we list the fixed amounts the participants received in each study according to the time we calculated they would expend to complete the full task.

3.2. *Participants*

For the lab study, participants were recruited through adverts posted on university social network groups. For the crowdsourcing studies, participants were recruited through MTurk, with three restrictions. First, they were only allowed to take part in the study if their location was *United States, Canada, or Australia*, to avoid issues related to English comprehension. Second, recruitment was limited to participants with a task approval rate equal to 100% (this is the approval from those who advertise the tasks[6]), as rejection on MTurk often indicates that workers do not take tasks seriously. Finally, we did not accept participants who took part in previous studies by tagging them once they completed one of our studies. In Table A2, we list the number of participants we recruited for each study and their demographic information.

[6] `https://www.turkprime.com/Home/FrequentlyAskedQuestions`

Only 5 participants reported knowing Filipino, out of the 192 who took part in the 7 studies which involved that language[7]. The data collected from these 5 participants did not appear to be different from the rest of the sample, so in what follows we consider this data together with the rest.

The number of participants recruited on MTurk was kept relatively low: 8 participants per condition. This choice was made to keep the number of participants consistent between the lab study and the following MTurk studies, and hence make the results more easily comparable.

3.3. *Equipment*

The lab study was run in a room at a university, with the participant sitting next to the investigator. The interfaces and the questionnaire were implemented as a simple Web application, using HTML5 and Python with the Django framework. The application was displayed on a 13" laptop, and served from the same computer. The animation we used was a GIF image, integrated in the Web application. For the crowdsourcing studies, the Web application was extended with an initial questionnaire to obtain the participants' demographic information, and it was served from a standard Web server. No restrictions were placed on the display size or resolution of MTurk participants.

3.4. *Procedure*

At the beginning of the lab study, participants received written instructions asking them to evaluate and compare the performance of the two systems. In the crowdsourcing studies, before participants accepted the task, they were told that the aim of the study was to compare different handwriting recognition or part-of-speech tagging systems, one at a time[8]. They were instructed to check the system's outcome and find possible mistakes that the system could have made. After the introduction, the participants can decide to either accept or reject the task. Once they decided to accept the task, an external link was displayed. The link opened a new window that showed a brief questionnaire, asking for the participants' demographic information.

Once the participants in the crowdsourcing studies completed the initial questionnaire, and in the lab study the participants received the approval, they were ready to start the task. The systems were presented one at a time, in sequence: half of the participants first experienced the *no-animation* condition, while the others experienced one of the *animation* conditions first, which shows the animations we designed for each study. In each condition the system was shown to participants for two minutes, so they had a limited time to compare input and output. After the participants had seen each system, they were asked to fill in a questionnaire to evaluate their performance, as described above.

4. Studies' Results

For conciseness, in this section we report the analysis and the results that were common across different studies. The rest of the results are reported and discussed in separate sections, one for each study. Across all studies, at most only 1 participant per study

[7] 1 participant in Study 6, and 2 in each of Study 7 and Study 8.

[8] this information was displayed in the tasks' description

provided a different answer when ignoring the financial reward, so this data is left out of the following analysis, and only reported in Appendix A. In other words, apart from the Appendix, in this paper we only refer to data based on the consensus-oriented reward mechanism described above.

After the post-task questionnaire, participants were also asked (on a separate web page) which system they considered to have the best performance without taking into account what the majority of the participants would choose, nor the reward.

4.1. *Performance ratings*

A Wilcoxon Signed-rank Test was used to analyse the data we gathered from participants' ratings of the systems' performance. Table A3 presents the results of all studies except Study 7 (as this study required a different statistical analysis, presented later). The same table indicates the median ratings for each group, and the participants' selections of the system they considered as performing best.

4.2. *Reasons for choosing one system over the other*

Participants' answers to the open questions about why they chose one system over the other were summarised and categorised through thematic analysis Braun & Clarke (2006) for all studies. The thematic analysis was performed by two persons for all the studies and each thematic analysis was done from scratch every time. Each response was associated to one or two of the following eight themes: *number of errors, type of errors, generic, animation, speed, others' opinion, random* and *order*. The theme *number of errors* was associated to responses where the participants reported finding fewer errors or mistakes in the output of one system than in the output of the other, such as "There were less mistakes in total", "It has mistaken less characters." and "I think both of them had about the same number of errors, however the second [*animation* condition] one's were more obvious [..]" Comments categorised as *type of error* were linked to situations when participants pointed out typographical errors they found, such as "only confuses a-o, b-h, ri-n whereas the second [*animation* condition] also confuses d-g" and "[..] algorithm only got mistakes when the words contain 'a' and 'o'." Comments such as "More sensitive recognition of lettering [...]", and "[...] Errors of the second program [*animation* condition] are easier to guess and find out." were categorised as *generic*. The category *animation* was used when comments were explicitly related to the animation, e.g.: "Actually seeing the words transcribed probably leaves a good impression." When participants talked about the performance of other participants during the task, such as: "The workers are warmed up and ready to do the job." and "I think the second [*animation* condition] works better because the workers are more prepared at that point." This comments were categorised as *others' opinion*. Comments categorised as *random* were associated when participants mentioned that their selection was random, such as "It's really a toss up. I saw the same potential errors on the same word in both programs, so I'm just picking one." and "I am not sure, all of them seemed to perform similarly, but it looks like the third was maybe the best? Not sure." Responses categorised as *speed* are related to comments when participants mentioned that the speed of the system was a reason for their choice. One example of these responses is "Seems that the first program [*no-animation* condition] was faster and presented a complete page at once." Finally, comments such as "I didn't find any errors in the first program [*animation* condition], and it is the first on the list." were

categorised as *order*.

We refer to the sections about each individual study for the frequencies of these categories, and the discussion of the results.

5. Study 1 - Animation-performance effect in the lab

We set out to assess the potential for animation cues to influence users' perception of the performance of smart systems.

5.1. *Study Conditions*

This study included two conditions: *no animation* and *animation*. In the *animation* condition, after the loading screen, an animation was shown: on the last two lines of the input, words were highlighted one by one, with a delay of 250 milliseconds; as each handwritten word was highlighted in red, the corresponding word on the output appeared. The first word highlighted by the animation was the first word of the penultimate row: "naging" (Figure A1). The animation can be seen in full in the video accompanying this paper: `https://vimeo.com/245121953`.

5.2. *Results*

The participants' selections of the system they considered to have the best performance, the performance ratings, and the results of the statistical analysis on these ratings are reported in Table A3. The performance ratings are also summarized in Figure A3. These were higher, in aggregate, in the *animation* condition compared to the *no-animation* condition. Figure A4 illustrates the frequencies of the themes emerged from the thematic analysis (described in Section 4.2).

5.3. *Discussion*

The results of Study 1 show that animation cues have an effect on participants' perception of the system's performance. The data shows clearly that the majority of participants considered the performance of the system in the *animation* condition to be better. It should be noted that this was the case despite the fact that one error was present in the sentence highlighted by the animation. In other words, even though the animation could have drawn the participants' attention to the mistake, for most of them the animation instead had the opposite effect. The qualitative data further supports this result, most participants seem to believe that the system in the *animation* condition made fewer errors or different kind of errors than the other system, despite the two systems producing the same number and kind of errors. Moreover, participants seemed to be unconscious of the effect: none of the comments referred explicitly to the animation.

As mentioned in Section 2.1, prior research in psychology revealed participants' unconscious bias when they were asked to rate the quality of physical products (e.g. nylon stockings): identical products were rated as having different quality, based on their smell Laird (1932) or on the position in which they were displayed de Camp Wilson & Nisbett (1978). The bias was unconscious in the sense that participants in these previous studies reported (almost always) a variety of *alternative* factors to justify

their rating of quality. The results from our Study 1 extend these: in a similar unconscious way, our participants' rating of the performance of the two smart systems that we presented to them appears to be biased by the presence of the animation. The concept of 'performance' of the system, defined generically as it was in our study, can be considered relatively similar to idea of 'quality' of a physical product. However, as described above, in our study the order of presentation of the stimuli was fully counterbalanced, so the source of bias is different than these studies. Closer to our work, our results also extend those from Garcia et al. (2016), who showed that motion can influence people's perception of the performance of vacuum cleaning robots. Our results indicate that the effect of motion does not apply only to physically moving systems, but also to graphical user interfaces through animation.

These results open up a number of follow up questions. Can this effect be observed in a less controlled environment? Can it be observed for a different type of smart system? The following two experiments were designed and carried out to address these two questions.

6. Study 2 - Animation-performance effect on MTurk

To assess whether similar results to those of Study 1 can also be observed in a less controlled environment than the lab, we decided to run the same experiment on a crowdsourcing platform: Amazon Mechanical Turk (MTurk)[9]. Crowdsourcing has become a widespread online tool that researchers and companies use to outsource micro-tasks that leverage human computation, gather distributed and unbiased data, or validate results (Difallah, Catasta, Demartini, Ipeirotis, & Cudré-Mauroux, 2015; Kazai, Kamps, & Milic-Frayling, 2013; Mason & Watts, 2010).

Crowdsourcing studies can be considered less controlled for three main reasons. First, participants take part in the study from their own computers or mobile devices, rather than in a lab – so there might be external distractors that the experimenters have no control over. Second, because the study does not take place in person, participants are not directly observed by a researcher, possibly limiting a Hawthorne effect (Mayo, 2004). Third, crowdsourcing experiments have also been reported to include more diverse participants (Buhrmester, Kwang, & Gosling, 2011; Germine et al., 2012).

6.1. *Results*

The participants' selections of the system they considered to have the best performance, the performance ratings, and the results of the statistical analysis on these ratings are reported in Table A3. The selection results are also illustrated in Figure A5, and the performance ratings in Figure A6. The performance ratings were higher in the *animation* condition than in the *no-animation* condition, with statistical significance. Figure A7 illustrates the frequencies of the themes emerged from the thematic analysis (described in Section 4.2).

6.2. *Discussion*

The results of this study confirm those of Study 1. Study 2 clearly shows that the positive effect of animation cues persists even in a less controlled environment. The

[9]https://www.mturk.com

majority of participants reported that the system which contained the animation performed better than the other system. The statistically significant differences in the Likert scales results, as well as the qualitative data, further confirm this finding. Moreover, similar to Study 1, the level of detail of the responses we collected clearly shows that the participants were committed to the task, giving credibility to the data. For example, participants referred not only to the number of errors that they found in the transcribed text but also to the type of errors (e.g., "I think the first program [*no-animation* condition] had more problems distinguishing the 'a' and 'o'."). Only two participants justified their selection in terms of others' opinions or impressions, and by explicitly referring to the animation. This finding can be interpreted as confirming that the effect of motion cues is mostly unconscious. It should be noted that these references to others' opinions and to the animation emerged in Study 2, but not in Study 1. This difference could be explained by the less controlled nature of Study 2 (e.g. seeing the animations differently on their screens), and perhaps the fact that MTurk users have more experience of research studies than the participants we recruited in the University, and hence they are more likely to think about other participants' answers. Indeed, the presence of such study-trained participants may be a key limitation of MTurk, even though it only affected 2, a relatively small number, of our participants.

The alignment of the results from Studies 1 and 2 also indicates that to further investigate this phenomenon, follow-up studies can be conducted on the MTurk platform, with considerable practical advantages. Having observed the effect of animation cues in a less controlled crowdsourcing environment, we turn to investigating whether this effect is specific to the handwriting recognition system we used so far, or whether the results can be generalized to a different type of smart system.

7. Study 3 - Animation-performance effect: part-of-speech tagging

Studies 1 and 2 tested the effect of animation cues using one particular system, a handwriting recognition system. Handwriting recognition is, in its very nature, a visual task, making us wonder whether this factor alone may explain our results. So we designed a new study to assess whether the same effect would occur with a different type of system, one which involves processing that is not visual in nature. We selected a part-of-speech (POS) tagging system, a system that analyses natural language sentences and tags each word according to its syntactic function, such as article, adjective, adverb, conjunctions, noun, preposition, pronoun, and verb. Part-of-speech tagging algorithms are readily available through open source libraries[10] and their application has been suggested for different types of interfaces and visualizations (Chuang, Manning, & Heer, 2012; Yatani, Novati, Trusty, & Truong, 2011). We decided to continue to use text as the type of data handled by the smart system, for continuity with the previous studies and hence to facilitate comparison of the results.

Because not everyone might be familiar with part-of-speech tagging as a grammatical exercise, we included a *validation task* when using the part-of-speech tagging system: participants had to tag a given sentence (in English) with the part-of-speech corresponding to each word. Only those who completed this validation task with less than 3 mistakes (out of 8 words) were allowed to proceed to the main task.

[10] e.g. `http://www.nltk.org/`.

7.1. *Study Conditions*

The study included two conditions, as the previous two. However, the animation matched the new UI for the part-of-speech tagging system. In this case, the tags for the two last sentences appeared 200 milliseconds one after the other under the corresponding word. The animation can be seen in full in the video accompanying this paper: `https://vimeo.com/245121953`.

7.2. *Results*

The participants' selections of the system they considered to have the best performance, the performance ratings, and the results of the statistical analysis on these ratings are reported in Table A3. The selection results are also illustrated in Figure A8, and the performance ratings in Figure A9. The performance ratings were higher in the *animation* condition than in the *no-animation* condition, with statistical significance. Figure A10 illustrates the frequencies of the themes emerged from the thematic analysis (described in Section 4.2).

7.3. *Discussion*

The results of Study 3 extend those of Study 2, which demonstrate that the effect of animation cues on participants' perception of system performance applies also to the part-of-speech tagging system we tested. The majority of participants selected the system in the *animation* condition as the one with the best performance, and the Likert-scale ratings for this system were higher, in aggregate, than those for the *no-animation* condition, with statistical significance. Similar to Studies 1 and 2, the qualitative data collected in Study 3 indicates that participants offered a variety of reasons to justify their selections, and only 1 participant provided different answers based on the financial incentives, suggesting that most answers were not based solely on the financial incentives. Moreover, the themes emerged from the qualitative data are the same as Studies 1 and 2, further confirming the similarity of the effect on part-of-speech and on handwriting recognition systems. Such an effect, then, appears to apply even if the task performed by the system is not inherently visual, and hence if the animation does not directly mimic the task performed by the smart system.

Having observed this effect both in the lab and on MTurk, and on two different systems, we turn to the question of *why* such an effect occurs. Given that both animations highlight one word at a time, in reading order (from left to right), one option could be that the animations give users the impression that the systems process text in the same way a person would process it. Is it possible that the similarity to humans may positively influence users' attitude towards the system? This, in turn, may lead them to evaluate its performance more favourably, perhaps somehow suggesting to them that the system is "as smart as a person". An alternative explanation might involve more generally the relationship between the animations in our studies and users' expectation, or "mental model," of how the system works. Users' mental models of interactive systems have been of interest in HCI for several decades Norman (2013). The animations might induce a mental model that leads them to rate the system performance more positively. To assess the validity of these possible explanations, we designed and carried out three follow-up studies that we report in the following.

8. Study 4 - Non-human-like animation

If it is the case that the effect is due to the animations making the system appear to process information like a human, then showing an animation where the order in which the words are processed is decisively not human-like should have no effect on participants' perception of the system performance. So we designed a fourth study to test whether an animation that is decisively not human-like would still cause the same effect as the animation used in the previous studies.

Study 3 revealed that the animation effect applies in a similar way to a part-of-speech tagging system as it does to a handwriting recognition system. For simplicity, we decided to conduct further experiments on the handwriting recognition system, as it does not require the additional training and validation task described above.

8.1. *Study Conditions*

The study included two conditions: *no-animation* and *non-human-like animation*. The animation used here was similar to what was used in Study 1 (Animation-performance effect in the lab), except that the words on the last two lines of text were animated in random order, rather than left-to-right. The new animation was compared to the no-animation condition, rather than to the old animation because our interest is on whether the non-human-like animation would influence participants' perception of the system performance (compared to a system without animation), rather than on the relative strength of the two animations on the participants' perception. The animation can be seen in full in the video accompanying this paper: `https://vimeo.com/245121953`.

8.2. *Results*

The participants' selections of the system they considered to have the best performance, the performance ratings, and the results of the statistical analysis on these ratings are reported in Table A3. The selection results are also illustrated in Figure A11, and the performance ratings in Figure A12. The performance ratings were higher in the *animation* condition than in the *no-animation* condition, with statistical significance. Figure A13 illustrates the frequencies of the themes which emerged from the thematic analysis (described in Section 4.2).

8.3. *Discussion*

The majority of participants in Study 4 selected the system in the *non-human-like animation* condition as the one with the best performance, and the Likert-scale ratings for this system were higher, in aggregate, than those for the *no-animation* condition, with statistical significance. The analysis of qualitative data is also very much in line with that of our previous studies. These results indicate that the effect we observed in previous studies can be observed also for an animation that can be interpreted as non-human-like. Therefore the tentative explanation suggested above, that the effect of animations in Studies 1 to 3 may be related to making the system appear more human-like can be rejected. While this study only considered one type of non-human-like animation, numerous other options for this kind of animation could be considered, such as animations where the text appears to be processed "by letter" (e.g. system

first processes all os, then all us, then all as, then all ps, and so on). So further research could be conducted to study the effects of other types of non-human-like animations on participants' perception of system performance. In what follows we turn to the option that the effect of animations may be due to the more general relationship between the animation and a user's mental model of the smart system.

9. Study 5 - Animation-performance effect and mental model salience

A new study was designed to investigate the relationship between users' mental models of handwriting recognition systems and the animations we displayed in earlier studies. In particular, in this study participants were asked one open question about their idea of how a handwriting recognition system works, to check whether these explanations are compatible with the animations used in our prior studies. The additional question was asked after the initial questionnaire about demographic information and before the main task. In an attempt to prevent spurious answers, participants were required to submit answers containing at least 20 words. The study then followed the same structure as the one we explained in studies' methodology section, with the addition that at the end we also asked participants whether their experience of using the system matched their initial idea of how it works.

9.1. *Results*

The participants' selections of the system they considered to have the best performance, the performance ratings, and the results of the statistical analysis on these ratings are reported in Table A3. The selection results are also illustrated in Figure A14, and the performance ratings in Figure A15. The performance ratings were higher in the *animation* condition than in the *no-animation* condition, with statistical significance. Figure A16 illustrates the frequencies of the themes emerged from the thematic analysis (described in Section 4.2).

9.1.1. How people think a handwriting recognition system works

The responses to the question regarding how participants think that the handwriting recognition works were analysed through thematic analysis. Two themes emerged in our analysis: *match with database* and *image recognition*. The theme *match with database* included responses that mention using a database to compare the words or characters identified in the handwritten text, such as "The program analyses the written text. It then compares each character to a database loaded into it [...]". The theme *image recognition* was associated to responses that mention how the program processes images to extract characters and words, such as: "It scans the handwriting into an image and then the program look[s] at the image pixel by pixel to match each individual letter [...]'.

The answers of the participants suggest that the majority seems to have a shared mental model of how they expect the system to work. In general, participants agree that somehow the system has to detect the words or letters to digitise them. Of the 16 participants, 4 stated that the system matches the words and letters using some form of optical recognition. Three participants mentioned that the handwritten text needs to be matched with a 'collection' of some kind, a database or library, containing labelled examples of handwritten text, to find the corresponding letter or word. 8 participants

considered that the system needs a combination of recognizing an image and compare it with a collection. This suggests that most participants have a mental model that could provide a plausible explanation for non-experts on how a handwriting system works which is reinforced by the animation.

9.1.2. The system worked as participants expected

All participants reported that both systems successfully transcribed the handwritten text to typed text, and so they considered that the systems worked as they expected. Additionally, only three participants mentioned in their comments the animation (e.g. "For the second program, it showed how the program scanned each word in red. It was computing for the e-text").

9.2. *Discussion*

The explanations that participants provided about how a handwriting recognition system works seem to be quite in line with the animation that we implemented, even though the explanations were provided before seeing it. This seems to be true regardless of the participants' education level. The match, however, is not always an exact one: 14 out of the 16 participants explained that the recognition would happen character by character, so in the same way a human would actually type handwritten text into a computer. In contrast, the animation implemented in the previous studies can be interpreted as processing the text word by word rather than character by character.

The results from Study 5 seem to be in stark contrast to those from previous studies. Only 5 participants selected the system in the *animation* condition as the one with the best performance level, and the analysis of the likert-scale ratings did not reveal statistically significant differences between the conditions, despite the sample size being the same as in the earlier studies. The different results can be attributed to the additional question about participants' mental model of handwriting recognition systems asked at the beginning of the study.

Arguably, asking participants how they think a handwriting recognition system works, makes their mental model for this kind of system *salient* to them. This salience seems to contrast the effect of the animation that we observed in earlier studies. Perhaps, then, making participants aware of how the system works has an effect similar to that of the animation in our earlier studies. In other words, these results suggest that in our earlier studies the animations reminded participants of how the smart system works, instead in Study 5 the preliminary question had the same effect, so it seems to have replaced the effect of the animation (for both conditions). While there might be alternative interpretations of these findings, the one we put forward resonates with studies in psychology which demonstrated that making a bias salient to participants may remove the effect of the bias. In particular Schwarz & Clore (1983) demonstrated through a well known study about the effect of weather on mood that asking participants about the weather (and hence making the weather salient to them) removes the effect that weather has on mood (at least in the case of bad weather). Similarly, in our study asking participants about how the system works seems to remove the effect of the animation[11]. We further explore the relationship between

[11]This effect may be reminiscent of 'priming,' however, we prefer to refer to 'salience' rather than priming because priming generally refers to the situation where participants are shown a stimulus that causes an implicit memory effect (Meyer & Schvaneveldt, 1971). In our study participants are explicitly asked to think, or remember, how a handwriting recognition system could work, so we do not consider this as related to implicit

mental models and the effect of animations on perceptions of performance in the following study.

10. Study 6 - Animation-performance effect with alternative mental models

The results of the previous study suggest that the animation used in previous studies is aligned to participants' mental models of how handwriting recognition systems work. Based on this finding, a possible explanation of the results from earlier studies is that the animations we displayed "reassured" participants that the system works as they expected. As such, the animation raised their confidence in the system and enhanced their perception of its performance. Which impact would it have on people's perception, if we give them an explanation of how the system works that does not match the animation that is shown?

To test this, for Study 6, we formulated two explanations of how the system works for two different animations. The *original* animation is the animation used in the previous studies. We also designed a new animation, which we refer to as the *alternative* animation, and we formulated a corresponding explanation. The *alternative* animation consisted of enclosing each word with a rectangle and inverting its colour, before displaying the corresponding word on the right hand side of the screen. Figure A18 shows a frame of the *alternative* animation, while the full animation can be seen in the video accompanying this paper: `https://vimeo.com/245121953`. This alternative animation was designed to be at odds with the explanations collected from participants in Study 5 about how a handwriting recognition works. For consistency, both animations, original and alternative, included exactly the same transcription errors. Figure A17 shows the two explanations participants received before the task. The explanations were designed to be superficial and to relate to the animations, rather than to provide a realistic or in depth description of how the system actually works. This is because, *for our study*, participants do not really need to understand how such a system works, but only think that they understand. How handwriting recognition works in reality is much more complex and it might not be possible to explain it with a simple animation.

10.1. *Study Design*

A fully counterbalanced order, between-participants design was used. The two animations and the two explanations described above define 4 conditions in a 2×2 fashion: (original animation, original explanation), (original animation, alternative explanation), (alternative animation, alternative explanation), and (alternative animation, original explanation). In the first and third conditions, animation and explanation are *matching*, while in the second and fourth they are *mismatching*. Each participant was assigned to one of these 4 conditions. Similar to the methodology we explain in previous sections, each participant was asked to evaluate and compare the performance of two handwriting recognition systems: one involving an animation (*animation* condition) and one with no animation (*no-animation*). The *no-animation* condition, which was similar to the previous studies, was the same for all participants. The *animation* condition would involve either the *original* or the *alternative* animation, depending on

---

memory

the assigned group of the participant.

In summary, then, the study included:
64 participants in
4 groups

- orig. animation, orig. explanation
- orig. animation, alt. explanation
- alt. animation, alt. explanation
- alt. animation, orig. explanation

2 conditions each

- *animation* vs.
- *no-animation*.

10.2. *Results*

Matching conditions. The participants' selections of the system they considered to have the best performance, the performance ratings, and the results of the statistical analysis on these ratings are reported in Table A3. The performance ratings were higher for the *animation* condition than for the *no-animation* condition, with statistical significance. Figure A19 shows participants rating of the performance of the systems. Moreover, Figure A20 illustrates the frequencies of the themes emerged from the thematic analysis (described in Section 4.2).

Mismatching conditions. The participants' selections of the system they considered to have the best performance, the performance ratings, and the results of the statistical analysis on these ratings are reported in Table A3. In this case the difference was not statistically significant. In addition, Figure A20 illustrates the frequencies of the themes emerged from the thematic analysis (described in Section 4.2).

10.2.1. System working according to expectations

Matching conditions. Overall, 27 out of the 32 participants in the matching conditions indicated that the systems worked as they expected from the explanation given at the beginning of the study, while the remaining 5 stated that it did not. In more detail, participants considered that the system compares each word with a database. As such, 3 participants believed the system would process the data word by word rather than character by character based on the errors they found (e.g. "It seemed the program did it letter by letter, not by the word as described. But then again I don't know the language, so changing one letter like the programs did may still have been recognizing a word."). Other 2 participants mentioned that they believed that the systems did not work because only one system showed the animation (e.g. "The second computer program highlights the words as it transcribes them. The first didn't appear to do that."). In the free text comments, 10 participants mentioned the animation explicitly as a reason of why they considered the system to work as they expected (e.g. "I could see the text being highlighted and picked apart", and "Because you could see the process of transcription as it was happening.").

Mismatching conditions. Overall, 25 of the 32 participants indicated that the system worked as expected from the explanation given at the beginning of the study, while the remaining 7 stated it did not. In the free text comments, 10 participants mentioned the animation. In more detail, 3 participants mentioned that the anima-

tions mismatched the explanation. Moreover, the 3 participants who reported that the system did not work as they expected mentioned that they thought the system transcribed the handwritten text word by word rather than character by character because of errors they found in the typed text. In contrast, other participants felt that the system transcribed the handwritten text better than they expected. Because of this, they felt that the system worked correctly.

Participants' responses to why they considered that the systems worked according to their expectations (or not) were categorised through thematic analysis. Each response was associated with one or two themes, with eight themes in total: *generic, animation, faith, disbelief, correctness, analysis, experience,* and *technology.* Figure A21 illustrates the frequencies of these themes for the **matching conditions** and **mismatching conditions**. The theme *generic* was associated to responses where participants did not provide full explanation or misunderstood the question, such as "I believe they translated the handwritten text into a digital computer font". The theme *animation* was used when participants talked about why the animation affected their consideration of whether the systems are working or not, such as: "Because you could see the process of transcription as it was happening". We grouped a response into the theme *faith* if it is related to the participants believing the explanation provided: "I had no reason to doubt the explanation, it seemed perfectly reasonable". Comments grouped into *disbelief* is the opposite, and instead it's related to situations where the participants do not believe in the explanation provided: "I don't see how changing the color of the handwritten text to match the color of the paper as a way to convert the text to etext [...]". Responses related to the accuracy of the output, such as "Yes, it translated the characters of the handwritten text correctly", were categorised as *correctness.* The theme *analysis* was used to categorise comments that talk about the actual transcription process, such as "It appears that the program goes through each letter and tries to identify which letter it is". The theme *experience* was used for any comments in which the participant talk about his/her own personal experience with handwriting recognition systems: "I've used OCR [Optical Character Recognition] programs before and they were never as accurate as this one was, so I don't believe it actually exists". Finally comments such as "Technology and artificial intelligence is growing at an exponential rate" were categorised as *technology.*

10.3. *Discussion*

The results of this study seem to confirm what was suggested by the findings of Study 5: animation cues influence the perception of the system performance only if they are consistent with the participant's mental model. More in general, taken together with the results of Study 5, these results allow us to propose the following explanation for the effect:

> Animation cues suggest the user an explanation of how the smart system works. If this explanation is largely consistent with the user's own mental model of the system, i.e. if this explanation appears plausible to the user, they then feel reassured about how the system functions, and therefore are inclined to have higher confidence in the system output, compared to an alternative system for which they have no cues about how it may work.

We used the word 'largely' because the results of Study 5 indicate that the explanations for how the system works provided by participants do not match the animation *exactly.* However, if the explanation is radically different, as in the *mismatching* con-

ditions of Study 6, the effect disappears.

Moreover, as mentioned above, the results of Studies 1 to 4 suggest that this effect takes place largely unconsciously – most participants did not mention the animation in their justification for the selection of the system with the best performance. Similarly, in Study 5 we observed that if the explanation of how the system works is made salient to participants, in that case through an initial question, the effect of the animation disappears. Arguably, the explanations that participants provided in Study 5 could apply to *both* systems that they evaluated, so that process reminded them, or made them aware, of how *both* systems work. Hence, their judgement was not biased towards either of them. It should be noted that in Study 6 participants were also exposed to an explanation of how a handwriting recognition system works at the beginning of the study. However, in this case the explanation was *presented* to participants, who just had to read it, which requires less effort and reflection than having to come up with an explanation and writing it down. Moreover, in the *matching* condition, the explanations provided in Study 6 matched very closely the animation shown to them. These differences *may explain* why the effect of animation was still observed in Study 6, but not in Study 5.

Even though the importance of mental models in HCI has been discussed for at least three decades (e.g. Kieras & Bovair (1984); Norman (2013)), most prior work focussed on the effect of mental models on *users' performance* when using an interactive system. In contrast, our results suggest a relationship between mental models and users' perception of the *system's performance*. Reflecting on the physical motion cues presented by Garcia et al. (2016), it would be interesting to explore if people also rated the robot's performance higher because seeing the motion made them think that they understand how the robot works.

In the qualitative data, we did not find comments of people explaining that they perceived a match or mismatch between the explanation that elicits a mental model and the animation they received. This behaviour suggests that the effect of animation cues happens unconsciously. Thus, we can argue that the participants' comments and evaluation make visible that indeed the fact that the animation matches participants' mental model affects their perception.

Now that we found a possible explanation for why the *animation* cues influence people's perception on how they perceive smart systems' performance, we move to further characterize this effect with the following two studies.

11. Study 7 - Amount of animation detail

Through the previous studies, we found that animation cues can influence how people evaluate the performance of screen-based systems. As a subsequent step, we evaluate whether the amount of detail of a displayed motion, so how much animation needs to be shown in all the elements related to the system's task (e.g. handwritten text and e-text), can have an impact on participants' perception. We expect to find a relationship between the amount of motion displayed and the perceived performance.

To explore this, we designed a new animation, which involves less motion than the animations used in previous studies.

11.1. *Study Conditions*

The study included 3 conditions: *animation*, *partial-animation*, and *no-animation*. The number of participants recruited was 16 for each of these 3 conditions, i.e. 48 in total. Similar to prior studies, each condition corresponded to a system that participants were asked to evaluate and compare in terms of performance. The new *partial-animation* condition is similar to the *animation* condition, except that instead of involving the animation on both the input and output parts of the UI (i.e. on both the handwritten and typed text), it only applies to the output part of the UI (approximately the right half of the screen), while the input part of the UI remains static.

11.2. *Results*

11.2.1. Performance ratings

Median values for the performance evaluation for the *animation*, *partial-animation*, and *no-animation* were $Mdn = 4.5$, $Mdn = 4$, and $Mdn = 4$, respectively. A CHI-squared test revealed statistically significant differences in the performance ratings, $\chi^2(2) = 9.73, p = 0.008$. Post-hoc analysis through pairwise Wilcoxon Signed-ranks tests with significance level set at $p < 0.05$, revealed significant differences between the *animation* and the *no-animation* conditions ($Z = 3.037, p = 0.002, r = 0.31$), and also between the *partial-animation* and the *no-animation* conditions ($Z = 2.64, p = 0.014, r = 0.25$). No significant differences were instead found between the *animation* and *partial-animation* conditions ($Z = 0.89, p = 0.375, r = 0.09$). Figure A23 shows participants evaluation of the performance of the systems.

11.2.2. Selection of the system with the best performance

Overall, 26 of the 48 participants (54%) selected the system in the *animation* condition as the one with the best performance, 9 participants (19%) selected instead the system in the *partial-animation* condition, 7 participants (15%) the system in the *no-motion* condition, while the remaining 6 participants (12%) suggested that the three systems had the same performance. These results are illustrated in Figure A22. Figure A24 illustrates the frequencies of the themes emerged from the thematic analysis (described in Section 4.2).

11.3. *Discussion*

The results of Study 7 suggest that *any amount* of animation seems to influence users' perception of the performance of the system: statistically significant differences in the Likert-scale ratings were found both between *no-animation* and *animation* and between *no-animation* and *partial-animation*, while no statistically significant differences were found between *animation* and *partial-animation*. However, in terms of choosing the system with the best performance, the majority of participants opted for the *animation* condition, rather than any of the other 3 options, regardless of the experimental reward. Indeed, almost three times as many participants opted for the system in the *animation* condition compared to the one in the *partial-animation* one. In contrast to the Likert-scale results, the selection results suggest that the amount of animation does play some role in users' perception of performance. So perhaps the lack of statistical significance mentioned above could be a limitation of our sample

size.

Once again, similar to prior studies the qualitative data from Study 7 indicates that participants took the task seriously and engaged with it.

12. Study 8 - Animation-performance effect vs real systems performance

In all studies reported so far, the presence of animation was the only difference across the systems our participants evaluated. The performance of the various systems, defined in terms of number of errors produced by the system was kept constant. To further characterise the effect we identified, we decided to test what level of imbalance in the performance level of the system being compared would "break the illusion" created by the animation. In other words: how many additional errors can the animation cover? Study 8 was designed to address this question, by comparing pairs of systems with different numbers of mistakes.

12.1. *Study Design*

The study design was based on the previous studies, but we increased the number of errors in the text by one unit at the time, with a separate group of participants for each number. As detailed below, with 10 errors the effect was no longer observed, so we stopped at this number. In other words, the study included only two groups: *9-errors* group and *10-errors* group. The *no-animation* condition always included just 8 errors, as in previous studies (in earlier studies the number of errors was the same across the conditions). The new errors were added randomly around the two paragraphs without adding more errors in the last two sentences were the animation is displayed. Moreover, the new errors were similar to the previous ones, we substituted letters, such as 'a' with 'o' or 'g' with "q'.

12.2. *Results*

For the *9-errors* group, the majority of participants selected the system in the *animation* condition as the one with the best performance, and the performance ratings were higher in the *animation* condition than in the *no-animation* condition, with statistical significance. In the *10-errors* group no statistically significant differences were found between the conditions.

The data summary and the results of the statistical analysis for both are reported in Table A3. The selection results are also illustrated in Figure A25, and the performance ratings in Figure A27. Figure A26 illustrates the frequencies of the themes emerged from the thematic analysis for the *9-error* group (described in Section 4.2). The themes for the *10-error* group are illustrated in Figure A10.

12.3. *Discussion*

The results of Study 8 indicate that the effect of animation cues on participants' perception of the system performance holds, to some extent, even when comparing two systems that have different performance levels. In particular, within the 9-errors group most participants selected the system in the *animation* condition, even when it produced one additional mistake compared to the system in the *no-animation* condi-

tion (corresponding to a performance degradation of 12.5%). When the difference in number of errors produced by the two systems becomes 2 (the 10-errors group, which corresponds to a performance degradation of 25%), the animation system is no longer selected as the one with the best performance by the majority of participants, but only by 4 participants (25%). However, even in the 10-errors group 6 participants (37.5%) suggested that the two systems have the same performance, and that's as many as those who correctly selected the system in the *no-animation* condition as the one with the best performance. This finding is reinforced by the qualitative data, which shows that in both the 9-errors group and the 10-errors group, some participants suggested that there are fewer errors in the *animation* condition compared to the *no-animation* condition. More in general, from this study, we can learn that the positive effect of animation cues can persist even when a system's performance is degraded. In other words, our findings show that the *animation* cues tend to hide a possible malfunction of the system.

13. Summary and General Discussion

Our initial three studies revealed that animation cues integrated into the GUI of a smart system can affect people's perception of the system performance, extending and generalising the results reported by Garcia et al. (2016) for physical motion cues and vacuum cleaning robots. In particular, in Study 1 – Animation-performance effect in the lab (N=16) participants reported a handwriting recognition system to perform better when animation cues are displayed than when they are not. Study 2 – Animation-performance effect on MTurk (N=16) replicated the same experiment on MTurk, extending the initial results to a less controlled environment, and demonstrating that further studies could be conducted on the online platform. Study 3 – Animation-performance effect: part-of-speech tagging (N=16) demonstrates that the effect of animation is not specific to the type of smart system used in Studies 1 and 2, similar results were observed also for a part-of-speech tagging system, one which involves a type of data processing that is less inherently visual than the handwriting recognition system.

Studies 4, 5 and 6 were designed to look for an explanation for this effect. In particular, Study 4 – Non-human-like animation (N=16) examined and ruled out the possibility that the animation cues may induce users to perceive that the system recognises the handwriting as a person would, and so appears to be "as smart as a person." Study 5 - Animation-performance effect and mental model salience (N=16) provides an initial exploration of the relationship between the animation and participants' mental model of the system. Study 6 – Animation-performance effect with alternative mental models (N=64) probed such relationship further: its results suggest that animation cues affect participants' perception of the system performance, only if the animation matches their mental model of the system. More in general, the results of Studies 5 and 6 suggest that if the animation cues are largely compatible with a user's mental model of the system, they then act as a *reminder* for how the system works and they can increase the user's confidence in the system (at least compared to alternative systems for which they have no cues about how it works).

Once we found the reason behind why *animation* cues influence participants' perception of smart systems, we designed and conducted two further studies to characterize the observed phenomenon more in detail. In Study 7 – Amount of animation detail (N=48), we analysed whether the amount of animation shown would influence partic-

ipants' perception of a system's performance. The results indicate that any amount of animation seem to have potential to influence users perception of a system performance. Finally, Study 8 – Animation-performance effect vs real systems performance (N=32) assessed the effect of *animation* cues when a system's performance actually decreases, compared to the alternative system where no animation is integrated. The results of Study 8 show that that the effect of animation cues on participants' perception of the system performance holds, to some extent, even in this case. In particular, most participants still favour the system with animation even when it makes one error more than the the system with no animation. However, when the number of extra errors becomes 2, the effect decreases.

14. Implications

Our studies bear implications for the design of user interfaces for smart systems, and in particular the design of visual feedback around such systems. Overall, our results imply that designers should be aware that including animations in the UI can bias users' perception of how well a smart system works. In particular, we found that if the animations are largely consistent with users' mental models of how the system works, they can lead users to have a more positive perception of the system performance. There seems to be a risk, then, that animations may inadvertently lead users to rely on the results of a smart system more than they should. Indeed, our last study indicates that animations can even 'cover up' some of the errors made by the system. Such over-reliance can have undesirable consequences, if not even dramatic, especially for safety-critical applications (Parasuraman & Riley, 1997).

Moreover, our results seem to suggest that if animations are in conflict with users' mental model of the system, they could have a negative effect on the perception of the system performance. Particular care, then, needs to be taken to make sure that a user's mental model can be reliably predicted, so that animations can be made compatible with it. However, this may be particularly difficult for systems which involve more complex forms of machine learning than the handwriting recognition and part-of-speech tagging systems that we studied. On one hand, users may not be able to easily guess how these kinds of systems work. On the other hand, it may be difficult to map mental models for such more advanced systems to a simple animation.

The effects reported in this paper could potentially affect mundane animations such as loading screens, transitions or even decorative animations. These are often found on web and desktop applications, especially when data is being processed or loaded from remote servers. Beyond handwriting recognition and part-of-speech tagging systems, similar effects could be expected, for example, for mobile and web applications like translators, image recognition systems, recommender systems or chatbots in, e.g., automated customer support. For example, translators can show that they are translating the text by typing the translated text on the screen. In an image recognition scenario, the system can visualize that it is detecting the contour of an object and then tag the appropriate object. Another possible animation that systems could use is detecting and then isolating the object and showing that it is making a comparison with another object of its database. Recommender systems could display how users' information is compared with a database to generate a recommendation.

15. Further Research Opportunities

While our results show an evident effect, they also open up a number of new research questions. For example, we found that people's mental model is the reason why animation cues influence people's perception. However, what happens with incomplete or incorrect mental models that users may form for systems that are based on complex machine learning approaches? Do animations still influence performance perception? If not, how simple does a system need to be so that the effect of the animation can be observed? Systems that are more complex than the ones in our studies may require longer lasting and more complex animations, in order to convey the *right* mental model. Would such more complex animations produce similar results to the ones we observed? More research is needed to address these questions.

More in general, further work is needed to assess whether these results may apply to other types of smart systems, since our studies focussed only on functionality related to text. For example, the relationship between mental model and performance perception could be probed for robots, for example robotic vacuum cleaners. While prior work Garcia et al. (2016) demonstrated that seeing a robot moving leads participants to perceive it as performing better than a similar robot which they do not see in motion, it falls short of evaluating whether such an effect is related to the mental model of how the robot works.

Our work so far focussed only on short-term effects. While this is a needed initial step, future work should look into whether there are any long-term effects. Addressing this question would require field deployments of systems augmented with animations – in such context artificially controlling the performance level of the system being tested may be particularly difficult. Moreover, if an animation ended up being integrated in the majority of the systems that people interact with in their everyday life activities, would there be a saturation effect? In particular, people could feel overwhelmed by all the visual feedback they would receive. In any case, short term effects would still be relevant for some applications where users interact with the system only once, or infrequently (e.g., seeing self-checkout machines processing users' purchases, ATM processing people's transactions, or information kiosks processing people's request).

Another strand of potential further work is related to specific characteristics of the animations, such as their speed: does varying the animation time frame change how people perceive systems' performance? Moreover, in our studies, only the last two lines of text of a document containing 12 lines were animated. This corresponds to approximately 17% of the content being animated. Further work should assess whether the amount of content that is animated, relative to the total content displayed, makes a difference. Such variations may be particularly important when considering variations in the real system performance (as in our *Study 8: Animation-performance effect vs real systems performance*): can animations "cover up" decreases in system performance only when they draw the participants' attention to a small part of the data?

16. Conclusion

In this paper, we presented eight studies, conducted mainly on the crowdsourcing platform Amazon Mechanical Turk, which explored whether visual animation cues can change people's perception of how well smart systems perform their task. In these studies, we further investigated the characteristics, which are crucial for the animations to change perception. Indeed, our results suggest that displaying a high detail of

animations that match people's mental model, can influence people's perception of the performance of smart systems that we tested: one based on a handwriting recognition software and the other a part-of-speech tagging system. We were able to show that this effect holds, even when these systems have a minimal decrease in their performance.

While this modality has the potential to improve users' ratings of a smart system's performance, it may also pose the danger of making a system appear to work better than it actually does. We present a number of design implications to guide researchers. Furthermore, we describe a number of further research opportunities that could enhance our understanding of the effects of this modality. We expect that the results presented in this paper will stimulate designers to consider integrating animations as a feedback of their systems, and researchers to explore this area further.

17. Acknowledgements

This work was supported by a PhD scholarship by CONACyT, SICYT Morelos, and by the EPSRC A-IoT project (EP/N014243/2). Study 1 was approved by the University of Southampton Ethics Committee (ref: 17155), while the rest of the studies were approved by the UCLIC Ethics Committee (ref: UCLIC/1617/017/Staff Costanza/Nowacka/Yang) at University College London. The data referred to in this paper can be found at http://dx.doi.org/10.5258/SOTON/397976. ANONYMOUS.

References

Ariely, D. (2008). *Predictably irrational*. HarperCollins New York.

Bakhshi, S., Shamma, D. A., Kennedy, L., Song, Y., de Juan, P., & Kaye, J. J. (2016). Fast, cheap, and good: Why animated gifs engage us. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 575–586). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2858036.2858532` doi:

Bellotti, V., & Edwards, K. (2001, December). Intelligibility and accountability: Human considerations in context-aware systems. *Hum.-Comput. Interact.*, *16*(2), 193–212. Retrieved from `http://dx.doi.org/10.1207/S15327051HCI16234_05` doi:

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. Retrieved from `http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa` doi:

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk. *Perspectives on Psychological Science*, *6*(1), 3-5. Retrieved from `http://dx.doi.org/10.1177/1745691610393980` (PMID: 26162106) doi:

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Chang, B.-W., & Ungar, D. (1993). Animation: From cartoons to the user interface. In *Proceedings of the 6th annual acm symposium on user interface software and technology* (pp. 45–55). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/168642.168647` doi:

Chuang, J., Manning, C. D., & Heer, J. (2012, October). Without the clutter of unimportant words: Descriptive keyphrases for text visualization. *ACM Trans. Comput.-Hum. Interact.*, *19*(3), 19:1–19:29. Retrieved from `http://doi.acm.org/10.1145/2362364.2362367` doi:

de Camp Wilson, T., & Nisbett, R. E. (1978). The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology*, *41*(2), 118-131. Retrieved from `http://www.jstor.org/stable/3033572`

Detenber, B. H., & Reeves, B. (1996). A bio-informational theory of emotion: Motion and image size effects on viewers. *Journal of Communication*, *46*(3), 66–84. Retrieved from http://dx.doi.org/10.1111/j.1460-2466.1996.tb01489.x doi:

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., & Cudré-Mauroux, P. (2015). The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th international conference on world wide web* (pp. 238–247). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2736277.2741685

Dittrich, W. H., & Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception*, *23*(3), 253-268. Retrieved from http://pec.sagepub.com/content/23/3/253.abstract doi:

Dragicevic, P., Bezerianos, A., Javed, W., Elmqvist, N., & Fekete, J.-D. (2011). Temporal distortion for animated transitions. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2009–2018). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1978942.1979233 doi:

Fritz Heider, M. S. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243-259. Retrieved from http://www.jstor.org/stable/1416950

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*(12), 1845-1853. Retrieved from http://pss.sagepub.com/content/21/12/1845.abstract doi:

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 669. doi:

Garcia, P. G., Costanza, E., Ramchurn, S. D., & Verame, J. K. M. (2016). The potential of physical motion cues: Changing people's perception of robots' performance. In *Proceedings of the 2016 acm international joint conference on pervasive and ubiquitous computing* (pp. 510–518). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2971648.2971697 doi:

Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. *Causal cognition: A multidisciplinary debate*, 150–184.

Germine, L., Nakayama, K., Duchaine, B., Chabris, C., Chatterjee, G., & Wilmer, J. (2012). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847-857. Retrieved from http://dx.doi.org/10.3758/s13423-012-0296-9 doi:

Harrison, C., Amento, B., Kuznetsov, S., & Bell, R. (2007). Rethinking the progress bar. In *Proceedings of the 20th annual acm symposium on user interface software and technology* (pp. 115–118). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1294211.1294231 doi:

Harrison, C., Yeo, Z., & Hudson, S. E. (2010). Faster progress bars: Manipulating perceived duration with visual augmentations. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1545–1548). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1753326.1753556 doi:

Ju, W., & Takayama, L. (2009). Approachability: How people interpret automatic door movement as gesture. *International Journal of Design*, *3*(2).

Kazai, G., Kamps, J., & Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, *16*(2), 138-178. Retrieved from http://dx.doi.org/10.1007/s10791-012-9205-0 doi:

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, *8*(3), 255–273. Retrieved from http://dx.doi.org/10.1207/s15516709cog0803 doi:

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 5686–5697). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2858036.2858529 doi:

Laird, D. A. (1932). How the consumer estimates quality by subconscious sensory impressions.

Journal of Applied psychology, *16*(3), 241.

Lim, B. Y., & Dey, A. K. (2011a). Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (pp. 157–166). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2037373.2037399 doi:

Lim, B. Y., & Dey, A. K. (2011b). Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on ubiquitous computing* (pp. 415–424). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2030112.2030168 doi:

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2119–2128). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1518701.1519023 doi:

Lyons, J. (2013). Being transparent about transparency: A model for human-robot interaction.. Retrieved from http://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5712

Mason, W., & Watts, D. J. (2010, May). Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, *11*(2), 100–108. Retrieved from http://doi.acm.org/10.1145/1809400.1809422 doi:

Mayo, E. (2004). *The human problems of an industrial civilization.* Routledge.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227. doi:

Michotte, A. (1963). *The perception of causality.* Basic Books.

Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition.* Basic books.

Nowacka, D., Hammerla, N. Y., Elsden, C., Plötz, T., & Kirk, D. (2015). Diri - the actuated helium balloon: A study of autonomous behaviour in interfaces. In *Proceedings of the 2015 acm international joint conference on pervasive and ubiquitous computing* (pp. 349–360). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2750858.2805825 doi:

O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., & Höllerer, T. (2008). Peerchooser: Visual interactive recommendation. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1085–1088). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1357054.1357222 doi:

Padrao, G., Gonzalez-Franco, M., Sanchez-Vives, M. V., Slater, M., & Rodriguez-Fornells, A. (2016). Violating body movement semantics: Neural signatures of self-generated and external-generated errors. *NeuroImage*, *124*, 147 - 156. Retrieved from http://www.sciencedirect.com/science/article/pii/S1053811915007314 doi:

Pantelis, P. C., & Feldman, J. (2012). Exploring the mental space of autonomous intentional agents. *Attention, Perception, & Psychophysics*, *74*(1), 239–249.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253.

Park, D., & Lee, J.-H. (2010a). Investigating the affective quality of motion in user interfaces to improve user experience. In H. S. Yang, R. Malaka, J. Hoshino, & J. H. Han (Eds.), *Entertainment computing - icec 2010: 9th international conference, icec 2010, seoul, korea, september 8-11, 2010. proceedings* (pp. 67–78). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-15399-0 doi:

Park, D., & Lee, J.-H. (2010b, December). Understanding how the affective quality of motion is perceived in the user interface. *Comput. Entertain.*, *8*(2), 14:1–14:11. Retrieved from http://doi.acm.org/10.1145/1899687.1899696 doi:

Popović, J., Seitz, S. M., & Erdmann, M. (2003, October). Motion sketching for control of rigid-body simulations. *ACM Trans. Graph.*, *22*(4), 1034–1054. Retrieved from http://doi.acm.org/10.1145/944020.944025 doi:

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press Cambridge, UK.

Schlottmann, A., & Surian, L. (1999). Do 9-month-olds perceive causation-at-a-distance? *Perception*, *28*(9), 1105-1113. Retrieved from `http://pec.sagepub.com/content/28/9/1105.abstract` doi:

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology*, *45*(3), 513.

Takayama, L., Dooley, D., & Ju, W. (2011). Expressing thought: Improving robot readability with animation principles. In *Proceedings of the 6th international conference on human-robot interaction* (pp. 69–76). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1957656.1957674` doi:

Talbot, J., Lee, B., Kapoor, A., & Tan, D. S. (2009). Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1283–1292). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1518701.1518895` doi:

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943-951. Retrieved from `http://pec.sagepub.com/content/29/8/943.abstract` doi:

Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 31–40). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1240624.1240630` doi:

Tversky, A., & Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Environmental impact assessment, technology assessment, and risk analysis* (pp. 107–129). Springer.

Verame, J. K. M., Costanza, E., & Ramchurn, S. D. (2016). The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 4908–4920). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2858036.2858369` doi:

Vermeulen, J. (2010). Improving intelligibility and control in ubicomp. In *Proceedings of the 12th acm international conference adjunct papers on ubiquitous computing - adjunct* (pp. 485–488). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1864431.1864493` doi:

Vermeulen, J., Luyten, K., Coninx, K., & Marquardt, N. (2014). The design of slow-motion feedback. In *Proceedings of the 2014 conference on designing interactive systems* (pp. 267–270). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2598510.2598604` doi:

Ware, C. (2012). *Information visualization: perception for design*. Elsevier.

Yang, R., & Newman, M. W. (2013). Learning from a learning thermostat: Lessons for intelligent systems for the home. In *Proceedings of the 2013 acm international joint conference on pervasive and ubiquitous computing* (pp. 93–102). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2493432.2493489` doi:

Yatani, K., Novati, M., Trusty, A., & Truong, K. N. (2011). Review spotlight: A user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1541–1550). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1978942.1979167` doi:

Appendix A. Responses independent of the reward

As mentioned above, at the end of each study we asked participants which system they considered as having the best performance, without taking into account what the majority of others would select, and independently of the reward they would receive. This was done to check whether the consensus-oriented financial incentive would distort considerably our participants' answers. Because in most cases participants gave the same answer as to the previous question, for the sake of brevity we excluded this data from the main body of the paper, but we report it in Table A4 for completeness.

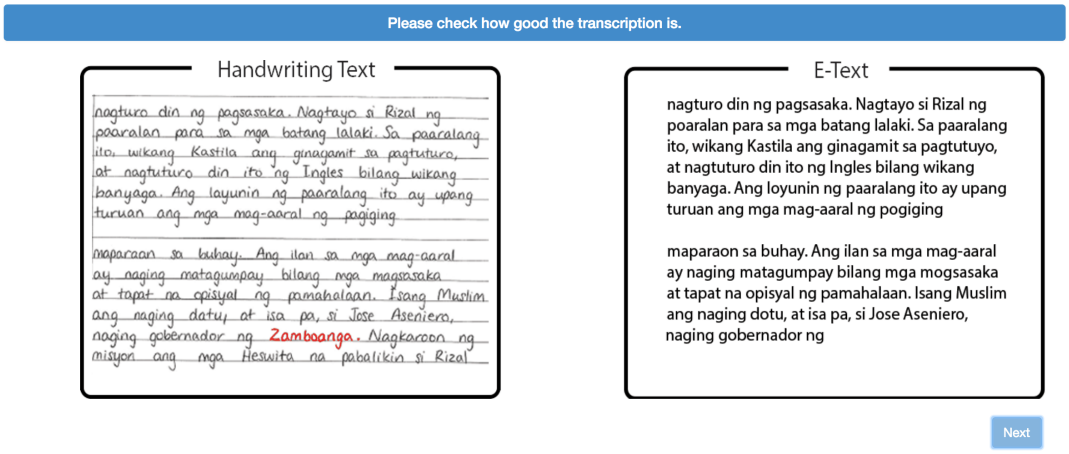

Figure A1. The interface of the handwriting recognition system presented in all studies except Study 3.

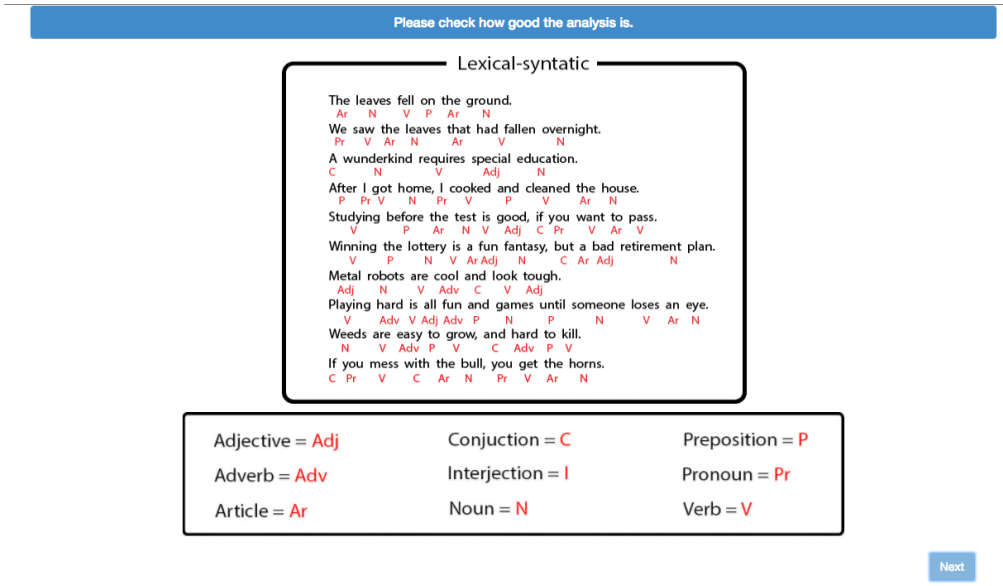

Figure A2. The interface of the part-of-speech tagging system presented in Study 3

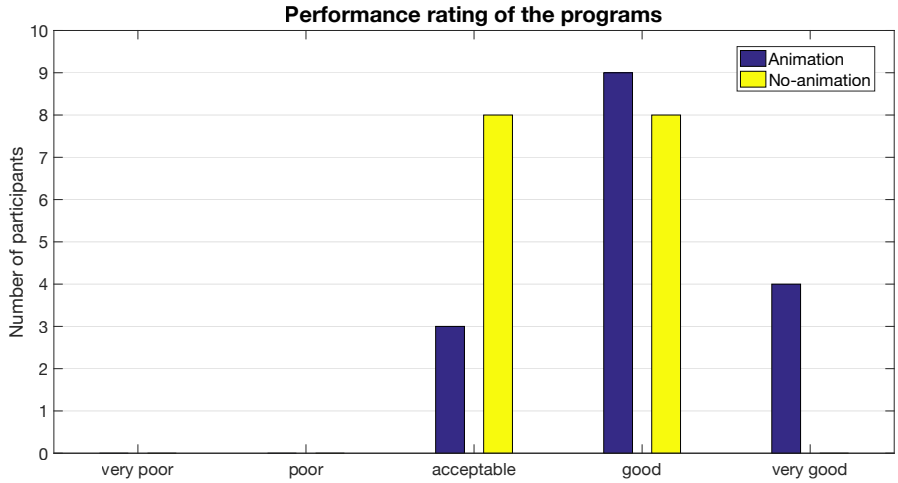

Figure A3. Participants' rating (on a Likert scale) of the system performance in each condition in Study 1.

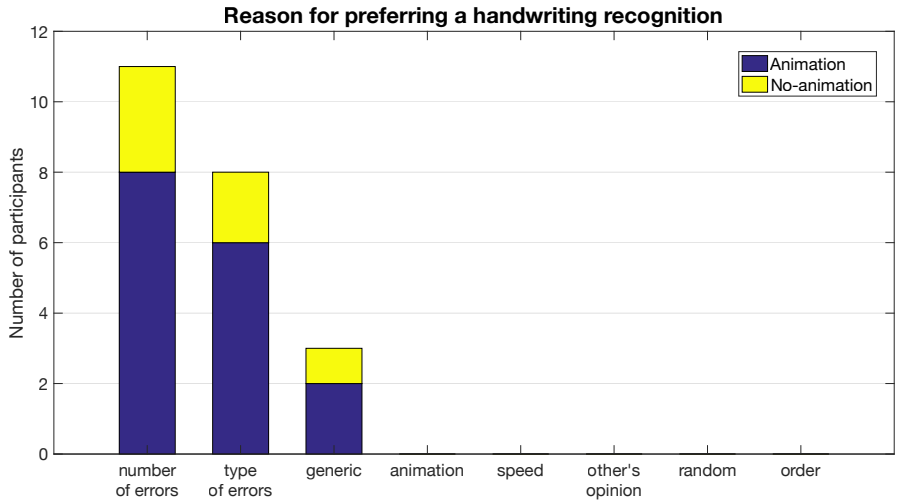

Figure A4. Reasons expressed by participants for selecting the system they considered as having the best performance in Study 1.

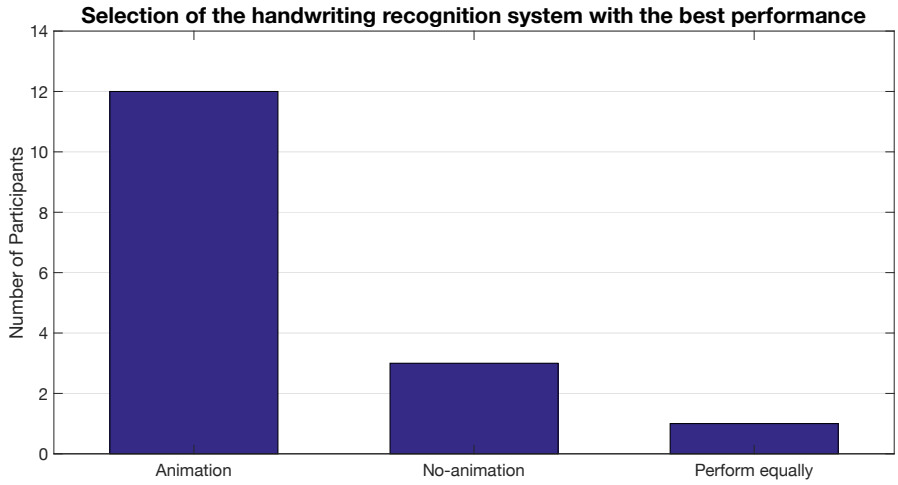

Figure A5. Selections of the system considered to have the best performance in Study 2.

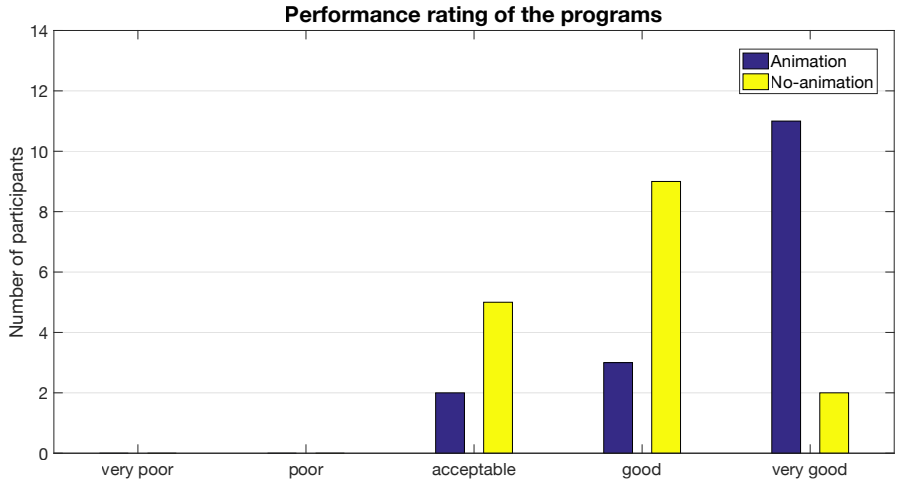

Figure A6. Participants' rating (on a Likert scale) of the system performance in each condition in Study 2.

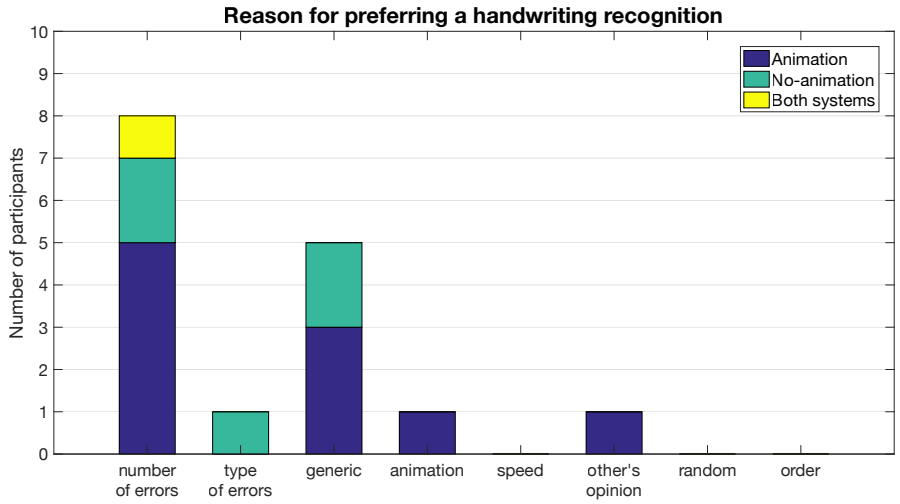

Figure A7. Reasons expressed by participants for selecting the system they considered as having the best performance in Study 2.

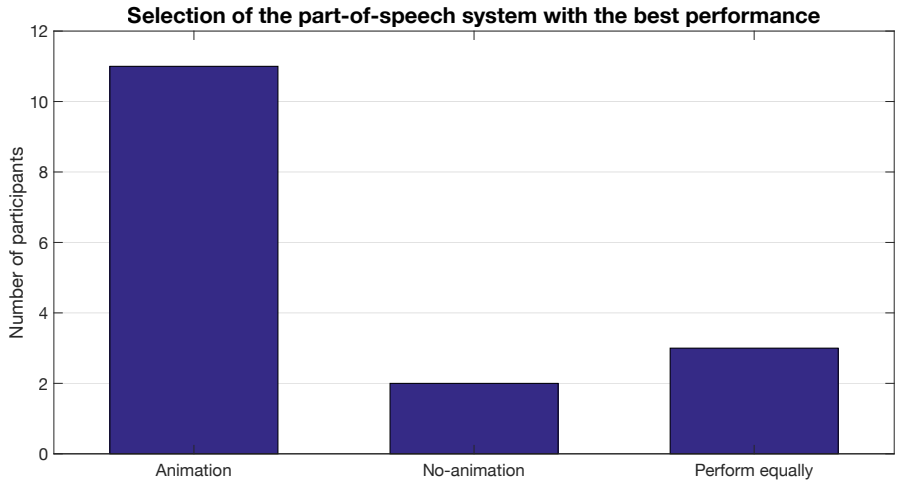

Figure A8. Selections of the system considered to have the best performance in Study 3.

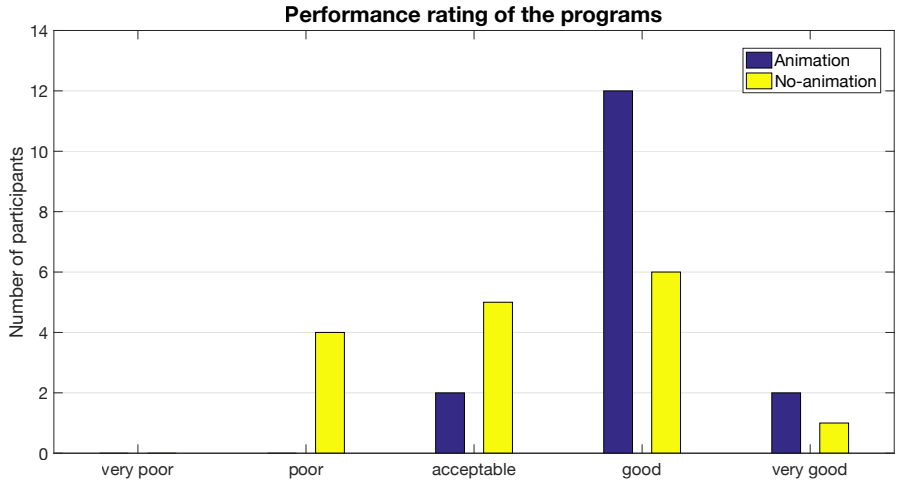

Figure A9. Participants' rating (on a Likert scale) of the system performance in each condition in Study 3.

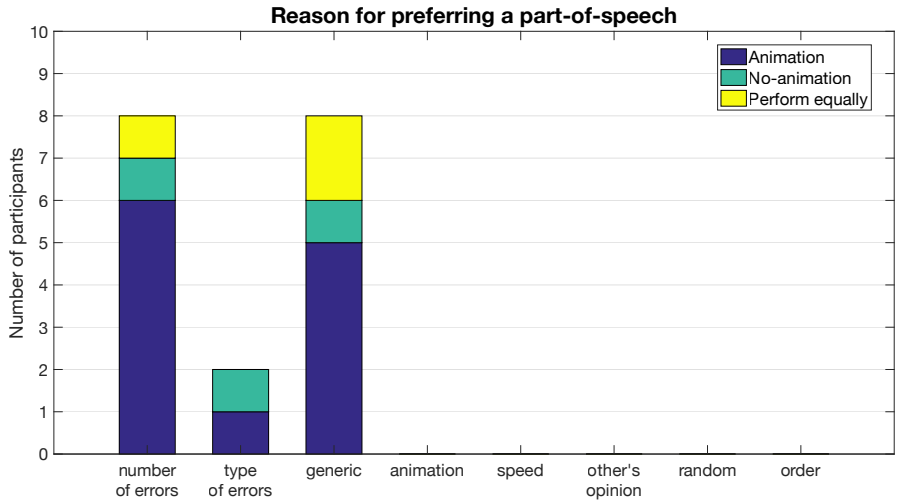

Figure A10. Reasons expressed by participants for selecting the system they considered as having the best performance in Study 3.

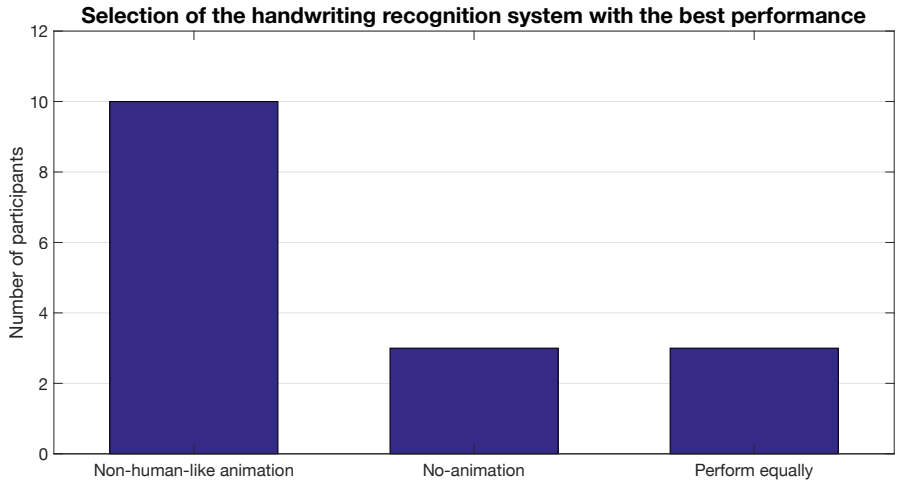

Figure A11. Selections of the system considered to have the best performance in Study 4.

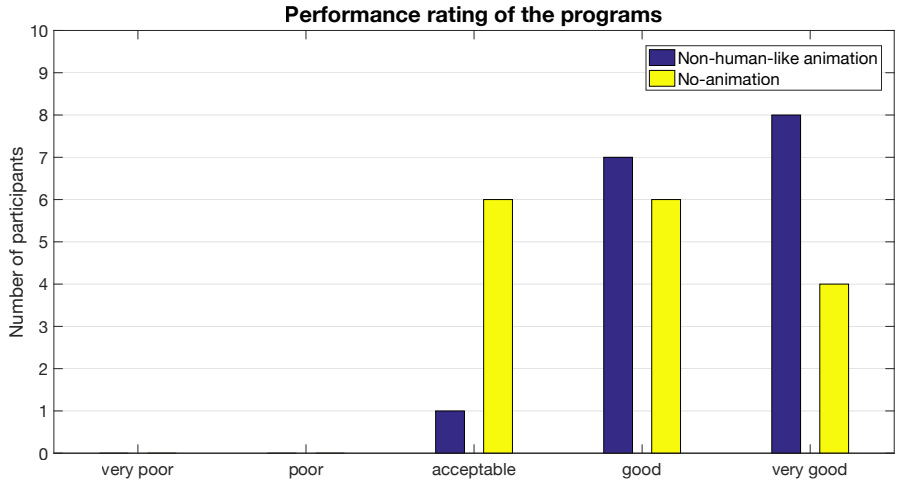

Figure A12. Participants' rating (on a Likert scale) of the system performance in each condition in Study 5.

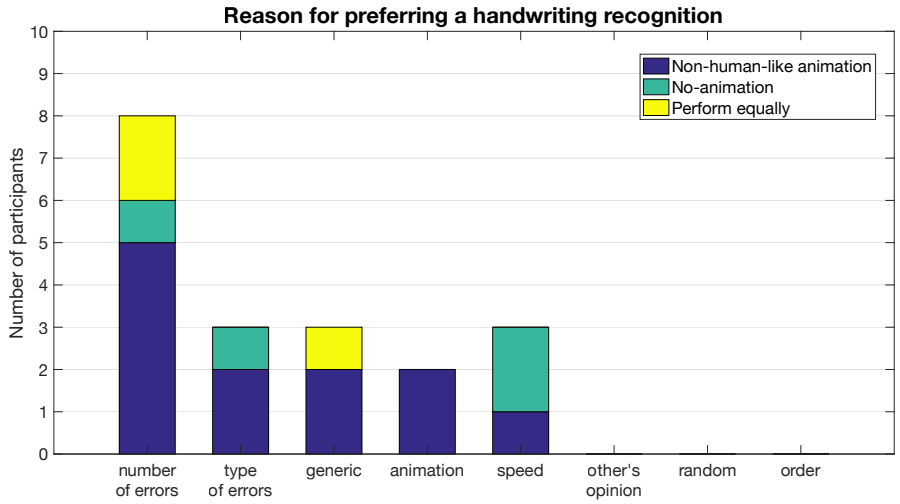

Figure A13. Reasons expressed by participants for selecting the system they considered as having the best performance in Study 5.

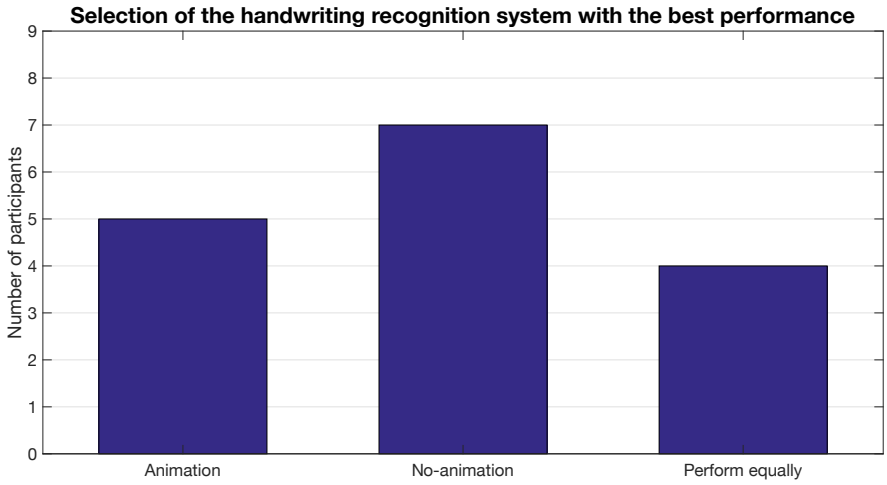

Figure A14. Selections of the system considered to have the best performance in Study 5.

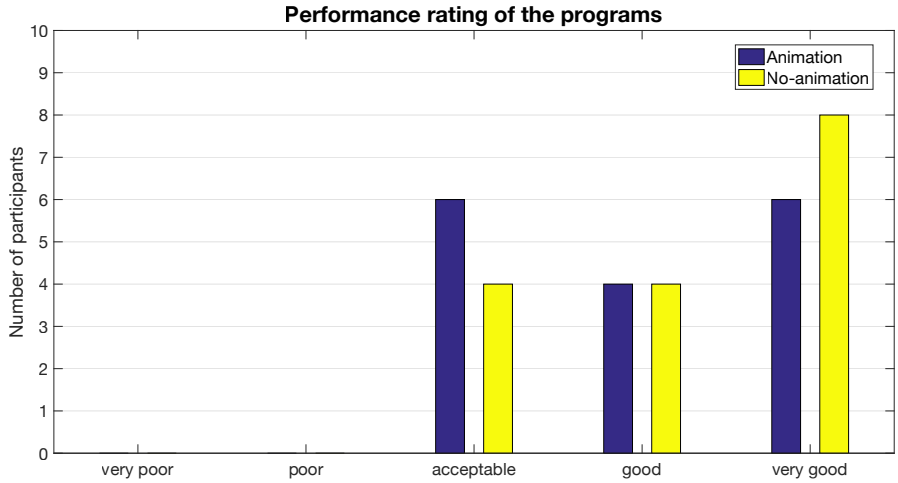

Figure A15. Participants' rating (on a Likert scale) of the system performance in each condition in Study 5.

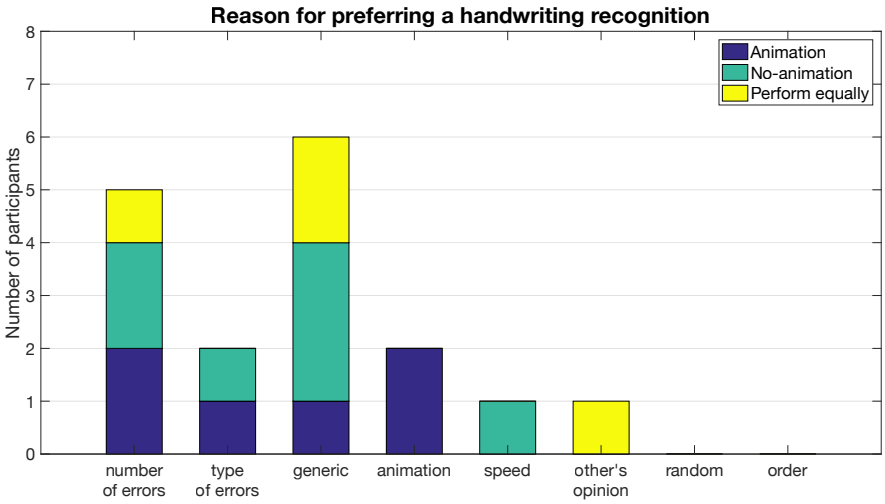

Figure A16. Reasons expressed by participants for selecting the system they considered as having the best performance in Study 5.

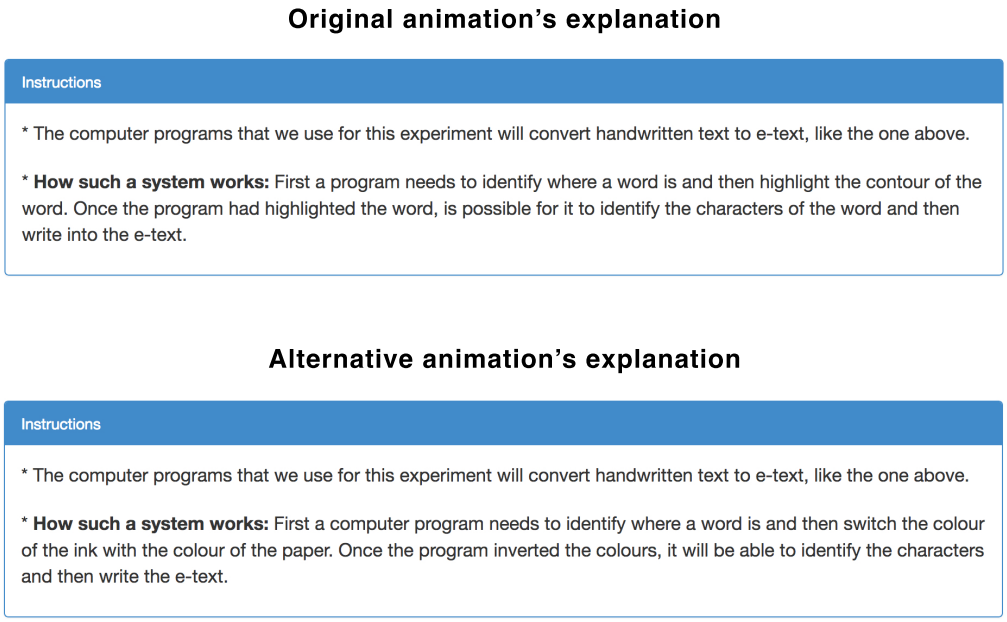

Figure A17. Explanations participants received of how the animations worked in Study 6.

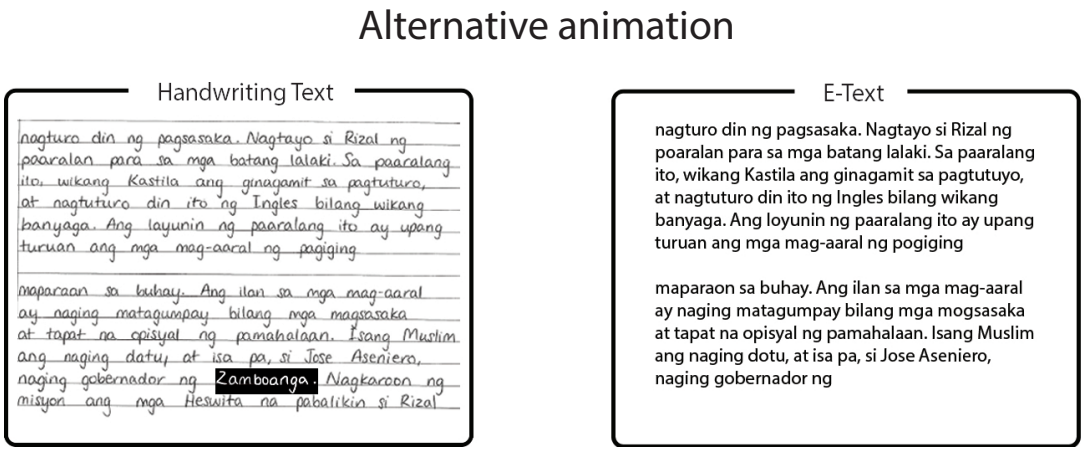

Figure A18. The *alternative* animation used in Study 6.

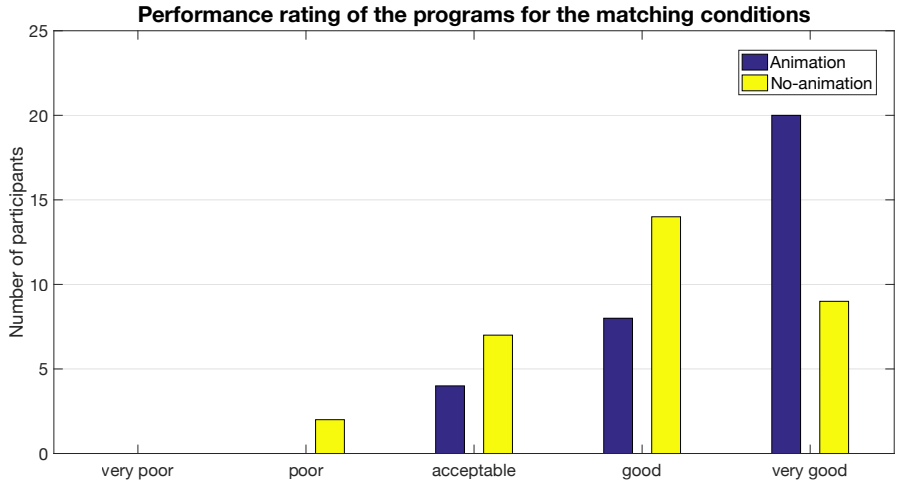

Figure A19. Participants' rating (on a Likert scale) of the system performance for the matching conditions in Study 6.

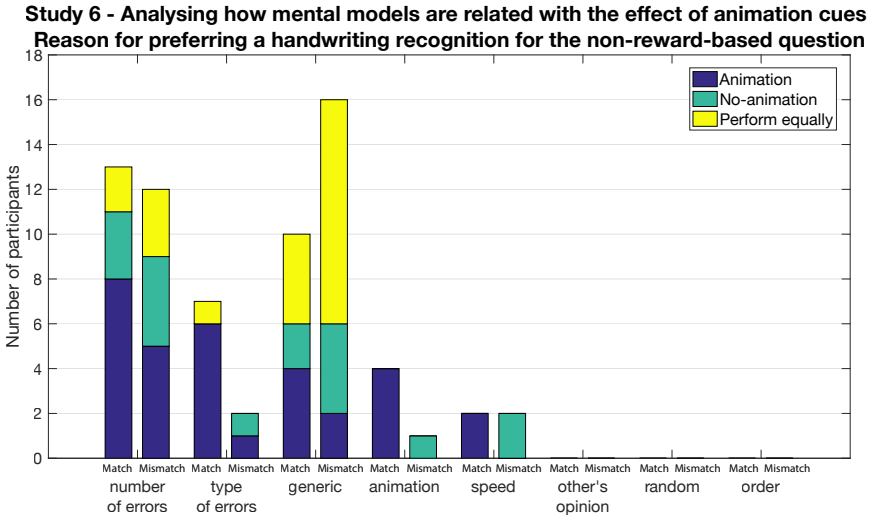

Figure A20. Reasons expressed by participants for selecting the system they considered as having the best performance for the matching and mismatching conditions in Study 6.

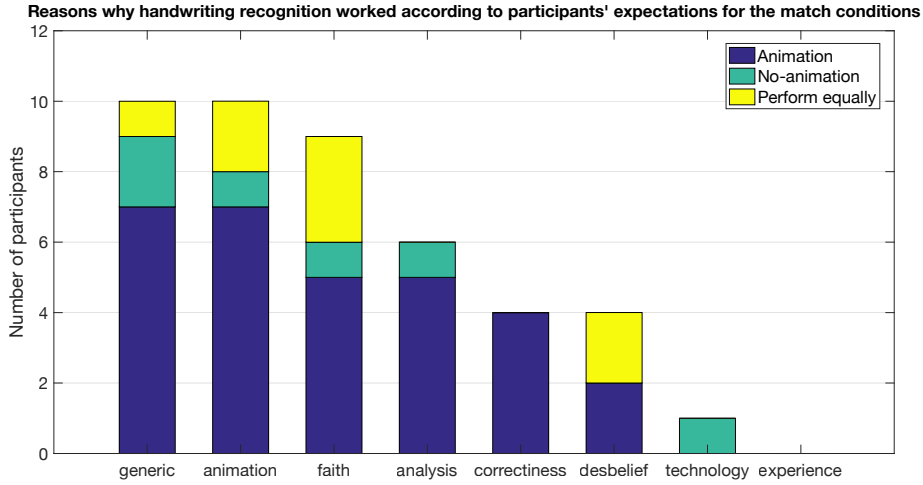

Figure A21. Reasons expressed by participants for why they considered that the systems worked according to their expectations for the matching conditions in Study 6.

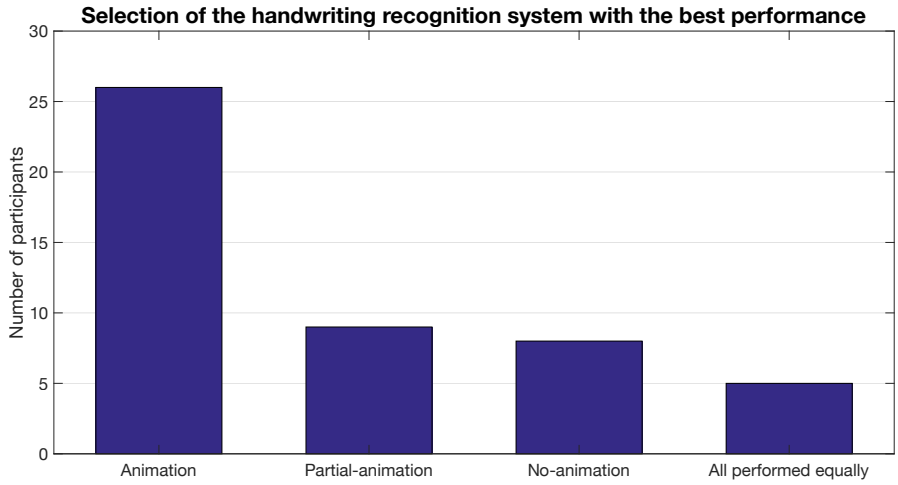

Figure A22. Selections of the system considered to have the best performance in Study 7.

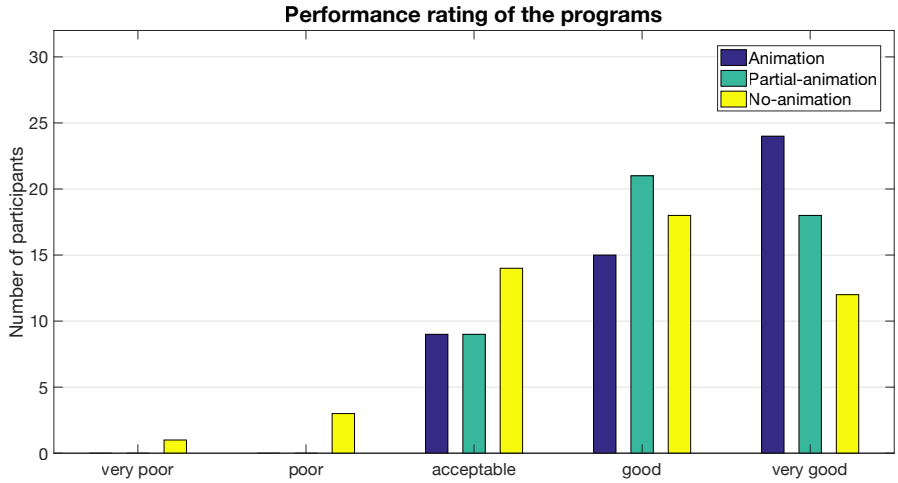

Figure A23. Participants' rating (on a Likert scale) of the system performance in each condition in Study 7.

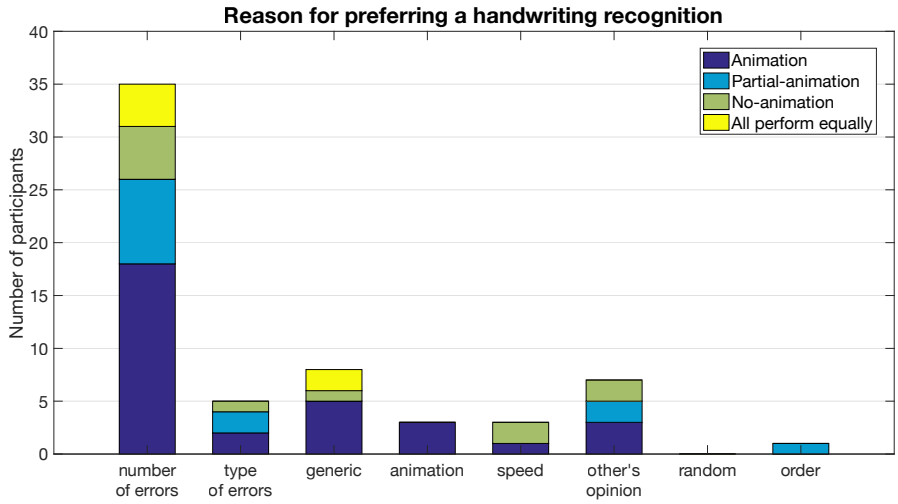

Figure A24. Reasons expressed by participants for why they considered that the systems worked according to their expectations in Study 7.

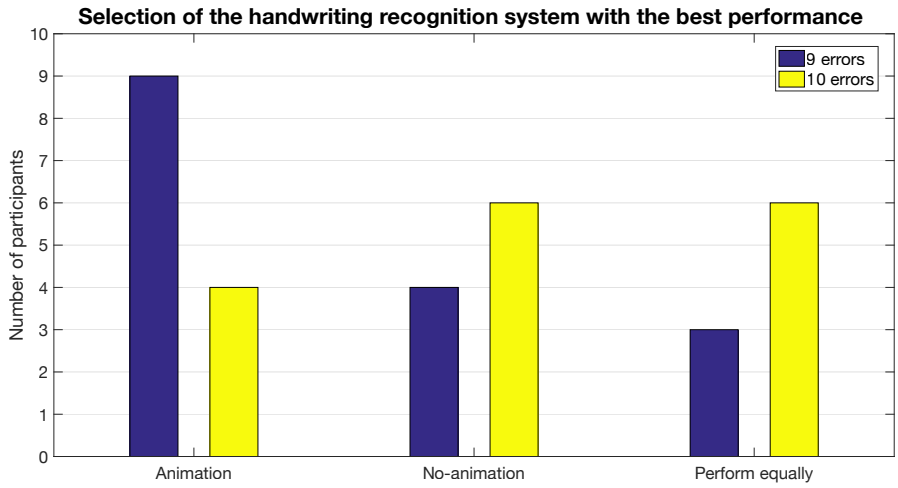

Figure A25. Selections of the system considered to have the best performance in Study 8 for the 9-errors and 10-errors groups.

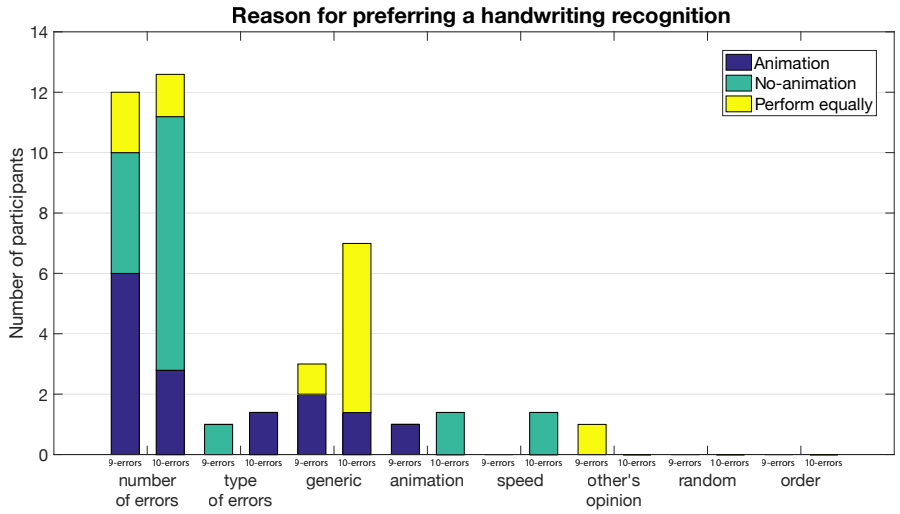

Figure A26. Reasons expressed by participants for selecting the system they considered as having the best performance for the 9-errors and 10-errors groups in Study 8.

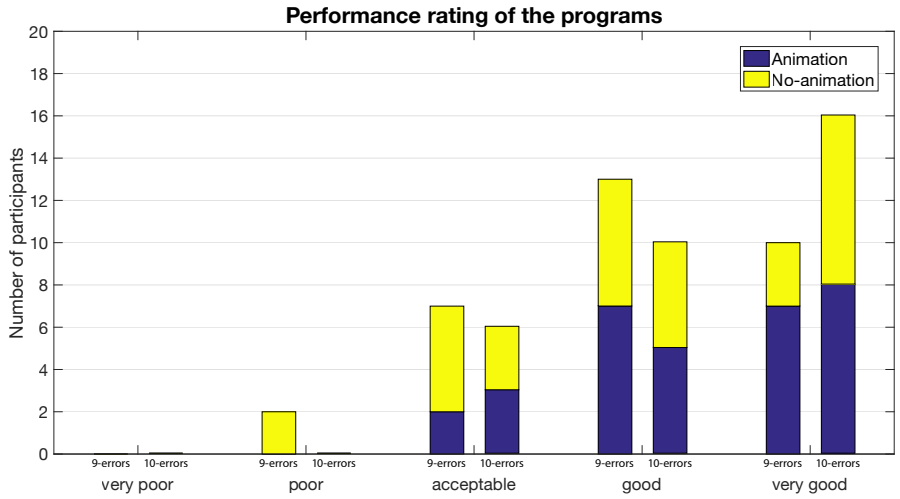

Figure A27. Participants' rating (on a Likert scale) of the system performance in each condition for the 9-errors and 10-errors groups in Study 8.

Study	Time	Fixed amount
Study 2	7 minutes	$1.17
Study 3	8 minutes	$1.33
Study 4	7 minutes	$1.17
Study 5	8 minutes	$1.33
Study 6	8 minutes	$1.33
Study 7	12 minutes	$2.00
Study 8	7 minutes	$1.17

Table A1. Rewards participants received to take part in the crowdsourcing studies. In each study, participants also received a bonus if they selected the system which the majority identified as the one with the best performance.

Study	N (N males)	Age	Education	Nationality
Study 1 – Animation-performance effect in the lab	16 (10)	18 to 24 $M = 20.68$ $SD = 1.70$	undergraduate and postgraduate students	n.a.
Study 2 – Animation-performance effect on MTurk	16 (10)	21 to 44 $M = 33$ $SD = 5.94$	1 master's degree 10 university degree 5 secondary school	16 US
Study 3 – Animation-performance effect: part-of-speech tagging	16 (5)	21 to 54 $M = 26$ $SD = 10.02$	9 university degree 4 secondary school 3 primary school	12 US 1 KR 1 BD 1 CA 1 BE
Study 4 – Non-human-like animation	16 (13)	22 to 44 $M = 31$ $SD = 7.16$	7 university degree 7 secondary school 2 primary school	15 US 1 DE
Study 5 – Animation-performance effect and mental model salience	16 (11)	22 to 37 $M = 29.5$ $SD = 4.76$	2 master's degree 9 university degree 5 secondary school	15 US 1 KR
Study 6 – Animation-performance effect with alternative mental models	64 (34)	20 to 61 $M = 32$ $SD = 11.22$	1 doctoral degree 5 masters' degree 39 university degree 19 secondary school	62 US 1 KR 1 CA
Study 7 – Amount of animation detail	48 (31)	22 to 44 $M = 34$ $SD = 9.53$	2 master's degree 30 university degree 15 secondary school 1 primary school	46 US 1PL 1UK
Study 8 – Animation-performance effect vs real systems performance	16 (11)	18 to 61 $M = 263.5$ $SD = 11.35$	1 masters degree 7 university degree 7 secondary school 1 primary school	16 US

Table A2. Participants' demographic information

Study	Best performing system			Median rating		Wilcoxon Signed-rank Test
	Anim.	No-anim.	Same	Anim.	No-anim.	
Study 1 - Animation-performance effect in lab	12 (75%)	4 (25%)	0	$Mdn = 4$	$Mdn = 3.5$	$Z = 2.07, p = 0.039, r = 0.37$
Study 2 - Animation-performance effect on MTurk	12 (75%)	3 (19%)	1 (6%)	$Mdn = 5$	$Mdn = 4$	$Z = 2.45, p = 0.014, r = 0.43$
Study 3 - Animation-performance effect: part-of-speech tagging	11 (69%)	2 (12%)	3 (19%)	$Mdn = 4$	$Mdn = 3$	$Z = 2.55, p = 0.011, r = 0.45$
Study 4 Non-human-like animation	10 (62%)	3 (19%)	3 (19%)	$Mdn = 5$	$Mdn = 4$	$Z = 2.07, p = 0.039, r = 0.37$
Study 5 Animation-performance effect and mental model salience	5 (31%)	7 (44%)	4 (25%)	$Mdn = 4$	$Mdn = 4.5$	$Z = 1.03, p = 0.31, r = 0.18$
Study 6 Animation-performance effect with alternative mental models						
- matching conditions	21 (66%)	5 (15%)	6 (19%)	$Mdn = 5$	$Mdn = 4$	$Z = 2.94, p = 0.003, r = 0.37$
- mismatching conditions	10 (31%)	12 (38%)	10 (31%)	$Mdn = 4$	$Mdn = 4.5$	$Z = 2.94, p = 0.003, r = 0.37$
Study 8 Animation-performance effect vs real system's performance						
- 9-erros group	9 (56%)	4 (25%)	3 (19%)	$Mdn = 4$	$Mdn = 4$	$Z = 2.183, p = 0.029, r = 0.39$
- 10-erros group	4 (25%)	6 (37.5%)	6 (37.5%)	$Mdn = 4$	$Mdn = 4$	$Z = 0, p = 1, r = 0$

Table A3. Experimental results from all the studies, except Study 7 (because of a different structure), and statistical test results. *Best performing system* refers to the system which participants selected as the one performing best.

Study	Best performing system			Best performing system for the non-reward question		
	Anim.	No-anim.	Same	Anim.	No-anim.	Same
Study 1 - Animation-performance effect in lab	12 (75%)	4 (25%)	0	12 (75%)	3 (19%)	1 (6%)
Study 2 - Animation-performance effect on MTurk	12 (75%)	3 (19%)	1 (6%)	11 (69%)	4 (25%)	1 (6%)
Study 3 - Animation-performance effect: part-of-speech tagging	11 (69%)	2 (12%)	3 (19%)	12 (75%)	2 (12%)	2 (12%)
Study 4 Non-human-like animation	10 (62%)	3 (19%)	3 (19%)	10 (62%)	3 (19%)	3 (19%)
Study 5 Animation-performance effect and mental model salience	5 (31%)	7 (44%)	4 (25%)	5 (31%)	8 (50%)	3 (19%)
Study 6 Animation-performance effect with alternative mental models - matching conditions - mismatching conditions	 21 (66%) 10 (31%)	 5 (15%) 12 (38%)	 6 (19%) 10 (31%)	 21 (66%) 8 (25%)	 5 (15%) 12 (38%)	 6 (19%) 12 (38%)
Study 8 Animation-performance effect vs real system's performance - 9-erros group - 10-erros group	 9 (56%) 4 (25%)	 4 (25%) 6 (37.5%)	 3 (19%) 6 (37.5%)	 8 (50%) 4 (25%)	 4 (25%) 6 (37.5%)	 4 (25%) 6 (37.5%)

Table A4. Experimental results from all the studies, except Study 7 (because of different structure), and statistical test results. *Best performing system* refers to the system participants selected as the one having the best performance *without* taking into account what the majority of participants would choose.